SIAM J. MATH. ANAL.
Vol. 35, No. 1, pp. 1–32

ON THE SPECTRA OF THREE-DIMENSIONAL LAMELLAR SOLUTIONS OF THE DIBLOCK COPOLYMER PROBLEM*

XIAOFENG REN† AND JUNCHENG WEI‡

Abstract. One-dimensional free energy local minimizers are viewed as three-dimensional lamellar-type critical points in a box. To determine whether they model the lamellar phase of diblock copolymers in the strong segregation region, we analyze their spectra. We obtain the asymptotic expansions of their eigenvalues and eigenfunctions. Consequently we find that they are stable, i.e., are local minimizers in space, only if they have sufficiently many interfaces. Interestingly the one-dimensional global minimizer is near the borderline of three-dimensional stability.

Key words. spectrum, three-dimensional stability, lamellar solution, diblock copolymer

AMS subject classifications. 35J55, 34D15, 45J05, 82D60

DOI. 10.1137/S0036141002413348

1. Introduction.

In a diblock copolymer melt a molecule is a linear chain consisting of two subchains grafted covalently to each other. The first subchain has N_A type A monomer units, and the second subchain has N_B type B monomer units. In polymer systems even a weak repulsion between unlike monomers A and B induces a strong repulsion between subchains. With many chain molecules in a polymer melt the different type subchains tend to segregate below some critical temperature, but as they are chemically bonded in chain molecules, even a complete segregation of subchains cannot lead to a macroscopic phase separation. Only a local microphase separation occurs: microdomains rich in A and B are formed. These microdomains form morphological patterns/phases in a larger scale. The commonly observed phases include the spherical, cylindrical, and lamellar, depicted in Figure 1.

We consider a scenario in which a diblock copolymer melt is placed in a domain D and maintained at fixed temperature. D is scaled to have unit volume in space. Let $a = N_A/(N_A + N_B) \in (0,1)$ be the relative number of the A monomers in a chain molecule. Similarly $b = N_B/(N_A+N_B)$, so $a+b = 1$. The relative A monomer density field u is an order parameter. $u \approx 1$ stands for high concentration of A monomers. The melt is incompressible so the relative B monomer density is $1-u$, and $u \approx 0$ stands for high concentration of B monomers.

Ohta and Kawasaki [10] introduced an equilibrium theory in which the free energy of the system is a functional of the relative A monomer density:

$$I(u) = \int_D \left\{ \frac{\epsilon^2}{2}|\nabla u|^2 + \frac{\sigma}{2}|(-\Delta)^{-1/2}(u-a)|^2 + W(u) \right\}, \tag{1.1}$$

defined in $X_a = \{u \in W^{1,2}(D) : \overline{u} = a\}$, where $\overline{u} := \frac{1}{|D|}\int_D u$ is the average of u on D. The original formula in [10] is given for the whole space. The expression here on a bounded domain D first appeared in Nishiura and Ohnishi [8].

*Received by the editors August 26, 2002; accepted for publication (in revised form) December 13, 2002; published electronically June 10, 2003.
http://www.siam.org/journals/sima/35-1/41334.html

†Department of Mathematics and Statistics, Utah State University, Logan, UT 84322-3900 (ren@math.usu.edu).

‡Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong (wei@math.cuhk.edu.hk). This author was supported in part by a direct grant from CUHK and an earmarked grant of RGC of Hong Kong.

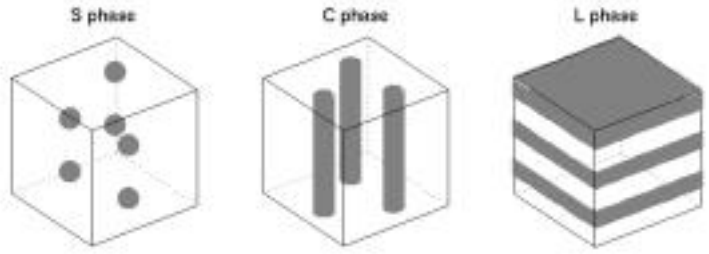


FIG. 1. *The spherical, cylindrical, and lamellar morphology phases commonly observed in diblock copolymer melts. The dark color indicates the concentration of type A monomer, and the white color indicates the concentration of type B monomer.*

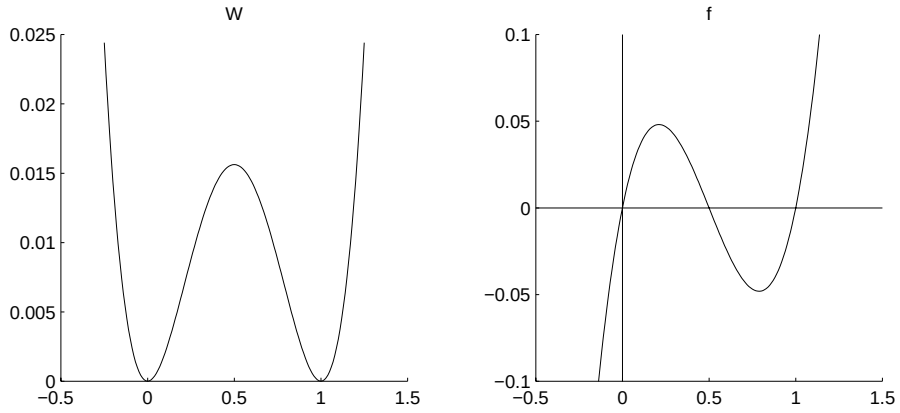


FIG. 2. *The graphs of W and $f = W'$.*

ϵ and σ are positive dimensionless parameters that depend on various physical quantities such as N_A, N_B, the average distance between two adjacent monomers in a chain, the interaction between monomers, the temperature, and the size of the sample. In the strong segregation region where morphology patterns form, ϵ is very small. The size of σ in this paper is chosen to be of order ϵ; i.e., there is a fixed positive constant γ so that $\sigma = \epsilon\gamma$. This particular parameter range is realized if we take the sample size to be of the $(N_A + N_B)^{2/3}$ order.[1]

The local function W is smooth and has the shape of a double well, as depicted in Figure 2. It has the global minimum value 0 at two numbers: 0 and 1. To avoid unnecessary technical difficulties we assume that $W(p) = W(1-p)$. The two global minimum points are nondegenerate, i.e., $W''(0) = W''(1) \neq 0$. A simple example is $W(u) = \frac{1}{4}((u - \frac{1}{2})^2 - \frac{1}{4})^2$.

The most mathematically interesting part in equation (1.1) is the nonlocal term $(-\Delta)^{-1/2}(u-a)$ in the integrand. Let $(-\Delta)^{-1}(u-a)$ be the solution v of

$$-\Delta v = u - a \text{ in } D, \quad \partial_\nu v = 0 \text{ on } \partial D, \quad \overline{v} = 0,$$

where $\partial_\nu v$ is the outward normal derivative of v. $(-\Delta)^{-1/2}$ in I is the square root of the positive operator $(-\Delta)^{-1}$ from $\{w \in L^2(D) : \overline{w} = 0\}$ to itself. If we let

[1] See Choksi and Ren [3] for more on these parameters.

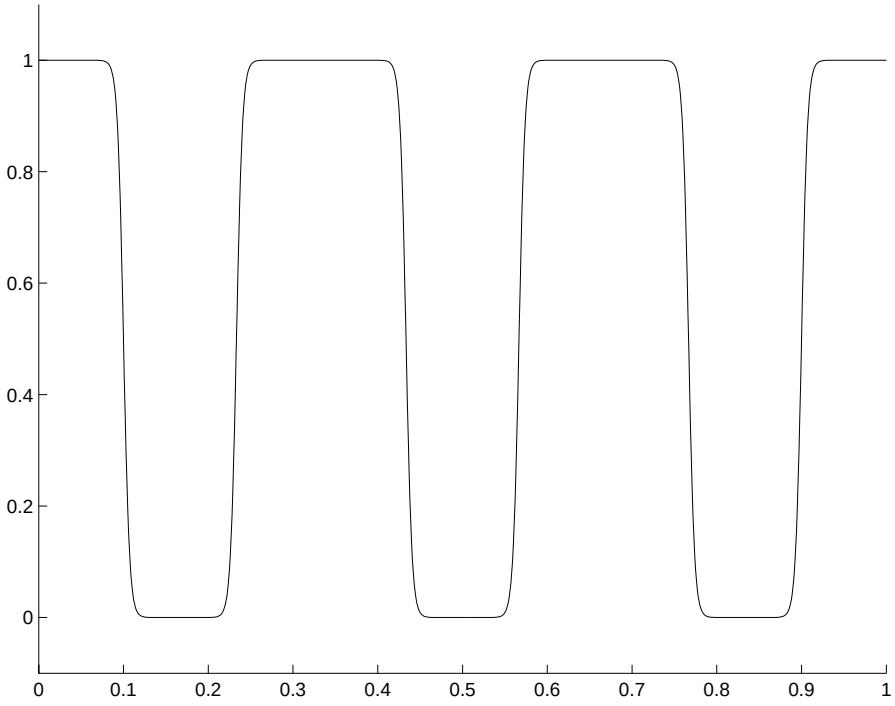


FIG. 3. *A one-dimensional local minimizer u with $K = 6$. The regions where u is close to 1 are microdomains with high concentration of A monomers, and the regions where u is close to 0 are microdomains with high concentration of B monomers.*

$v = (-\Delta)^{-1}(u - a)$, then an often more useful formula is

$$I(u) = \int_D \left\{ \frac{\epsilon^2}{2}|\nabla u|^2 + \frac{\epsilon\gamma}{2}|\nabla v|^2 + W(u) \right\}.$$

Let $f(u) = W'(u)$ as in Figure 2. For the particular $W(u) = \frac{1}{4}((u - \frac{1}{2})^2 - \frac{1}{4})^2$, $f(u) = u(u - \frac{1}{2})(u - 1)$. The Euler–Lagrange equation of I is

$$-\epsilon^2\Delta u + \epsilon\gamma(-\Delta)^{-1}(u - a) + f(u) - \overline{f(u)} = 0, \quad \partial_\nu u = 0 \text{ on } \partial D. \tag{1.2}$$

The term $\overline{f(u)}$ is equal to the Lagrange multiplier corresponding to the constraint $\overline{u} = a$.

It is proved in Ren and Wei [13] using the Γ-limit theory that when $D = (0, 1)$ for any positive integer K there exists a local minimizer u with K interfaces and $K + 1$ microdomains if ϵ is small enough.[2] An example of u with $K = 6$ is shown in Figure 3. u is close to 0 in three regions and close to 1 in four regions. These regions are separated by sharp interfaces. That such u is energetically favored is not too difficult to explain. Note that the W term in I likes to have $u \approx 0$ or $u \approx 1$. The gradient term penalizes oscillation of u, but since it has a small coefficient it tolerates a number of sharp interfaces. The best profile for the nonlocal term is $u \approx a$. But this is impossible due to the presence of the W term and the fact $0 < a < 1$. The second

[2] See Theorem 2.1.

best profile for the nonlocal term is for u to have wild oscillation about a. When all the three terms are present in I, a compromise must be reached, and u as in Figure 3 emerges as a local minimizer.

Now we place such a one-dimensional (1-D) local minimizer in a three-dimensional (3-D) box through trivial extension. The extended u becomes a 3-D critical point of I, i.e., a solution of (1.2). We ask whether this u is a good model of the lamellar phase depicted in plot 3 of Figure 1. In general a morphology phase must be at least *metastable* in the sense that it is described by a local minimizer of I in space. Such a 3-D local minimizer is also called a stable solution of (1.2). We take $D = (0,1)\times(0,1)\times(0,1)$ and study the spectrum of u, i.e., the second variation of I at u. The linearized operator at u is

$$\begin{aligned} &L(\phi) := -\epsilon^2\Delta\phi + \epsilon\gamma(-\Delta)^{-1}\phi + f'(u)\phi - \overline{f'(u)\phi}, \\ &\partial_\nu\phi = 0 \text{ on } \partial D, \quad \phi \in W^{2,2}(D), \quad \overline{\phi} = 0. \end{aligned} \tag{1.3}$$

This is an unbounded self-adjoint operator defined densely on $\{\phi \in L^2(D) : \overline{\phi} = 0\}$ whose spectrum consists of eigenvalues only.

We will obtain detailed information on the spectrum of u when ϵ is small. In particular we will find the asymptotic expansions of the important eigenvalues of small absolute values in terms of ϵ. We will also derive asymptotic expansions of the corresponding eigenfunctions. The analysis in this paper culminates in the following theorem.

THEOREM 1.1. *The eigenvalues λ of L are classified into λ_m by $m = (m_1, m_2)$, which is a pair of nonnegative integers. The following three statements hold when ϵ is sufficiently small:*

1. *There exists $M(K)$, depending on K but not ϵ, so that when $|m| := \sqrt{m_1^2 + m_2^2}$ $\geq M(K)$, $\lambda_m \geq C\epsilon^2$ for some $C > 0$ independent of ϵ.*
2. *When $m = (0,0)$, there are K small positive $\lambda_{(0,0)}$'s. One of them is of order ϵ whose only eigenfunction is approximately $\sum_j (h_j(x) - \overline{h_j})$. The other $K-1$ $\lambda_{(0,0)}$'s are of order ϵ^2. Their only eigenfunctions are approximately $\sum_j c_j^0 h_j(x)$ for some vectors c^0 satisfying $\sum_j c_j^0 = 0$. The remaining $\lambda_{(0,0)}$'s are positive and bounded below by a positive constant independent of ϵ.*
3. *When $m \neq (0,0)$ and $|m| < M(K)$, there are K λ_m's of order ϵ^2, which are not necessarily positive, whose only eigenfunctions are approximately $\sum_j c_j^0 h_j(x)\cos(m_1\pi y_1)\cos(m_2\pi y_2)$. The remaining λ_m's are positive and bounded below by a positive constant independent of ϵ. Only when K is sufficiently large or γ is sufficiently small are all the eigenvalues of L positive and u stable.*

Here a point in D is denoted by (x, y_1, y_2), where x is in the direction perpendicular to the interfaces of a lamellar phase, the up direction in plot 3, Figure 1. The functions h_j are defined in (3.5), and the c^0 vectors are given in sections 5 and 7. The $\lambda_{(0,0)}$ eigenvalues are just the eigenvalues in the 1-D problem. That they are positive, as noted in statement 2, is consistent with the fact that u is a 1-D local minimizer.

The most exciting discovery is apparently statement 3. The presence of the λ_m's there is a 3-D phenomenon. A 1-D local minimizer is *not necessarily* a local minimizer in three dimensions. Not all 1-D local minimizers may be used to model the lamellar phase of diblock copolymers. Only the ones with sufficiently many interfaces, or in other words with sufficiently thin microdomains, are suitable candidates.

Of particular interest is the 1-D global minimizer, which is one of the 1-D local minimizers with $K \approx (\frac{a^2b^2\gamma}{3\tau})^{1/3}$, where τ is a positive number specified in (2.7). Since

its energy is lower than that of any other 1-D local minimizer, it is thermodynamically more preferred. But if it were unstable in three dimensions, then the lamellar phase would only be a transient metastable phase. Thermal fluctuation would eventually destroy any metastable lamellar phase. It turns out that the 1-D global minimizer has a delicate spectral property. It actually lies near the *borderline* of the stability of lamellar solutions.[3]

The stability of a solution of (1.2) may also be defined by a dynamic problem. As observed in [8] one may consider negative gradient flows of I in various function spaces. The simplest one is probably

$$u_t = \epsilon^2 \Delta u - \epsilon\gamma(-\Delta)^{-1}(u-a) - f(u) + \overline{f(u)}, \quad \partial_\nu u = 0 \text{ on } \partial D \times (0,\infty). \tag{1.4}$$

A physically more realistic dynamic model is the Cahn–Hilliard-like [1] fourth order problem:

$$u_t = \Delta(-\epsilon^2\Delta u + \epsilon\gamma(-\Delta)^{-1}(u-a) + f(u)), \quad \partial_\nu \Delta u = \partial_\nu u = 0 \text{ on } \partial D \times (0,\infty). \tag{1.5}$$

The stability of steady states of (1.4) or (1.5) agrees with our static stability definition that a stable solution of (1.2) is a local minimizer of I.

Some preliminary work is done in section 2. We derive inner and outer asymptotic expansions of the lamellar solution u in a rigorous way. The first statement of the theorem is proved in section 3, the second in sections 4 and 5, and the third in sections 6 and 7. In the last section we discuss the spectrum of the 1-D global minimizer.

To avoid clumsy notation a quantity's dependence on ϵ is usually suppressed. For example, we write u, the lamellar solution, instead of u_ϵ. On the other hand we often emphasize a quantity's independence of ϵ with a superscript 0. For example, the limit of a lamellar solution u as $\epsilon \to 0$ is denoted by u^0. In estimates, C is always a positive constant independent of ϵ. Its value may vary from line to line. The shorthand *e.s.* stands for a quantity that is exponentially small, i.e., equals $O(e^{-C/\epsilon})$. The L^2 inner product is denoted by $\langle \cdot, \cdot \rangle$ and the L^p norm by $\| \cdot \|_p$.

References on the mathematical aspects of the block copolymer theory include, in addition to the ones cited already, Ohnishi et al. [9], Choksi [2], Fife and Hilhorst [4], Henry [6], and Ren and Wei [11, 14] on diblock copolymers. On triblock copolymers we refer to Ren and Wei [15, 16].

2. The lamellar solution u. The lamellar solutions we consider in this paper were constructed in [13] by the Γ-limit theory.

THEOREM 2.1 (Ren and Wei [13]). *In* 1-D *for each positive integer K the functional*

$$I_1(u) := \int_0^1 \left\{ \frac{\epsilon^2}{2} \left(\frac{du}{dx}\right)^2 + \frac{\epsilon\gamma}{2} \left| \left(-\frac{d^2}{dx^2}\right)^{-1/2} (u-a) \right|^2 + W(u) \right\} dx \tag{2.1}$$

in $\{u \in W^{1,2}(0,1) : \overline{u} = a\}$ has a local minimizer u near u^0, under the L^2 norm, when ϵ is sufficiently small. It satisfies the Euler–Lagrange equation

$$-\epsilon^2 u'' + f(u) - \overline{f(u)} + \epsilon\gamma G_0[u-a] = 0, \quad u'(0) = u'(1) = 0$$

[3] This phenomenon compares well with the *marginal* stability, observed in Muratov [7], of the corresponding 1-D global minimizer in the Γ-limit.

and the properties

$$\lim_{\epsilon\to 0}\|u-u^0\|_2=0 \quad \textit{and} \quad \lim_{\epsilon\to 0}\epsilon^{-1}I_1(u)=\tau K+\frac{\gamma}{2}\int_0^1|(v^0)'|\,dx.$$

Here u^0 is a step function defined to be

$$u^0(x)=1 \text{ on } (0,x_1^0),\quad 0 \text{ on } (x_1^0,x_2^0),\quad 1 \text{ on } (x_2^0,x_3^0),\quad 0 \text{ on } (x_3^0,x_4^0),\quad 1 \text{ on } (x_4^0,x_5^0),\dots$$

with (recall $b=1-a$)

$$x_1^0=\frac{a}{K},\quad x_2^0=\frac{1+b}{K},\quad x_3^0=\frac{2+a}{K},\quad x_4^0=\frac{3+b}{K},\quad x_5^0=\frac{4+a}{K},\dots,$$

$v^0=G_0[u^0-a]$. G_0 is the solution operator of $-v''=g$, $v'(0)=v'(1)=0$, $\overline{v}=0$. The constant τ is positive and defined in (2.7).

There is another K-interface lamellar solution whose limiting value as $\epsilon\to 0$ is 0 instead of 1 on the first interval $(0,b/K)$. This solution has the same properties as u does, so we focus on u, the solution of the first type, only.

Remark 2.2. This second solution is just $1-\tilde{u}$, where $\tilde{u}$ is a solution of the first type, but with $\overline{\tilde{u}}=1-a$.

In this section we learn more about u. In particular u is periodic.

THEOREM 2.3. *When ϵ is small, for every $x\in(0,1/K)$,*

$$\begin{aligned}u(x)&=u\left(\frac{2}{K}-x\right)=u\left(x+\frac{2}{K}\right)=u\left(\frac{4}{K}-x\right)=u\left(x+\frac{4}{K}\right)=\cdots\\&=\begin{cases}u(1-x) & \textit{if } K \textit{ is even,}\\ u(x+\frac{K-1}{K}) & \textit{if } K \textit{ is odd.}\end{cases}\end{aligned}$$

Moreover when ϵ is small, u is the unique local minimizer of I_1 in an L^2 neighborhood of u^0. If u on $((j-1)/K,j/K)$ for some $j=1,2,\dots,K$ is scaled to a function on $(0,1)$, then it is exactly a one-layer local minimizer of (2.1) *with ϵ and γ replaced by $\tilde{\epsilon}=\epsilon K$ and $\tilde{\gamma}=\gamma/K^3$.*

The nuts and bolts needed to prove this theorem are available in [11]. We give the proof in Appendix A, so the reader may skip it first in order to focus on the spectral properties of u in the following sections.

For that purpose we need asymptotic expansions of u in terms of ϵ. By Lemma A.1 in Appendix A there exist exactly K points x_j, $j=1,2,\dots,K$, in $(0,1)$ so that $u(x_j)=1/2$. These K points identify the interfaces of u. Theorem 2.3 implies that $x_2=\frac{2}{K}-x_1$, $x_3=\frac{4}{K}-x_2$, $x_4=\frac{6}{K}-x_3$, etc. The first approximation of u is

$$\begin{aligned}w(x)&=H\left(-\frac{x-x_1}{\epsilon}\right)+H\left(\frac{x-x_2}{\epsilon}\right)+H\left(-\frac{x-x_3}{\epsilon}\right)-1+\cdots\\&\quad+\begin{cases}H(\frac{x-x_K}{\epsilon}) & \text{if } K \text{ is even,}\\ H(-\frac{x-x_K}{\epsilon})-1 & \text{if } K \text{ is odd.}\end{cases}\end{aligned}\tag{2.2}$$

Here H is the heteroclinic solution of

$$-H''+f(H)=0,\quad H(-\infty)=0,\quad H(\infty)=1,\quad H(0)=1/2.$$

In the case $W(u)=\frac{1}{4}((u-\frac{1}{2})^2-\frac{1}{4})^2$, it is explicitly known that $H(t)=(1/2)(\tanh\frac{t}{2\sqrt{2}}+1)$. $H(t)$ converges to 1 as $t\to\infty$ (and to 0 as $t\to-\infty$) exponentially fast. Also $H'(t)$ and $H''(t)$ decay to 0 exponentially fast as $t\to\pm\infty$. H,

or $H(-\cdot)$, gives the profile of interfaces between the microdomains of u. At every $x \neq x_j^0$, $j = 1, 2, \ldots, K$, $\lim_{\epsilon\to 0} w(x) = u^0(x)$.

Next we define

$$z^0(x) = -\frac{\gamma(v^0(x) - v^0(x_j^0))}{f'(0)}. \tag{2.3}$$

Let us compute

$$(v^0)'(x) = \begin{cases} (a-1)x & \text{on } (0, x_1^0), \\ a(x - \frac{1}{K}) & \text{on } (x_1^0, x_2^0), \\ (a-1)(x - \frac{2}{K}) & \text{on } (x_2^0, x_3^0), \\ a(x - \frac{3}{K}) & \text{on } (x_3^0, x_4^0), \\ (a-1)(x - \frac{4}{K}) & \text{on } (x_4^0, x_5^0), \\ \cdots. \end{cases} \tag{2.4}$$

If we integrate $(v^0)'$ over an interval (x_{j-1}^0, x_j^0), we get 0. So $v^0(x_j^0)$ is independent of j, and the definition of z^0 makes sense. Note that z^0 is independent of ϵ.

LEMMA 2.4. *Let z be defined by $u = w + \epsilon z$. Then $\|z - z^0\|_\infty = O(\epsilon)$.*

Proof. Combine Lemma A.3 in Appendix A and Theorem 2.3. □

LEMMA 2.5. *There exists a constant $C > 0$ independent of ϵ so that $|\epsilon^{-1} z(x_j + \epsilon t)| \le C(1 + |t|)$ for all $t \in (-\frac{x_j}{\epsilon}, \frac{1-x_j}{\epsilon})$. $\epsilon^{-1} z(x_j + \epsilon\cdot)$ converges to P in $C^2_{loc}(-\infty, \infty)$, where $P(t)$ is the solution of*

$$-P'' + f'(H)P = -\gamma(v^0)'(x_j^0)t, \quad P \perp H'$$

in $(-\infty, \infty)$.

There are two different P's depending on whether j is odd or even. But they just differ by a sign, and it is always easy to tell from the context which one is referred to. Once j is given, there exists a unique P since the right side of its equation is perpendicular to the kernel H'.

Proof. Without the loss of generality we assume that j is even. Define $Z(t) = z(x_j + \epsilon t)$. Lemma 2.4 implies $Z = O(1)$ and hence, with the help of Lemma A.4, $\overline{f(u)} = O(\epsilon)$. From the 1-D Euler–Lagrange equation in Theorem 2.1, which u satisfies, we find the equation for Z:

$$-Z'' + f'(H)Z + O(\epsilon)Z^2 + \gamma G_0[u - a] - \epsilon^{-1}\overline{f(u)} = 0.$$

From this equation we also have $Z'' = O(1)$ and $Z' = O(1)$. Multiply the equation by H' and integrate. Set $v(x) = G_0[u - a](x)$. Then

$$\begin{aligned} \text{e.s.} &= \int_{-x_j/\epsilon}^{(1-x_j)/\epsilon} (-Z''H' + f'(H)ZH')\, dt \\ &= \int_{-x_j/\epsilon}^{(1-x_j)/\epsilon} (-O(\epsilon)Z^2 - \gamma v(x_j + \epsilon t) + \epsilon^{-1}\overline{f(u)})H'\, dt \\ &= -\gamma v(x_j) + \epsilon^{-1}\overline{f(u)} + O(\epsilon). \end{aligned}$$

Hence $\gamma v(x_j) - \epsilon^{-1}\overline{f(u)} = O(\epsilon)$ and $\gamma v(x) - \epsilon^{-1}\overline{f(u)} = O(\epsilon) + O(\epsilon)t$. The equation for Z is now simplified to

$$-Z'' + f'(H)Z + O(\epsilon) + O(\epsilon)t = 0.$$

As $\epsilon \to 0$, $Z \to cH'$ in $C^2_{loc}(-\infty,\infty)$ for some c. But $Z(0) = $e.s. implies $c = 0$ since $H'(0) \neq 0$. Therefore $Z \to 0$ in $C^2_{loc}(-\infty,\infty)$.

Next we study $\epsilon^{-1}Z$, whose equation is written as

$$-(\epsilon^{-1}Z)'' + f'(H)(\epsilon^{-1}Z) + O(1)Z^2 + \gamma\epsilon^{-1}v - \epsilon^{-2}\overline{f(u)} = 0.$$

We again multiply by H' and integrate:

$$\begin{aligned}
\text{e.s.} &= \int_{-x_j/\epsilon}^{(1-x_j)/\epsilon} (-\epsilon^{-1}Z''H' + f'(H)\epsilon^{-1}ZH')\,dt \\
&= \int_{-x_j/\epsilon}^{(1-x_j)/\epsilon} (-O(1)Z^2 - \gamma\epsilon^{-1}v(x_j+\epsilon t) + \epsilon^{-2}\overline{f(u)})H'\,dt \\
&= \int_{-x_j/\epsilon}^{(1-x_j)/\epsilon} (-O(1)Z^2 - \gamma\epsilon^{-1}v(x_j) - \gamma v'(x_j)t + O(\epsilon)t^2 + \epsilon^{-2}\overline{f(u)})H'\,dt \\
&= -\gamma\epsilon^{-1}v(x_j) + \epsilon^{-2}\overline{f(u)} + o(1),
\end{aligned}$$

where we have used the facts that $Z \to 0$ locally and $\int_{-x_j/\epsilon}^{(1-x_j)/\epsilon} tH'\,dt = $e.s. Hence $\gamma\epsilon^{-1}v(x_j) - \epsilon^{-2}\overline{f(u)} = o(1)$, which simplifies the equation for $\epsilon^{-1}Z$ to

$$-(\epsilon^{-1}Z)'' + f'(H)(\epsilon^{-1}Z) + O(1)Z^2 + \gamma v'(x_j)t + O(\epsilon)t^2 + o(1) = 0. \tag{2.5}$$

Next we show that $|\epsilon^{-1}Z(t)| \le C(1+|t|)$. Without the loss of generality we consider $t > 0$. Let $\epsilon^{-1}Z(t) = (1+t)R(t)$, where R satisfies

$$-R'' - \frac{2R'}{1+t} + f'(H)R + O(1) = 0 \ \text{ in } \left(0, \frac{1-x_j}{\epsilon}\right), \quad R(0) = \text{e.s.}, \quad R\left(\frac{1-x_j}{\epsilon}\right) = O(1).$$

Suppose that $R = O(1)$ is invalid. We let $\hat{R} = R/\|R\|_{L^\infty}$, which satisfies

$$-\hat{R}'' - \frac{2\hat{R}'}{1+t} + f'(H)\hat{R} + o(1) = 0, \quad \hat{R}(0) = \text{e.s.}, \quad \hat{R}\left(\frac{1-x_j}{\epsilon}\right) = o(1).$$

From this equation we see that $|\hat{R}|$ must attain its maximum value 1 in a bounded region around 0. In the limit $\hat{R}$ approaches in $C^2_{loc}[0,\infty)$ to a nonzero, bounded solution of

$$-\hat{R}_\infty'' - \frac{2\hat{R}_\infty}{1+t} + f'(H)\hat{R}_\infty = 0 \ \text{ in } (0,\infty), \quad \hat{R}_\infty(0) = 0.$$

Then $(1+t)\hat{R}_\infty$ satisfies

$$-((1+t)\hat{R}_\infty)'' + f'(H)(1+t)\hat{R}_\infty = 0 \ \text{ in } (0,\infty).$$

Thus, $(1+t)\hat{R}_\infty(t) = cH'(t)$ for some c. This is because $|(1+t)\hat{R}_\infty(t)|$ grows at most like t, and any other solution, independent of H', of the last equation grows exponentially fast. Since $(1+0)\hat{R}_\infty(0) = 0$ and $H'(0) \neq 0$, we derive $c = 0$ and $\hat{R}_\infty = 0$, a contradiction.

Since $|\epsilon^{-1}Z(t)| \le C(1+|t|)$, we may send $\epsilon \to 0$ in (2.5) and find that $\epsilon^{-1}Z$ approaches in $C^2_{loc}(-\infty,\infty)$ to a solution of

$$-P'' + f'(H)P = -\gamma(v^0)'(x_j^0)t \ \text{ in } (-\infty,\infty).$$

We write the solution family as $P + cH'$ with $P \perp H'$. Here $P(0) = 0$, and $P(0) + cH'(0) = cH'(0)$, where $H'(0) \neq 0$. Since $\epsilon^{-1}Z(0) =$ e.s., we must have $c = 0$, and $\epsilon^{-1}Z \to P$ in $C^2_{loc}(-\infty, \infty)$. □

In the language of singular perturbation theory, the last two lemmas assert that the outer expansion of u is $u^0 + \epsilon z^0 + \cdots$ and the inner expansion at x_j (when j is even) is $H + \epsilon^2 P + \cdots$. The fact that $z^0(x_j^0) = 0$ matches the absence of the ϵ order term in the inner expansion. The function w defined in (2.2) is the 0th order uniform approximation of u.

We close this section by defining two frequently used constants. The first one is

$$s := \int_{-\infty}^{\infty} f''(H(t))(H'(t))^2 P(t)\, dt = -\frac{\gamma ab}{K}. \tag{2.6}$$

Here P is associated with an even j. When P is associated with an odd j in this paper, $f''(H(t))$ will always be changed to $f''(H(-t)) = -f''(H(t))$, so s remains the same. To verify the equality in (2.6) we differentiate the equation for P, multiply by H', and integrate. The right side becomes $-\gamma(v^0)'(x_2^0)$. The left side becomes

$$\int_{-\infty}^{\infty} (-P'''H' + f'(H)P'H' + f''(H)(H')^2 P)\, dt = \int_{-\infty}^{\infty} f''(H)(H')^2 P\, dt,$$

where the first two terms on the left side cancel after integration by parts and using $-H''' + f'(H)H' = 0$, which follows after differentiating the equation for H. From (2.4) we find $(v^0)'(x_2^0) = ab/K$ and $s = -\gamma ab/K$.

The second constant is

$$\tau := \int_{-\infty}^{\infty} (H'(t))^2\, dt > 0. \tag{2.7}$$

Because the equation for H has a first integral $-\frac{(H')^2}{2} + W(H) = 0$, then $\tau = \int_{-\infty}^{\infty} \sqrt{2W(H(t))} H'(t)\, dt = \int_0^1 \sqrt{2W(p)}\, dp$.[4] In the special case $W(u) = \frac{1}{4}((u - \frac{1}{2})^2 - \frac{1}{4})^2$, $\tau = \frac{\sqrt{2}}{12}$.

3. Linearization at u. The 1-D local minimizer u of I_1 is now viewed as a function on D through extension to the second and third dimensions trivially, so $u(x, y_1, y_2) = u(x)$. It is a solution of (1.2) and $I_1(u) = I(u)$.

For an eigenpair (λ, ϕ) of (1.3) we separate variables so that

$$\phi(x, y_1, y_2) = \sum_{m_1, m_2 = 0}^{\infty} \phi_m(x) \cos(m_1 \pi y_1) \cos(m_2 \pi y_2). \tag{3.1}$$

We set $m = (m_1, m_2)$ and let $m^2 = m_1^2 + m_2^2$. Note that

$$(-\Delta)^{-1}\{\phi_m(x) \cos(m_1 \pi y_1) \cos(m_2 \pi y_2)\} = X(x) \cos(m_1 \pi y_1) \cos(m_2 \pi y_2),$$

where X is the solution of

$$-X'' = \phi_{(0,0)}, \quad X'(0) = X'(1) = 0, \quad \overline{X} = 0 \quad \text{if } m = (0, 0)$$

or

$$-X'' + m^2 \pi^2 X = \phi_m, \quad X'(0) = X'(1) = 0 \quad \text{if } m \neq (0, 0).$$

[4] In [13, 14, 16] this constant is defined by the last integral.

The solution operator of the first equation is G_0, already defined. Let $G_m[\cdot]$ be the solution operator of the second equation. They are identified with the Green functions $G_m(\cdot,\cdot)$ in this paper. Therefore $X = G_m[\phi_m]$. The eigenvalue problem $L\phi = \lambda\phi$ now becomes

$$\sum_m \{-\epsilon^2(\phi_m'' - m^2\pi^2\phi_m) + \epsilon\gamma G_m[\phi_m] + f'(u)\phi_m\}\cos(m_1\pi y_1)\cos(m_2\pi y_2) - \overline{f'(u)\phi_0}$$
$$= \lambda\sum_m \phi_m(x)\cos(m_1\pi y_1)\cos(m_2\pi y_2).$$

Here we have used the fact that $\overline{f'(u)\phi_m(x)\cos(m_1\pi y_1)\cos(m_2\pi y_2)} = 0$ if $m \neq (0,0)$.

Multiplying the equation by $\cos(m_1\pi y_1)\cos(m_2\pi y_2)$ and integrating with respect to y_1 and y_2, we find two cases:

1. When $m = (0,0)$,

$$-\epsilon^2\phi_{(0,0)}'' + \epsilon\gamma G_0[\phi_{(0,0)}] + f'(u)\phi_{(0,0)} - \overline{f'(u)\phi_{(0,0)}} = \lambda\phi_{(0,0)},$$
$$\phi_{(0,0)}'(0) = \phi_{(0,0)}'(1) = \overline{\phi_{(0,0)}} = 0. \tag{3.2}$$

2. When $m \neq (0,0)$,

$$-\epsilon^2(\phi_m'' - m^2\pi^2\phi_m) + \epsilon\gamma G_m[\phi_m] + f'(u)\phi_m = \lambda\phi_m,$$
$$\phi_m'(0) = \phi_m'(1) = 0. \tag{3.3}$$

Because the λ's are classified by m, we use λ_m to denote an eigenvalue that is associated with m. The corresponding eigenfunction is $\phi_m(x)\cos(m_1\pi y_1)\cos(m_2\pi y_2)$.

Proof of Theorem 1.1, *statement* 1. We first consider the local eigenvalue problem

$$E(\phi) := -\epsilon^2\phi'' + f'(u)\phi = \nu\phi, \quad \phi'(0) = \phi'(1) = 0. \tag{3.4}$$

In this proof an eigenpair of (3.4) is denoted by (ν,ϕ). We will prove that $\nu \geq -C\epsilon^2$ for some $C > 0$.

Claim 1. If $\nu \to \nu^0$ as $\epsilon \to 0$, then $\nu^0 \geq 0$.

Suppose on the contrary that $\nu^0 < 0$. Let $y \in [0,1]$ so that $\phi(y) = \max|\phi| = 1$. Then $y - x_j = O(\epsilon)$ for some j. Otherwise $-\epsilon^2\phi''(y) \geq 0$, $f'(u(y))\phi(y) > 0$, $\nu\phi(y) < 0$, and hence (3.4) is not satisfied. Then we consider $\Phi(t) = \phi(x_j + \epsilon t)$, which satisfies $-\Phi'' + f'(u)\Phi = \nu\Phi$ in $(-x_j/\epsilon, (1-x_j)/\epsilon)$. As $\epsilon \to 0$, Φ approaches $\Phi_\infty \not\equiv 0$ in $C^2_{loc}(-\infty,\infty)$, which satisfies $-\Phi_\infty'' + f'(H)\Phi_\infty = \nu^0\Phi$ in $(-\infty,\infty)$. But this is impossible since the last equation has no negative eigenvalues. This proves the claim.

The case $\nu^0 > 0$ does not concern us, so we assume $\nu \to 0$. We introduce, for $j = 1, 2, \ldots, K$,

$$h_j(x) = H'\left(\frac{x - x_j}{\epsilon}\right) + \text{e.s.} \tag{3.5}$$

Here e.s. is an exponentially small correction term. It is chosen so that $h_j(0) = h_j(1) = h_j'(0) = h_j'(1) = 0$, $\|h_j' - \epsilon^{-1}H''(\frac{\cdot - x_j}{\epsilon})\|_\infty = \text{e.s.}$, and $\|h_j'' - \epsilon^{-2}H'''(\frac{\cdot - x_j}{\epsilon})\|_\infty = \text{e.s.}$

Remark 3.1. Should we weaken the condition $W(p) = W(1-p)$, H' would no longer be even and we would set

$$h_j(x) = \begin{cases} H'(\frac{x-x_j}{\epsilon}) + \text{e.s.} & \text{if } j \text{ is even,} \\ H'(-\frac{x-x_j}{\epsilon}) + \text{e.s.} & \text{if } j \text{ is odd.} \end{cases}$$

Consider the subspace of $L^2(0,1)$ generated by h_j. Decompose $\phi = \sum_j c_j h_j + \psi$, so that $h_j \perp \psi$ for each $j = 1, 2, \ldots, K$. Note that

$$E(h_j) = (f'(u) - f'(H))h_j + \text{e.s.},$$

and by Lemma 2.5,

$$\begin{aligned} |(f'(u) - f'(H))h_j| &= |(f'(w(x_j + \epsilon t) + \epsilon z(x_j + \epsilon t)) - f'(H(t)))H'(t)| + \text{e.s.} \\ &= |f''(H(t))\epsilon z(x_j + \epsilon t)H'(t)| + O(\epsilon^4) \ = \ O(\epsilon^2). \end{aligned} \tag{3.6}$$

Hence we deduce

$$E(h_j) = O(\epsilon^2). \tag{3.7}$$

We write (3.4) as

$$\sum_{j=1}^{K} c_j E(h_j) + E(\psi) = \nu \sum_j c_j h_j + \nu\psi. \tag{3.8}$$

Claim 2. $\langle E(\psi), \psi\rangle \geq C\|\psi\|_2^2$ for some $C > 0$ independent of ϵ.

When we minimize the quotient $\frac{\langle E(\tilde\psi), \tilde\psi\rangle}{\|\tilde\psi\|_2^2}$ among nonzero $\tilde\psi$ subject to $\tilde\psi \perp h_j$ for every j, the minimizer, denoted by $\tilde\psi$ in this paragraph, satisfies

$$-\epsilon^2\tilde\psi'' + f'(u)\tilde\psi = \iota\tilde\psi + \sum_j d_j h_j. \tag{3.9}$$

The constant $\iota = \frac{\langle E(\tilde\psi), \tilde\psi\rangle}{\|\tilde\psi\|_2^2}$. Suppose that Claim 2 is false. Then $\lim_{\epsilon\to 0} \iota = \iota^0 \leq 0$. We multiply $\tilde\psi$ by a proper constant so there exists $y \in [0,1]$ such that $\tilde\psi(y) = \max|\tilde\psi| = 1$. Now we multiply (3.9) by h_k and integrate.

$$\langle E(h_k), \tilde\psi\rangle = \sum_j d_j \langle h_j, h_k\rangle.$$

The left side is $O(\epsilon^2)$ by (3.7). The right side is

$$\int_0^1 \sum_j d_j h_j h_k = \sum_j \epsilon d_j \tau \delta_{jk} + \text{e.s.}|d| = \epsilon\tau d_k + \text{e.s.}|d|,$$

where $\delta_{jk} = 1$ if $j = k$ and 0 otherwise, and $|d| = \sqrt{d_1^2 + d_2^2 + \cdots + d_K^2}$ is the norm of the vector d. Therefore $d_k = O(\epsilon)$. As in the proof of Claim 1, $y - x_j = O(\epsilon)$ for some j. Moreover we consider $\Psi(t) = \tilde\psi(x_j + \epsilon t)$, which satisfies $-\Psi'' + f'(u)\Psi = \iota\Psi + o(1)$. Passing to the limit we find a nonzero Ψ_∞ which satisfies $-\Psi_\infty'' + f'(H)\Psi_\infty = \iota^0\Psi_\infty$ in $(-\infty, \infty)$. Therefore $\iota^0 = 0$ and Ψ_∞ is proportional to H'. But on the other hand $\tilde\psi \perp h_j$ implies $\Psi_\infty \perp H'$. Hence $\Psi_\infty = 0$, contradicting the fact that Ψ_∞ is nonzero. This proves Claim 2.

We now return to (3.8). Multiply it by ψ and integrate. Use (3.7) to deduce

$$|c|O(\epsilon^2)\|\psi\|_2 + \langle E(\psi), \psi\rangle = \nu \int_0^1 \psi^2.$$

Then Claim 2 implies

$$\|\psi\|_2 = O(\epsilon^2)|c|. \tag{3.10}$$

Next we multiply (3.8) by h_k and integrate. The left side is

$$\begin{aligned}&\int_0^1 \left\{ E(h_k)\psi + \sum_j c_j E(h_j)h_k \right\} \\ &= \int_0^1 \left\{ (f'(u) - f'(H))h_k\psi + \text{e.s.}\psi + \sum_j c_j((f'(u) - f'(H))h_j h_k + \text{e.s.}) \right\},\end{aligned} \tag{3.11}$$

in which

$$\left| \int_0^1 (f'(u) - f'(H))h_k\psi \right| \le \|(f'(u) - f'(H))h_k\|_\infty \|\psi\|_2 = O(\epsilon^4)|c|$$

by (3.6) and (3.10), and

$$\begin{aligned}&\int_0^1 (f'(u) - f'(H))h_j h_k \\ &= \epsilon \int_{-x_j/\epsilon}^{(1-x_j)/\epsilon} \{f'(w(x_j + \epsilon t) + \epsilon z(x_j + \epsilon t)) - f'(H(t))\} H'(t)H'(t + (x_j - x_k)/\epsilon)\, dt + \text{e.s.} \\ &= \epsilon \int_{-x_j/\epsilon}^{(1-x_j)/\epsilon} \{f''(H(t))\epsilon z(x_j + \epsilon t) + O(\epsilon^2)z^2(x_j + \epsilon t)\} H'(t)H'(t + (x_j - x_k)/\epsilon)\, dt + \text{e.s.} \\ &= \epsilon^3 \int_{-x_j/\epsilon}^{(1-x_j)/\epsilon} f''(H(t))P(t)H'(t)H'(t + (x_j - x_k)/\epsilon)\, dt + o(\epsilon^3) \\ &= \epsilon^3 s\delta_{jk} + o(\epsilon^3)\end{aligned} \tag{3.12}$$

by Lemma 2.5. The above argument applies to the case when j is even. When j is odd, $f''(H(t))$ becomes $f''(H(-t)) = -f''(H(t))$ and $P(t)$ has a different sign, but the final result remains unchanged. Hence (3.11) becomes $\epsilon^3 s c_k + o(\epsilon^3)|c|$. The right side of (3.8) multiplied by h_k and integrated is $\nu\epsilon\tau c_k + \text{e.s.}|c|$. Equating the last two quantities, we find that for every k

$$sc_k + o(1)|c| = \frac{\nu\tau}{\epsilon^2} c_k.$$

Therefore $\nu \ge -C\epsilon^2$ for some $C > 0$ independent of ϵ.

Since G_m is a bounded, positive operator in the eigenvalue problem

$$-\epsilon^2\phi'' + \epsilon\gamma G_m[\phi] + f'(u)\phi = \nu\phi, \quad \phi'(0) = \phi'(1) = 0, \tag{3.13}$$

we again have $\nu \ge -C\epsilon^2$ for some $C > 0$ independent of ϵ. This can be seen easily by comparing the variational characterization of the principle eigenvalue of (3.13),

$$\inf \left\{ \int_0^1 \{\epsilon^2(\phi')^2 + \epsilon\gamma G_m[\phi]\phi + f'(u)\phi^2\}\, dx : \phi \in W^{1,2}(0,1),\ \|\phi\|_2 = 1 \right\},$$

to a similar one without the $\epsilon\gamma G_m[\phi]\phi$ term for (3.4). Finally, in (3.3), by setting m^2 large enough, we find $\lambda_m \ge C\epsilon^2$. □

4. $m = (0,0)$ eigenvalues. Here we study the $m_1 = m_2 = 0$ problem (3.2). Denote the linear operator there by L_0. An eigenpair of (3.2) is denoted by (λ, ϕ) in this section. Since (3.2) is precisely the linearized operator of the 1-D problem I_1 defined in (2.1) at a 1-D local minimizer u, we have $\lambda \geq 0$. The case $\lambda \to \lambda^0 > 0$ as $\epsilon \to 0$ does not concern us. So we assume $\lambda \to 0$ along a subsequence throughout the rest of this section.

We decompose $\phi = \sum_j c_j(h_j - \overline{h_j}) + \psi$, where $\psi \perp h_j - \overline{h_j}$ for every j. Note that

$$L_0(h_j - \overline{h_j}) = (f'(u) - f'(H))h_j + \epsilon\gamma G_0[h_j - \overline{h_j}] + (\overline{f'(u)} - f'(u))\overline{h_j} - \overline{f'(u)h_j} + \text{e.s.}$$

A few terms on the right side are estimated once and for all.

$$\epsilon\gamma G_0[h_j - \overline{h_j}](x) = \gamma\epsilon^2 G_0[(h_j - \overline{h_j})/\epsilon](x) = \gamma\epsilon^2 G_0(x, x_j) + O(\epsilon^3). \tag{4.1}$$

$$\begin{aligned}
\overline{f'(u)h_j} &= \epsilon \int_{-x_j/\epsilon}^{(1-x_j)/\epsilon} f'(w(x_j + \epsilon t) + \epsilon z(x_j + \epsilon t))H'(t)\, dt + \text{e.s.} \\
&= \epsilon \int_{-x_j/\epsilon}^{(1-x_j)/\epsilon} (f'(H(t)) + f''(H(t))\epsilon z(x_j + t) + O(\epsilon^4))H'(t)\, dt + \text{e.s.} \\
&= \epsilon^2 \int_{-\infty}^{\infty} f''(H(t))z(x_j + \epsilon t)H'(t)\, dt + O(\epsilon^5) \;=\; O(\epsilon^3),
\end{aligned} \tag{4.2}$$

where the last line follows from Lemma 2.5. The next estimate is not the sharpest.

$$|(\overline{f'(u)} - f'(u))\overline{h_j}| = |(\overline{f'(u)} - f'(u))|(\epsilon + \text{e.s.}) = O(\epsilon). \tag{4.3}$$

So based on the last three estimates and (3.6) we find

$$L_0(h_j - \overline{h_j}) = O(\epsilon). \tag{4.4}$$

We also need an L^1 version of (4.3):

$$\|(\overline{f'(u)} - f'(u))\overline{h_j}\|_1 = O(\epsilon)\|\overline{f'(u)} - f'(u)\|_1 = O(\epsilon)\|\overline{f'(w)} - f'(w) + O(\epsilon)\|_1 = O(\epsilon^2),$$

so we obtain

$$\|L_0(h_j - \overline{h_j})\|_1 = O(\epsilon^2). \tag{4.5}$$

Rewrite (3.2) as

$$\sum_j c_j L_0(h_j - \overline{h_j}) + L_0\psi = \lambda \sum_j c_j(h_j - \overline{h_j}) + \lambda\psi. \tag{4.6}$$

LEMMA 4.1. *$\langle L_0(\psi), \psi\rangle \geq C\|\psi\|_2^2$ for some $C > 0$ independent of ϵ.*

Proof. When we minimize the quotient $\frac{\langle L_0(\tilde\psi), \tilde\psi\rangle}{\|\tilde\psi\|_2^2}$ among nonzero $\tilde\psi$ of zero average subject to $\tilde\psi \perp h_j - \overline{h_j}$ for every j, the minimizer, denoted by $\tilde\psi$ in this proof, satisfies

$$-\epsilon^2\tilde\psi'' + \epsilon\gamma G_0[\tilde\psi] + f'(u)\tilde\psi - \overline{f'(u)\tilde\psi} = \iota\tilde\psi + \sum_j d_j(h_j - \overline{h_j}).$$

The constant $\iota = \frac{\langle L_0(\tilde\psi), \tilde\psi\rangle}{\|\tilde\psi\|_2^2}$. Suppose the lemma is false. Then $\lim_{\epsilon\to 0} \iota = \iota^0 \leq 0$. We multiply $\tilde\psi$ by a proper constant so there exists $y \in [0,1]$ such that $\tilde\psi(y) = \max|\tilde\psi| = 1$.

Now we multiply the last equation by $h_k - \overline{h_k}$ and integrate: $\langle L_0(h_k - \overline{h_k}), \tilde{\psi}\rangle = \sum_j d_j \langle h_j - \overline{h_j}, h_k - \overline{h_k}\rangle$. The left side is $O(\epsilon^2)$ by (4.5). The right side is

$$\int_0^1 \sum_j d_j (h_j - \overline{h_j})(h_k - \overline{h_k}) = \epsilon\tau d_k + O(\epsilon^2)|d|.$$

Therefore $d_k = O(\epsilon)$. The rest of the proof is the same as that of Claim 2 in section 3, since the additional terms in the equation satisfy

$$\epsilon\gamma G_0[\tilde{\psi}] = O(\epsilon),$$
$$\overline{f'(u)\tilde{\psi}} = \overline{(f'(u) - f'(0))\tilde{\psi}} \ = \ O(1)\|f'(u) - f'(0)\|_1 \ = \ O(\epsilon).$$

A minor difference is that $\tilde{\psi} \perp h_j$ here is a consequence of $\tilde{\psi} \perp h_j - \overline{h_j}$ and $\overline{\tilde{\psi}} = 0$. □

Multiply (4.6) by ψ and integrate. Using (4.4) we find

$$|c|O(\epsilon)\|\psi\|_2 + \langle L_0(\psi), \psi\rangle = \lambda\|\psi\|_2^2.$$

Lemma 4.1 implies that

$$\|\psi\|_2 = O(\epsilon)|c|. \tag{4.7}$$

Remark 4.2. As a comparison we compute

$$\left\|\sum_j c_j(h_j - \overline{h_j})\right\|_2 = \left\{\sum_j c_j^2 \int_0^1 h_j^2 + O(\epsilon^2)|c|^2\right\}^{1/2} = \{\epsilon\tau|c|^2 + O(\epsilon^2)|c|^2\}^{1/2} \sim \epsilon^{1/2}|c|.$$

So in the decomposition of ϕ, $\sum_j c_j(h_j - \overline{h_j})$ is more prominent than ψ.

Multiply (4.6) by $h_k - \overline{h_k}$ and integrate:

$$\int_0^1 L_0\psi(h_k - \overline{h_k}) + \sum_j c_j \int_0^1 L_0(h_j - \overline{h_j})(h_k - \overline{h_k}) = \lambda \sum_j c_j \int_0^1 (h_j - \overline{h_j})(h_k - \overline{h_k}). \tag{4.8}$$

The first term on the left side is written as

$$\int_0^1 L_0(\psi)(h_k - \overline{h_k}) \ = \ \int_0^1 L_0(h_k - \overline{h_k})\psi$$
$$= \int_0^1 \{(f'(u) - f'(H))h_k\psi + \epsilon\gamma G_0[h_k - \overline{h_k}]\psi + (\overline{f'(u)} - f'(u))\overline{h_k}\psi - \overline{f'(u)h_k}\psi + \text{e.s. } \psi\}$$
$$= \int_0^1 \{(f'(u) - f'(H))h_k\psi + \epsilon\gamma G_0[\psi]h_k + (f'(0) - f'(u))\overline{h_k}\psi + \text{e.s. } \psi\}. \tag{4.9}$$

The four terms are estimated as follows:

$$\left|\int_0^1 (f'(u) - f'(H))h_k\psi\right| \le \|(f'(u) - f'(H))h_k\|_\infty\|\psi\|_2 \ = \ O(\epsilon^2)\|\psi\|_2 \ = \ O(\epsilon^3)|c|,$$
$$\int_0^1 \epsilon\gamma G_0[\psi]h_k = O(\epsilon)\|G_0[\psi]\|_\infty\|h_k\|_1 \ = \ O(\epsilon^2)\|\psi\|_2 \ = \ O(\epsilon^3)|c|,$$
$$\int_0^1 (f'(0) - f'(u))\overline{h_k}\psi = \|f'(u) - f'(0)\|_2 O(\epsilon)\|\psi\|_2 \ = \ O(\epsilon^{2.5})|c|,$$
$$\int_0^1 \text{e.s. } \psi = \text{e.s.}\|\psi\|_2 \ = \ \text{e.s.}|c|.$$

Note that the first estimate follows from (3.6). The second term on the left of (4.8) is, for each j, by (4.1) and (4.2),

$$\begin{aligned}
&\int_0^1 L_0(h_j-\overline{h_j})(h_k-\overline{h_k}) \;=\; \int_0^1 L_0(h_j-\overline{h_j})h_k \\
&= \int_0^1 \{(f'(u)-f'(H))h_jh_k + \epsilon\gamma G_0[h_j-\overline{h_j}]h_k + (\overline{f'(u)}-f'(u))\overline{h_j}h_k - \overline{f'(u)h_j}h_k + \text{e.s.}\} \\
&= \epsilon^3 s\delta_{jk} + \gamma\epsilon^3 G_0(x_j,x_k) + \epsilon^2\overline{f'(u)} + o(\epsilon^3).
\end{aligned} \tag{4.10}$$

The last line follows from the estimates (3.12), (4.1), (4.2), and

$$\begin{aligned}
\int_0^1 (\overline{f'(u)}-f'(u))\overline{h_j}h_k &= (\epsilon+\text{e.s.})\left\{\overline{f'(u)}(\epsilon+\text{e.s.}) - \int_0^1 f'(u)h_k\right\} \\
&= \epsilon^2\overline{f'(u)} - \epsilon\int_0^1 f'(u)h_k + \text{e.s.} \\
&= \epsilon^2\overline{f'(u)} - \epsilon^2\int_{-x_k/\epsilon}^{(1-x_k)/\epsilon} (f'(H)+O(\epsilon^2))H'(t)\,dt + \text{e.s.} = \epsilon^2\overline{f'(u)} + O(\epsilon^4).
\end{aligned}$$

The right side of (4.8) is

$$\lambda\sum_j c_j\int_0^1 (h_j-\overline{h_j})(h_k-\overline{h_k}) = \lambda(\epsilon\tau c_k + O(\epsilon^2)|c|). \tag{4.11}$$

In summary, for every k

$$\epsilon^3 sc_k + \sum_j \{\gamma\epsilon^3 G(x_j,x_k) + \epsilon^2\overline{f'(u)} + o(\epsilon^3)\}c_j + O(\epsilon^{2.5})|c| = \lambda(\tau\epsilon c_k + O(\epsilon^2)|c|). \tag{4.12}$$

If we consider the c_k of the largest absolute value, since $\epsilon^2\overline{f'(u)} \sim \epsilon^2$, $\lambda = O(\epsilon)$. On the left side of (4.12), $\epsilon^2\overline{f'(u)}$ is the largest term. Because $\epsilon^2\overline{f'(u)}$ is multiplied by $\sum_j c_j$, a dichotomy appears at this point, unless $K = 1$.

Case 1. $\frac{\sum_k c_k}{|c|} \not\to 0$. Note that when $K = 1$, this is the only case. We rewrite (4.12) as

$$\overline{f'(u)}\sum_j c_j + O(\epsilon^{1/2})|c| = \frac{\lambda\tau}{\epsilon}c_k. \tag{4.13}$$

In the limit we have

$$f'(0)\sum_j c_j^0 = \eta\tau c_k^0, \qquad \sum_j c_j^0 \neq 0, \tag{4.14}$$

since $\lim_{\epsilon\to 0}\overline{f'(u)} = f'(0)$. Here $\eta = \lim_{\epsilon\to 0}\lambda/\epsilon$ and $\lim_{\epsilon\to 0} c_j = c_j^0$. Solving (4.14) we find

$$\eta = \frac{f'(0)K}{\tau}, \quad c_1^0 = c_2^0 = \cdots = c_K^0. \tag{4.15}$$

Thus we obtain the asymptotic expansions for one eigenpair $\lambda_{(0,0)} = \lambda$ and $\phi_{(0,0)} = \phi$ of (3.2):

$$\lambda_{(0,0)} = \frac{\epsilon f'(0)K}{\tau} + o(\epsilon), \quad \phi_{(0,0)} \approx \sum_j (h_j - \overline{h_j}). \tag{4.16}$$

Note that this $\lambda_{(0,0)}$ is positive.

Case 2. $\frac{\sum_k c_k}{|c|} \to 0$. This occurs when $K \geq 2$. To study this case, we rewrite $L_0(\phi) = \lambda\phi$ as

$$L_0(\psi) = -\sum_j c_j L_0(h_j - \overline{h_j}) + \lambda \sum_j c_j (h_j - \overline{h_j}) + \lambda\psi.$$

Note that

$$\sum_j c_j L_0(h_j - \overline{h_j}) = O(\epsilon^2)|c| + \sum_j c_j (\overline{f'(u)} - f'(u))\overline{h_j} = O(\epsilon^2)|c| + \left(\sum_j c_j\right) O(\epsilon) = o(\epsilon)|c|$$

by the assumption and by (4.1), (4.2), (3.6), and (4.3). Hence

$$L_0(\psi) = o(\epsilon)|c| + |\lambda| O(1)|c| + \lambda\psi. \tag{4.17}$$

LEMMA 4.3. $\|\psi\|_\infty = o(\epsilon)|c| + |\lambda| O(1)|c|$.

Proof. Suppose that the lemma is false. Replacing ψ by $\pm\frac{\psi}{\|\psi\|_\infty}$ in (4.17) we obtain $L_0(\psi) = o(1) + \lambda\psi$, where at some $y \in [0,1]$, $\psi(y) = \|\psi\|_\infty = 1$. We show that $y - x_j = O(\epsilon)$ for some j. Otherwise $-\epsilon^2\psi''(y) \geq 0$, $\epsilon\gamma G_0[\psi] = O(\epsilon)$, $f'(u)\psi(y) \to f'(0) > 0$, $-\overline{f'(u)\psi} = O(\epsilon)$, and $\lambda\psi(y) = O(\epsilon)$. Hence the equation $L_0(\psi) = o(1) + \lambda\psi$ is not satisfied. Then we set $\Psi(t) = \psi(x_j + \epsilon t)$, which satisfies $-\Psi'' + f'(u)\Psi = o(1)$ in $(-x_j/\epsilon, (1 - x_j)/\epsilon)$. As $\epsilon \to 0$, $\Psi \to \Psi_\infty \not\equiv 0$ in $C^2_{loc}(-\infty, \infty)$ and Ψ_∞ satisfies $-\Psi''_\infty + f'(H)\Psi_\infty = 0$. Hence Ψ_∞ is proportional to H'. On the other hand $\psi \perp h_j$ implies $\int_{-\infty}^{\infty} \Psi_\infty H' = 0$. Thus $\Psi_\infty = 0$, contradicting the fact that $\Psi_\infty \not\equiv 0$. □

With this lemma we return to (4.8) and recall

$$\int_0^1 (f'(u) - f'(H))h_k\psi = O(\epsilon^2)\|\psi\|_2, \quad \int_0^1 \epsilon\gamma G_0[\psi]h_k = O(\epsilon^2)\|\psi\|_2.$$

Rederive

$$\int_0^1 (f'(0) - f'(u))\overline{h_k}\psi = \|f'(u) - f'(0)\|_1 O(\epsilon)\|\psi\|_\infty \ = \ O(\epsilon^2)\|\psi\|_\infty.$$

Therefore

$$\int_0^1 L_0(\psi)(h_k - \overline{h_k}) = o(\epsilon^3)|c| + |\lambda| O(\epsilon^2)|c|, \tag{4.18}$$

and from (4.10)

$$\begin{aligned}\sum_j c_j \int_0^1 L_0(h_j - \overline{h_j})(h_k - \overline{h_k}) &= \epsilon^3 s c_k + \sum_j (\gamma\epsilon^3 G_0(x_j, x_k) + \epsilon^2\overline{f'(u)})c_j + o(\epsilon^3)|c| \\ &= O(\epsilon^3)|c| + \epsilon^2\overline{f'(u)}\left(\sum_j c_j\right) \ = \ o(\epsilon^2)|c|.\end{aligned}$$

This estimate and (4.18), (4.11) turn (4.8) into

$$|\lambda|O(\epsilon^2)|c| + o(\epsilon^2)|c| = \lambda(\epsilon\tau c_k + O(\epsilon^2)|c|).$$

From the c_k of the largest absolute value, we find, with the help of Lemma 4.3,

$$\lambda = o(\epsilon), \quad \|\psi\|_\infty = o(\epsilon)|c|. \tag{4.19}$$

So (4.18) is improved to $o(\epsilon^3)|c|$ and (4.8) reads

$$\epsilon^3 s c_k + \sum_j (\gamma\epsilon^3 G_0(x_j, x_k) + \epsilon^2\overline{f'(u)})c_j + o(\epsilon^3)|c| = \lambda\epsilon\tau c_k. \tag{4.20}$$

We sum over k. $\sum_k G_0(x_j^0, x_k^0)$ is independent of j, an issue further addressed in the next section, so we denote it by g. Then after dividing by $\epsilon^2|c|$ we obtain, using (4.19),

$$\frac{\sum_j c_j}{|c|}(\epsilon s + \gamma\epsilon g + \overline{f'(u)}K) + o(\epsilon) = o(1)\frac{\sum_j c_j}{|c|}.$$

Since $\overline{f'(u)} \sim 1$, $\frac{\sum_j c_j}{|c|} = o(\epsilon)$. Return to (4.20). Divide by ϵ^3. Since $\overline{f'(u)}\sum_j c_j = o(\epsilon)|c|$,

$$s c_k + \gamma \sum_j G_0(x_j, x_k)c_j + o(1)|c| = \frac{\lambda}{\epsilon^2}\tau c_k \tag{4.21}$$

for all k. In the limit we have

$$s c_k^0 + \gamma \sum_j G_0(x_j^0, x_k^0)c_j^0 = \eta\tau c_k^0, \quad \sum_j c_j^0 = 0. \tag{4.22}$$

Here $\eta = \lim_{\epsilon\to 0}\lambda/\epsilon^2$ and $c_j^0 = \lim_{\epsilon\to 0} c_j$. In the next section we will solve (4.22) to find $K-1$ pairs of η and c^0. Once they are determined we obtain the asymptotic expansions of $K-1$ eigenpairs $\lambda_{(0,0)} = \lambda$ and $\phi_{(0,0)} = \phi$ of (3.2):

$$\lambda_{(0,0)} = \epsilon^2\eta + o(\epsilon^2), \quad \phi_{(0,0)} \approx \sum_j c_j^0 h_j. \tag{4.23}$$

Here the $\overline{h_j}$ terms drop out in $\phi_{(0,0)}$ since $\sum_j c_j^0\overline{h_j} = \sum_j(1 + \text{e.s.})c_j^0 = \text{e.s.}|c|$ is negligible.

5. The spectrum of $[G_0(x_j^0, x_k^0)]$. To understand (4.22) we must find the spectrum of the K by K matrix $G_0(x_j^0, x_k^0)$. Suppose for every k that

$$\sum_j G_0(x_j^0, x_k^0)b_j = \Lambda b_k.$$

Note that from (4.22) $s + \gamma\Lambda = \tau\eta$. From the formula

$$G_0(x,y) = \begin{cases} \frac{x^2}{2} + \frac{(1-y)^2}{2} - \frac{1}{6} & \text{if } x < y, \\ \frac{(1-x)^2}{2} + \frac{y^2}{2} - \frac{1}{6} & \text{if } x > y \end{cases}$$

we see by straight computation that $\sum_k G_0(x_j^0, x_k^0)$ is independent of j. This number is an eigenvalue whose associated eigenvector is $(1, 1, \dots, 1)^T$, where the superscript ^{T} denotes the transpose of a vector. However, this eigenpair is discarded since by (4.22) we require that $\sum_j b_j = 0$.

To find other eigenpairs we let $\zeta = \sum_j G_0(x_j^0, \cdot) b_j$. Then ζ satisfies $-\zeta'' = \sum_j (\delta(\cdot - x_j^0) - 1) b_j = \sum_j \delta(\cdot - x_j^0)$, $\zeta'(0) = \zeta'(1) = 0$. Moreover $\zeta(x_k) = \Lambda b_k$ and $[-\zeta']_{x_j^0} = b_j$. Then for every k, $[-\zeta']_{x_j^0} = (1/\Lambda)\zeta(x_k^0)$. We need to express $[-\zeta']_{x_j^0}$ in terms of $\zeta(x_k^0)$. In other words we find a K by K matrix T so that $(T\vec{\zeta})_j = [-\zeta']_{x_j^0}$, where $\vec{\zeta} = (\zeta(x_1^0), \zeta(x_2^0), \dots, \zeta(x_K^0))^T$. This way the original eigenvalue problem is converted to

$$T\vec{\zeta} = \frac{1}{\Lambda}\vec{\zeta} \quad \text{with } b = \frac{1}{\Lambda}\vec{\zeta}. \tag{5.1}$$

To find T note that $-\zeta$ is affine between the x_j^0's. From $(0, x_1^0)$ we deduce

$$-\zeta'(x_1^0-) = \frac{-\zeta(x_1^0) + \zeta(x_0^0)}{a/K} = 0$$

since $\zeta'(0) = 0$. From (x_1^0, x_2^0) we obtain

$$-\zeta'(x_1^0+) = \frac{-\zeta(x_2^0) + \zeta(x_1^0)}{2b/K}.$$

Hence

$$[-\zeta']_{x_1^0} = \frac{K}{2b}\zeta(x_1^0) - \frac{K}{2b}\zeta(x_2^0).$$

On the other intervals we find

$$[-\zeta']_{x_j^0} = \left(\frac{K}{2a} + \frac{K}{2b}\right)\zeta(x_j^0) - \begin{cases} \dfrac{K}{2b}\zeta(x_{j-1}^0) - \dfrac{K}{2a}\zeta(x_{j+1}^0) & \text{if } j \text{ is even,} \\ \dfrac{K}{2a}\zeta(x_{j-1}^0) - \dfrac{K}{2b}\zeta(x_{j+1}^0) & \text{if } j \text{ is odd.} \end{cases}$$

When $K = 2$ we have

$$T = \begin{bmatrix} K/(2b) & -K/(2b) \\ -K/(2b) & K/(2b) \end{bmatrix}.$$

Therefore after discarding the 0 eigenvalue of the matrix, we find $\Lambda = \frac{b}{K}$. And for (4.22)

$$\eta = \frac{1}{\tau}\left(-\frac{\gamma a b}{K} + \frac{\gamma b}{K}\right), \quad c^0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{5.2}$$

Note that $\eta > 0$. When $K \geq 3$ we have $T = (\alpha + \beta) I_{K\times K} - Q$, where $I_{K\times K}$ is the K by K identity matrix, $\alpha = K/(2a)$, $\beta = K/(2b)$, and

$$Q = \begin{bmatrix} \alpha & \beta & & & \\ \beta & 0 & \alpha & & \\ & \alpha & 0 & \beta & \\ & & \beta & 0 & \alpha \\ & & & \cdots & \end{bmatrix}.$$

The K distinct eigenvalues of Q are found in (B.5) of Appendix B. One of them, $\alpha+\beta$, is discarded, for its eigenvector is $(1,1,\dots,1)^T$. If we denote the rest of them by q_1, $q_2,\dots,q_{K-1}$, we have $K-1$ Λ's:

$$\Lambda = \frac{1}{\alpha+\beta-q_j}, \quad j=1,2,\dots,K-1.$$

Therefore $K-1$ pairs of

$$\eta = \frac{1}{\tau}\left(-\frac{\gamma ab}{K} + \frac{\gamma}{\alpha+\beta-q_j}\right), \quad c^0 = \vec{\zeta} \tag{5.3}$$

for (4.22) are found.

When concerned with the positivity of η, we consider the smallest Λ, which is associated with the smallest q_j. According to equation (B.5), the smallest q_j is $-\sqrt{\alpha^2+\beta^2+2\alpha\beta\cos\theta}$, where $\theta = 2\pi/K$. Hence the smallest Λ is

$$\Lambda = \frac{1}{\alpha+\beta+\sqrt{\alpha^2+\beta^2+2\alpha\beta\cos\theta}} > \frac{1}{2(\alpha+\beta)} = \frac{ab}{K}.$$

Therefore the smallest η of (4.22) is

$$\eta = \frac{s+\gamma\Lambda}{\tau} > \frac{1}{\tau}\left(-\frac{\gamma ab}{K} + \frac{\gamma ab}{K}\right) = 0.$$

Thus the η's in both (5.2) and (5.3) are positive.

Finally, we show that L_0 has exactly K simple eigenpairs with the asymptotic expansions (4.16) and (4.23). Let F be the linear subspace generated by small eigenvalues. It is defined nonambiguously by $F = \mathrm{span}\{\phi \in L^2(0,1) : \overline{\phi} = 0,\ L_0(\phi) = \lambda\phi,\ |\lambda| < \epsilon^{1/2}\}$. Since the small eigenvalues of L_0 are of order ϵ^2 or ϵ, F addresses all the small eigenvalues when ϵ is small enough.

First dim F, the dimensional of F, is at most K. Suppose that this is not the case. There exist two distinct eigenpairs (λ,ϕ) and (λ',ϕ') with the same asymptotic behavior. That is,

$$\lambda = \epsilon^2\eta + o(\epsilon^2), \quad \lambda' = \epsilon^2\eta + o(\epsilon^2), \qquad \text{or} \qquad \lambda = \epsilon\eta + o(\epsilon), \quad \lambda' = \epsilon\eta + o(\epsilon),$$

$$\phi = \sum_j c_j(h_j - \overline{h_i}) + \psi, \quad \phi = \sum_j c'_j(h_j - \overline{h_i}) + \psi', \quad \lim_{\epsilon\to 0} c_j = \lim_{\epsilon\to 0} c'_j = c^0_j.$$

However, the two eigenfunctions must be orthogonal, so

$$\begin{aligned} 0 &= \langle\phi,\phi'\rangle \\ &= \sum_{j,k} c_j c_k \langle h_j - \overline{h_i}, h_k - \overline{h_k}\rangle + O(|c|)\|\psi\|_2\|h_j\|_2 + O(|c|)\|\psi'\|_2\|h_j\|_2 + \|\psi\|_2\|\psi'\|_2 \\ &= \sum_j c_j^2 \int_0^1 h_j^2\,dx + o(\epsilon)|c|^2 \ = \ \epsilon|c^0|^2 \int_{-\infty}^{\infty} (H'(t))^2\,dt + o(\epsilon)|c^0|^2 \end{aligned}$$

by Remark 4.2. This is obviously impossible when ϵ is sufficiently small.

Next dim F is at least K. Suppose instead that dim $F < K$. Define a subspace of $\{\phi \in L^2(0,1) : \overline{\phi} = 0\}$: $S = \mathrm{span}\{\sum_j c_j^0(h_j - \overline{h_j}) : \text{all } c^0 \text{ found in (5.3)}\}$. We use

a perturbation argument. The asymmetric distance between the closed subspaces S and F is

$$d(S,F)=\sup\{d(\varphi,F):\varphi\in S,\|\varphi\|_2=1\},$$

where $d(x,F)=\inf\{\|x-y\|_2:y\in F\}$. Since $\dim F<\dim S$, there exists $\sum_j c_j^0(h_j-\overline{h_j})\in S$ such that for every eigenvector in F which may be written as $\sum_j c_j'(h_j-\overline{h_j})+\psi$ with $\|\psi\|_2=O(\epsilon)|c'|$ according to (4.7), $\sum_j \frac{c_j'}{|c'|}\frac{c_j^0}{|c^0|}=o(1)$. Then a straight computation shows

$$\left\langle \frac{\sum_j c_j'(h_j-\overline{h_j})+\psi}{\|\sum_j c_j'(h_j-\overline{h_j})+\psi\|_2}, \frac{\sum_j c_j^0(h_j-\overline{h_j})}{\|\sum_j c_j^0(h_j-\overline{h_j})\|_2}\right\rangle=o(1).$$

So if we use

$$\varphi=\frac{\sum_j c_j^0(h_j-\overline{h_j})}{\|\sum_j c_j^0(h_j-\overline{h_j})\|_2},$$

$d(\varphi,F)=1-o(1)$ and $d(S,F)=1-o(1)$. The following lemma due to Helffer and Sjöstrand [5] will give us a contradiction.

LEMMA 5.1. *Let L be a self-adjoint operator on a Hilbert space H, let R be a compact interval in $(-\infty,\infty)$, and let $e_1,e_2,\dots,e_K$ be normalized linearly independent elements in the domain of L. Assume that the following are true:*

1. *$L(e_k)=p_ke_k+r_k$, $\|r_j\|\le\epsilon'$, and $p_j\in R$, $k=1,2,\dots,K$.*
2. *There is $\omega>0$ so that R is ω-isolated in the spectrum of L, i.e., $(\sigma(L)\backslash R)\cap(R+(-\omega,\omega))=\emptyset$.*

Then $d(S,F)\le\frac{K^{1/2}\epsilon'}{\omega\kappa^{1/2}}$, where $S=span\{e_1,\dots,e_K\}$, $F=$ the closed subspace associated with $\sigma(L)\cap R$, and κ equals the smallest eigenvalue of the matrix $[\langle e_j,e_k\rangle]$.

Here we take $L=L_0$, each $e_k\propto\sum_j c_j^0(h_j-\overline{h_j})$ for each one of the K vectors c^0, and S, F as before. ω and κ are positive and bounded away from 0 as $\epsilon\to0$. Set $p_k=\eta\epsilon^2$ or $\eta\epsilon$ depending on c^0 and $R=[-\epsilon^{1/2},\epsilon^{1/2}]$. From (4.4) we find

$$L_0\left(\sum_j c_j^0(h_j-\overline{h_j})\right)-p_k\sum_j c_j^0(h_j-\overline{h_j})=O(\epsilon)|c^0|,$$

and on the other hand $\|\sum_j c_j^0(h_j-\overline{h_j})\|_2\sim\epsilon^{1/2}|c^0|$, as discussed in Remark 4.2. Therefore $\|r_k\|_2=O(\epsilon^{1/2})$. Consequently $d(S,F)=o(1)$, a contradiction. Statement 2 of Theorem 1.1 is proved.

6. $m\neq(0,0)$ eigenvalues. Rewrite (3.3) as

$$L_m(\phi):=-\epsilon^2\phi''+\epsilon\gamma G_m[\phi]+f'(u)\phi=\mu\phi, \tag{6.1}$$

where $\mu=\lambda_m-\epsilon^2m^2\pi^2$. In this section an eigenpair of (6.1) is denoted by (μ,ϕ).

LEMMA 6.1. *If $\mu\to\mu^0$ as $\epsilon\to0$, then $\mu^0\ge0$.*

The proof of this lemma is almost identical to that of Claim 1 in section 3, and we skip it, because the extra term $\epsilon\gamma G_m[\phi]$ is of order $O(\epsilon)$. The case $\mu^0>0$ does not concern us, so we assume $\mu\to0$. Decompose $\phi=\sum_j c_jh_j+\psi$, where $\psi\perp h_j$, $j=1,2,\dots,K$. Note that

$$L_m(h_j)=(f'(u)-f'(H))h_j+\epsilon\gamma G_m[h_j]+\text{e.s.}$$

Because of (3.6) and

$$\epsilon\gamma G_m[h_j](x) = \gamma\epsilon^2 G_m\left[\frac{h_j}{\epsilon}\right](x) = \gamma\epsilon^2 G_m(x, x_j) + O(\epsilon^3),$$

we deduce

$$L_m(h_j) = O(\epsilon^2). \tag{6.2}$$

We write (6.1) as

$$\sum_{j=1}^{K} c_j L_m(h_j) + L_m(\psi) = \mu \sum_j c_j h_j + \mu\psi. \tag{6.3}$$

LEMMA 6.2. *$\langle L_m(\psi), \psi\rangle \geq C\|\psi\|_2^2$ for some $C > 0$ independent of ϵ.*

We skip the proof of this lemma since it is similar to that of Claim 2 in section 3. Multiply (6.3) by ψ and integrate. Use (6.2) to deduce

$$|c|O(\epsilon^2)\|\psi\|_2 + \langle L_m(\psi), \psi\rangle = \mu\|\psi\|_2^2.$$

Then Lemma 6.2 implies

$$\|\psi\|_2 = O(\epsilon^2)|c|. \tag{6.4}$$

Next we multiply (6.3) by h_k and integrate. The left side is

$$\int_0^1 \left\{ L_m(\psi)h_k + \sum_j c_j L_m(h_j)h_k \right\} = \int_0^1 \left\{ L_m(h_k)\psi + \sum_j c_j L_m(h_j)h_k \right\}$$

$$= \int_0^1 \{(f'(u) - f'(H))h_k\psi + \epsilon\gamma G_m[h_k]\psi + \text{e.s.}\,\psi\}$$

$$+ \sum_j c_j \int_0^1 \{(f'(u) - f'(H))h_j h_k + G_m[h_j]h_k + \text{e.s.}\,h_k\}. \tag{6.5}$$

All terms in (6.5) are estimated.

$$\left|\int_0^1 (f'(u) - f'(H))h_k\psi\right| \leq \|(f'(u) - f'(H))h_k\|_\infty\|\psi\|_2 = O(\epsilon^4)|c|$$

by (3.6) and (6.4).

$$\int_0^1 \epsilon\gamma G_m[h_k]\psi = O(\epsilon)\int_0^1 G_m[\psi]h_k = O(\epsilon)\|G_m[\psi]\|_\infty\|h_k\|_1 = O(\epsilon^2)\|\psi\|_2 = O(\epsilon^4)|c|$$

by (6.4). The rest of (6.5) are estimated as in section 4:

$$\int_0^1 (f'(u) - f'(H))h_j h_k = \epsilon^3 s\delta_{jk} + O(\epsilon^4), \quad \int_0^1 \epsilon\gamma G_m[h_j]h_k = \gamma\epsilon^3 G_m(x_j, x_k) + o(\epsilon^3).$$

Hence (6.5) becomes

$$\epsilon^3 s c_k + \sum_j c_j \gamma\epsilon^3 G_m(x_j, x_k) + o(\epsilon^3)|c|.$$

The right side of (6.3) multiplied by h_k and integrated is

$$\int_0^1 \mu \sum_j c_j h_j h_k = \sum_j \mu\epsilon c_j \tau \delta_{jk} + \text{e.s.}|c| = \mu\epsilon\tau c_k + \text{e.s.}|c|.$$

Equating the last two quantities, we find that $\mu = O(\epsilon^2)$ and, for every k,

$$sc_k + \gamma \sum_j G_m(x_j, x_k) c_j + o(1)|c| = \frac{\mu\tau}{\epsilon^2} c_k. \tag{6.6}$$

So in the limit

$$sc_k^0 + \gamma \sum_j G_m(x_j^0, x_k^0) c_j^0 = \eta\tau c_k^0. \tag{6.7}$$

Here $\eta = \lim_{\epsilon\to 0} \mu/\epsilon^2 =$ and $c_j^0 = \lim_{\epsilon\to 0} c_j$. In the next section we will solve (6.7) to determine η and c^0. Once they are found we obtain the asymptotic expansions for the eigenpair $\lambda_m = \mu + \epsilon^2 m^2\pi^2$ and $\phi_m = \phi$:

$$\lambda_m = \epsilon^2(\eta + m^2\pi^2) + o(\epsilon^2), \quad \phi_m \approx \sum_j c_j^0 h_j. \tag{6.8}$$

7. The spectrum of $[G_m(x_j^0, x_k^0)]$. When dealing with

$$\sum_j G_m(x_j^0, x_k^0) b_j = \Lambda b_k,$$

we first consider the simplest case $K = 1$. Then $\Lambda = G_m(x_1^0, x_1^0)$. On $(0, x_1^0)$

$$G_m(x, x_1^0) = \frac{G_m(x_1^0, x_1^0)}{\cosh \tilde{m} x_1^0} \cosh \tilde{m} x,$$

where $\tilde{m} = \pi\sqrt{m_1^2 + m_2^2}$, and on $(x_1^0, 1)$

$$G_m(x, x_1^0) = \frac{G_m(x_1^0, x_1^0)}{\cosh \tilde{m}(1 - x_1^0)} \cosh \tilde{m}(1 - x).$$

Then

$$1 = [-G_m'(\cdot, x_1^0)]_{x_1^0} = \left\{ \frac{\tilde{m} \sinh \tilde{m} x_1^0}{\cosh \tilde{m} x_1^0} + \frac{\tilde{m} \sinh \tilde{m}(1 - x_1^0)}{\cosh \tilde{m}(1 - x_1^0)} \right\} G_m(x_1^0, x_1^0).$$

Therefore

$$\Lambda = \frac{1}{\tilde{m}(\tanh \tilde{m} a + \tanh \tilde{m} b)},$$

and in (6.7)

$$\eta = \frac{1}{\tau} \left(-\gamma ab + \frac{\gamma}{\tilde{m}(\tanh \tilde{m} a + \tanh \tilde{m} b)} \right), \quad c^0 = 1. \tag{7.1}$$

To see the sign of λ_m, we recall

$$\lim_{\epsilon\to 0} \frac{\lambda_m}{\epsilon^2} = \eta + m^2\pi^2 = \frac{1}{\tau} \left(-\gamma ab + \frac{\gamma}{\tilde{m}(\tanh \tilde{m} a + \tanh \tilde{m} b)} \right) + m^2\pi^2.$$

The right side is positive for all $m \neq (0,0)$ if γ is small enough, because $m^2\pi^2$ dominates the negative term. However, when γ is sufficiently large, we may find some large $\tilde{m}$ that makes the right side negative. To see this we first take $\tilde{m}$ large enough so that sum of the two terms in the parentheses is negative. Then we take γ large enough so the entire right side is negative.

When $K \geq 2$, $\sum_j G_m(x_j^0, x_k^0) b_j = \Lambda b_k$ is a more complex problem. Let ζ be the solution of $-\zeta'' + m^2\pi^2\zeta = \sum_j \delta(\cdot - x_j^0) b_j$, $\zeta'(0) = \zeta'(1) = 0$. Hence $[-\zeta']_{x_k^0} = b_k$ and $\zeta(x_k^0) = \Lambda b_k$. Then for every k, $[-\zeta']_{x_k^0} = \frac{1}{\Lambda}\zeta(x_k^0)$. As in section 5 we express $[-\zeta']_{x_k^0} = (T\vec{\zeta})_k$ in order to convert to the new eigenvalue problem $T\vec{\zeta} = (1/\Lambda)\vec{\zeta}$. Away from x_j^0, $\zeta = g_1 \cosh \tilde{m}x + g_2 \sinh \tilde{m}x$. From here we write, in the matrix notation,

$$\begin{bmatrix} \zeta(x_{j-1}^0) \\ \zeta(x_j^0) \end{bmatrix} = \begin{bmatrix} \cosh \tilde{m}x_{j-1}^0 & \sinh \tilde{m}x_{j-1}^0 \\ \cosh \tilde{m}x_j^0 & \sinh \tilde{m}x_j^0 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}.$$

We denote the 2 by 2 matrix by A_L for the left of x_j^0. To the right we have similarly

$$\begin{bmatrix} \zeta(x_j^0) \\ \zeta(x_{j+1}^0) \end{bmatrix} = \begin{bmatrix} \cosh \tilde{m}x_j^0 & \sinh \tilde{m}x_j^0 \\ \cosh \tilde{m}x_{j+1}^0 & \sinh \tilde{m}x_{j+1}^0 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix},$$

with the 2 by 2 matrix denoted by A_R. Hence

$$\begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = A_L^{-1} \begin{bmatrix} \zeta(x_{j-1}^0) \\ \zeta(x_j^0) \end{bmatrix} \text{ on } (x_{j-1}^0, x_j^0), \quad \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = A_R^{-1} \begin{bmatrix} \zeta(x_j^0) \\ \zeta(x_{j+1}^0) \end{bmatrix} \text{ on } (x_j^0, x_{j+1}^0).$$

Then

$$-\zeta'(x_j^0-) = -\tilde{m}[\sinh \tilde{m}x_j^0, \cosh \tilde{m}x_j^0] A_L^{-1} \begin{bmatrix} \zeta(x_{j-1}^0) \\ \zeta(x_j^0) \end{bmatrix},$$

$$-\zeta'(x_j^0+) = -\tilde{m}[\sinh \tilde{m}x_j^0, \cosh \tilde{m}x_j^0] A_R^{-1} \begin{bmatrix} \zeta(x_j^0) \\ \zeta(x_{j+1}^0) \end{bmatrix},$$

and

$$[-\zeta']_{x_j^0} = \tilde{m}[\sinh \tilde{m}x_j^0, \cosh \tilde{m}x_j^0] \left\{ A_L^{-1} \begin{bmatrix} \zeta(x_{j-1}^0) \\ \zeta(x_j^0) \end{bmatrix} - A_R^{-1} \begin{bmatrix} \zeta(x_j^0) \\ \zeta(x_{j+1}^0) \end{bmatrix} \right\}.$$

We also compute

$$A_L^{-1} = \frac{1}{\sinh \tilde{m}(x_j^0 - x_{j-1}^0)} \begin{bmatrix} \sinh \tilde{m}x_j^0 & -\sinh \tilde{m}x_{j-1}^0 \\ -\cosh \tilde{m}x_j^0 & \cosh \tilde{m}x_{j-1}^0 \end{bmatrix},$$

$$A_R^{-1} = \frac{1}{\sinh \tilde{m}(x_{j+1}^0 - x_j^0)} \begin{bmatrix} \sinh \tilde{m}x_{j+1}^0 & -\sinh \tilde{m}x_j^0 \\ -\cosh \tilde{m}x_{j+1}^0 & \cosh \tilde{m}x_j^0 \end{bmatrix}.$$

Thus T is a triagonal matrix. The three entries of the jth row where $j \neq 1, K$ are

$$-\tilde{m}\operatorname{csch} \tilde{m}(x_j^0 - x_{j-1}^0), \quad \tilde{m} \coth \tilde{m}(x_j^0 - x_{j-1}^0) + \tilde{m} \coth \tilde{m}(x_{j+1}^0 - x_j^0),$$

$$-\tilde{m}\operatorname{csch} \tilde{m}(x_{j+1}^0 - x_j^0).$$

For the first row,

$$\zeta(x) = \frac{\zeta(x_1^0)}{\cosh \tilde{m} x_1^0} \cosh \tilde{m} x$$

and

$$[-\zeta']_{x_1^0} = \tilde{m}(\tanh \tilde{m} x_1^0 + \coth \tilde{m}(x_2^0 - x_1^0))\zeta(x_1^0) - \tilde{m}\mathrm{csch}\, \tilde{m}(x_2^0 - x_1^0)\zeta(x_2^0).$$

When $K = 2$ the matrix T is

$$\tilde{m} \begin{bmatrix} \tanh \tilde{m}a/2 + \coth \tilde{m}b & -\mathrm{csch}\, \tilde{m}b \\ -\mathrm{csch}\, \tilde{m}b & \tanh \tilde{m}b/2 + \coth \tilde{m}a \end{bmatrix}$$

$$= \tilde{m}(\coth \tilde{m}a + \coth \tilde{m}b) I_{K\times K} - \tilde{m} \begin{bmatrix} \mathrm{csch}\, \tilde{m}a & \mathrm{csch}\, \tilde{m}b \\ \mathrm{csch}\, \tilde{m}b & \mathrm{csch}\, \tilde{m}a \end{bmatrix}.$$

The two $(1/\Lambda)$'s are

$$\tilde{m}(\coth \tilde{m}a + \coth \tilde{m}b - \mathrm{csch}\, \tilde{m}a - \mathrm{csch}\, \tilde{m}b),$$
$$\tilde{m}(\coth \tilde{m}a + \coth \tilde{m}b - \mathrm{csch}\, \tilde{m}a + \mathrm{csch}\, \tilde{m}b),$$

which again lead to η for (6.7). To see the sign of λ_m, we take the smaller Λ so that the smaller λ_m satisfies

$$\lim_{\epsilon\to 0} \frac{\lambda_m}{\epsilon^2} = \eta + m^2\pi^2 = \frac{1}{\tau}\left(-\frac{\gamma ab}{2} + \frac{\gamma}{\tilde{m}(\coth \tilde{m}a + \coth \tilde{m}b - \mathrm{csch}\, \tilde{m}a + \mathrm{csch}\, \tilde{m}b)}\right) + m^2\pi^2.$$

As in the $K = 1$ case the right side is positive for all $m \neq (0, 0)$ if γ is small, and is negative for some m if γ is large.

When $K \geq 3$ we write $T = d\, I_{K\times K} - Q$ with

$$Q = \begin{bmatrix} \alpha & \beta & & & \\ \beta & 0 & \alpha & & \\ & \alpha & 0 & \beta & \\ & & \beta & 0 & \alpha \\ & & & \cdots & \end{bmatrix},$$

where

$$\alpha = \tilde{m}\mathrm{csch}\, \frac{2\tilde{m}a}{K}, \quad \beta = \tilde{m}\mathrm{csch}\, \frac{2\tilde{m}b}{K}, \quad d = \tilde{m} \coth \frac{2\tilde{m}a}{K} + \tilde{m} \coth \frac{2\tilde{m}b}{K}.$$

Because of diagonal domination the matrix $d\, I_{K\times K} - Q$ is positive definite. The K eigenvalues of Q are found in (B.5) of Appendix B. We again denote them by q_j, $j = 1, 2, \ldots, K$. Then $\Lambda = \frac{1}{d - q_j}$, and for (6.7), K eigenpairs

$$\eta = \frac{1}{\tau}\left(-\frac{\gamma ab}{K} + \frac{\gamma}{d - q_j}\right), \quad c^0 = \vec{\zeta} \tag{7.2}$$

are found. To see the sign of λ_m, we focus on the smallest η, which is associated with $q_j = -\sqrt{\alpha^2 + \beta^2 + 2\alpha\beta\cos\theta}$, where $\theta = 2\pi/K$. For this q_j

$$\lim_{\epsilon\to 0} \frac{\lambda_m}{\epsilon^2} = \eta + m^2\pi^2 = \frac{1}{\tau}\left(-\frac{\gamma ab}{K} + \frac{\gamma}{d - q_j}\right) + m^2\pi^2$$
$$= \frac{1}{\tau}\left(-\frac{\gamma ab}{K} + \frac{\gamma}{d + \sqrt{\alpha^2 + \beta^2 + 2\alpha\beta\cos\theta}}\right) + m^2\pi^2.$$

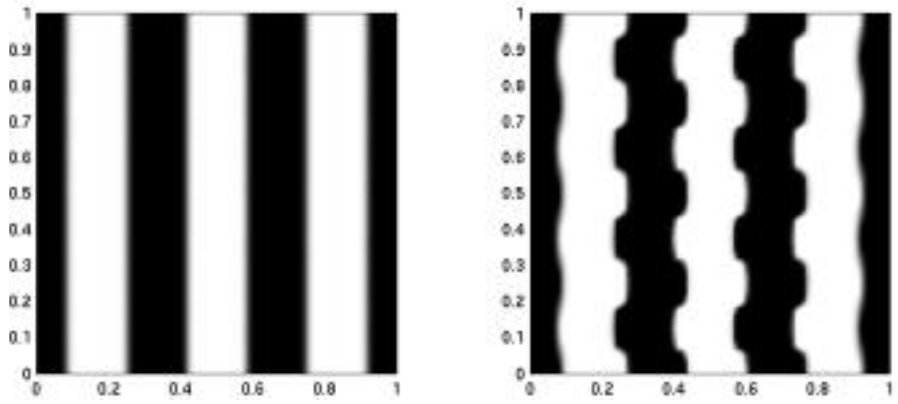


FIG. 4. *A lamellar solution u and its deformation in a cross section perpendicular to y_2.*

The dependence of the positivity of the right side on γ is still the same; i.e., the right side is positive for all m if γ is small and negative for some m if γ is large. The dependence of the positivity on K is also clear. When K is large, $m^2\pi^2$ dominates the negative term, so the whole quantity is positive for all m.

Remark 7.1. If γ and K are held fixed, then the last line is positive if $|m|$ is sufficiently large. This is consistent with statement 1 of Theorem 1.1.

We omit the proof that L_m has exactly K simple eigenpairs with small eigenvalues, because it is similar to that for L_0. This concludes the proof of statement 3, Theorem 1.1.

To visualize statement 3, Theorem 1.1, consider the example $a = 0.4$ and $K = 6$. We study $m = (8, 0)$ and find the c^0 associated with the smallest $\lambda_{(8,0)}$ by numerically diagonalizing Q:

$$c^0 = (0.0424,\ -0.4774,\ 0.5199,\ -0.5199,\ 0.4774,\ -0.0424)^T.$$

The eigenfunction of L associated with this $\lambda_{(8,0)}$ and with c^0 is approximately $\sum_j c_j^0 h_j(x)\cos(8\pi y_1)$. When γ is sufficiently large, we have $\lambda_{(8,0)} < 0$. Then the unstable lamellar solution u may easily be deformed in the direction of this eigenfunction. In Figure 4 we make a cross section of D, perpendicular to the y_2 direction. The first plot shows u on this cross section, where the black color indicates $u \approx 1$ and the white color indicates $u \approx 0$. The second plot shows u deformed by the eigenfunction. Note that under this deformation the straight interfaces in u become wriggled curves. See [7] for a heuristic argument for this change of shape.

8. The 1-D global minimizer. The integral $\int_0^1 |(v^0)'|^2\,dx$ in the conclusion of Theorem 2.1 may be calculated as

$$\begin{aligned}\int_0^1 |(v^0)'|^2\,dx &= K\int_0^{a/K} |(v^0)'|^2\,dx + K\int_{a/K}^{1/K} |(v^0)'|^2\,dx \\ &= K\int_0^{a/K} (1-a)^2x^2\,dx + K\int_{a/K}^{1/K} a^2\left(x-\frac{1}{K}\right)^2\,dx \\ &= \frac{a^3b^2}{3K^2} + \frac{a^2b^3}{3K^2} \;=\; \frac{a^2b^2}{3K^2}.\end{aligned}$$

Hence

$$\lim_{e\to 0}\epsilon^{-1}I_1(u)=\tau K+\frac{\gamma a^2b^2}{6K^2}. \tag{8.1}$$

It was shown in [13] that the 1-D global minimizer is a 1-D local minimizer whose number of interfaces K_* minimizes the right side of (8.1). Note that in some less likely cases two integers K_* and K_*+1 may both minimize the right side of (8.1). Then we may have two global minimizers with K_* and K_*+1 interfaces, respectively.[5] If we pretend that K is a positive real number and minimize the right side with respect to K, then the minimum is achieved at

$$K_*=\left(\frac{a^2b^2\gamma}{3\tau}\right)^{1/3}. \tag{8.2}$$

We set $t=\tilde{m}/K$. Consider the eigenvalue λ_m that is associated with the smallest η of section 7.

$$\lim_{\epsilon\to 0}\frac{\lambda_m}{\epsilon^2}\frac{\tau\tilde{m}}{\gamma}=\frac{\tau\tilde{m}(\eta+m^2\pi^2)}{\gamma}$$

$$=-abt+\frac{1}{\coth 2at+\coth 2bt+\sqrt{(\operatorname{csch}2at)^2+(\operatorname{csch}2bt)^2+2\operatorname{csch}2at\operatorname{csch}2bt\cos\theta}}+\frac{\tau K^3t^3}{\gamma}.$$

For the 1-D global minimizer, we use K_* in (8.2) for K to find

$$\lim_{\epsilon\to 0}\frac{\lambda_m}{\epsilon^2}\frac{\tau\tilde{m}}{\gamma}=\frac{\tau\tilde{m}(\eta+m^2\pi^2)}{\gamma}$$

$$=-abt+\frac{1}{\coth 2at+\coth 2bt+\sqrt{(\operatorname{csch}2at)^2+(\operatorname{csch}2bt)^2+2\operatorname{csch}2at\operatorname{csch}2bt\cos\theta}}+\frac{a^2b^2t^3}{3}. \tag{8.3}$$

Note that a natural lower bound for the second term in (8.3) is

$$\frac{1}{\coth 2at+\coth 2bt+\operatorname{csch}2at+\operatorname{csch}2bt}$$
$$=\frac{\sinh at\sinh bt}{\sinh t}=abt-\frac{a^2b^2t^3}{3}+\frac{(a^2b^2+2a^3b^3)t^5}{45}+\cdots$$

by replacing $\cos\theta$ by 1. This lower bound is sharp if K_* is large, i.e., γ is large. The first three terms of the Taylor expansion are given. We observe that the first two terms in the Taylor expansion are *exactly* canceled by the first and third terms in (8.3). This is certainly no coincidence. The fifth order term is positive. Our numerical tests confirm that all of (8.3) remains positive. The particular K_* of (8.2) is barely large enough to overcome the negative third order term in $\frac{\sinh at\sinh bt}{\sinh t}$.

To contemplate the physical significance of the shaky stability property of the 1-D global minimizer, we first note that the value (8.2) for K_* is only approximate. But the 1-D global minimizer is very close to the borderline of 3-D stability. 1-D local minimizers with larger K are likely to be stable in three dimensions, and 1-D local minimizers with smaller K are likely to be unstable in three dimensions. In the real

[5] Actually there are four global minimizers because of Remark 2.2 if we include solutions of both types.

physical system only the 3-D global minimizer, which is unlikely to be lamellar, is the thermal equilibrium. Other stable solutions of (1.2) are only transient, metastable states. In general lamellar phases, including the 1-D global minimizer, are transient. They are vulnerable to perturbations of the form $\sum_j c_j^0 h_j(x)\cos(m_1\pi y_1)\cos(m_2\pi y_2)$ found in statement 3 of Theorem 1.1, which push the straight interfaces in a lamellar state to a wriggled shape; see Figure 4. In a forthcoming paper [12] we will actually prove, using bifurcation analysis, that (1.2) admits wriggled solutions for some values of γ.

Appendix A. Proof of Theorem 2.3.

LEMMA A.1. *u has exactly K transition layers in the sense that there are exactly K points, $x_1, x_2, \ldots, x_K$, in $(0,1)$, so that $u(x_j) = 1/2$, $j = 1, 2, \ldots, K$, and $\lim_{\epsilon\to 0} x_j = x_j^0$.*

The proof of this lemma is similar to that of [11, Proposition 8.2].

LEMMA A.2. *The derivative of $v = G_0[u-a]$ has exactly $K-1$ zeros, denoted by $y_1, y_2, \ldots, y_{K-1}$, in $(0,1)$, such that $\lim_{\epsilon\to 0} y_j = j/K$.*

Proof. The derivative of $v^0 = G_0[u^0 - a]$ has zeros at $1/K$, $2/K$, ..., $(K-1)/K$. The convergence of v' to $(v^0)'$ implies that v' has exactly $K-1$ zeros y_j with the property $\lim_{\epsilon\to 0} y_j = j/K$. ∎

We set $y_0 = 0$ and $y_K = 1$. Let $l_i = y_i - y_{i-1}, i = 1, \ldots, K$. Between two zeros of v' we integrate the equation $-v'' = u - a$ and find $\frac{1}{l_i}\int_{y_{i-1}}^{y_i} u\,dx = a$. This allows us to localize the energy of u on (y_{i-1}, y_i). If we set $l_i\xi + y_{i-1} = x$, $\mathcal{U}_i(\xi) = u(x)$, and $\mathcal{V}_i(\xi) = l_i^{-2}v(x) = l_i^{-2}G_0[u-a](x)$, then $\int_0^1 \mathcal{U}_i\,dz = a$, $-\mathcal{V}_i'' = \mathcal{U}_i - a$, $\mathcal{V}_i'(0) = \mathcal{V}_i'(1) = 0$. More importantly,

$$\begin{aligned} I_1(u) &= \sum_{i=1}^K \int_{y_{i-1}}^{y_i} \left\{ \frac{\epsilon^2}{2}|u|^2 + \frac{\gamma\epsilon}{2}|v'|^2 + W(u) \right\} dx \\ &= \sum_{i=1}^K l_i \int_0^1 \left\{ \frac{\epsilon^2}{2l_i^2}|\mathcal{U}_i'|^2 + \frac{l_i^2\gamma\epsilon}{2}|\mathcal{V}_i'|^2 + W(\mathcal{U}_i) \right\} d\xi \;=\; \sum_{i=1}^K l_i J_{l_i}(\mathcal{U}_i) \end{aligned} \tag{A.1}$$

if we define a new variational problem:

$$J_l(\mathcal{U}) = \int_0^1 \left\{ \frac{\epsilon^2}{2l^2}|\mathcal{U}'|^2 + \frac{l^3\gamma}{2}\frac{\epsilon}{l}\left| \left(-\frac{d^2}{d\xi^2}\right)^{-1/2}(\mathcal{U}-a) \right|^2 + W(\mathcal{U}) \right\} d\xi, \quad \mathcal{U} \in X_a. \tag{A.2}$$

This new J_l is similar to the original I_1. l lies in a compact subinterval of $(0,1)$, so we take $l \sim 1$. We consider a one-layer local minimizer $\mathcal{U}$ that is close to $\mathcal{U}^0$, which is 0 on $(0, 1-a)$ and 1 on $(1-a, 0)$. The dependence of $\mathcal{U}$ on l and ϵ is suppressed in the notation. It is proved in Proposition 9.2 of [11] that this local minimizer is unique in an L^2 ball centered at $\mathcal{U}^0$ of radius δ. δ is small but independent of ϵ. Denote the transition point of $\mathcal{U}$ by χ, i.e., $\mathcal{U}(\chi) = 1/2$. This one-layer local minimizer has the following asymptotic expansion.

LEMMA A.3. *Let $\tilde\epsilon = \frac{\epsilon}{l}$ and $\tilde\gamma = l^3\gamma$. Then $\mathcal{U} = H(\frac{\cdot-\chi}{\tilde\epsilon}) + \tilde\epsilon\mathcal{Z}$ with $\|\mathcal{Z} - \mathcal{Z}^0\|_\infty = O(\tilde\epsilon)$. Here $\mathcal{Z}^0 = -\frac{\tilde\gamma(\mathcal{V}^0 - \mathcal{V}^0(1-a))}{f'(0)}$, $\mathcal{V}^0 = G_0[\mathcal{U}^0 - a]$. Note that $\mathcal{Z}^0(1-a) = 0$.*

Proof. See Proposition 8.3 in [11]. ∎

LEMMA A.4. *Let $F \in C^2(-\infty, \infty)$ be such that $F(0) = F(1) = 0$. Then*

$$\int_0^1 F(\mathcal{U})\,d\xi = \tilde\epsilon\int_{-\infty}^{\infty} F(H)\,dt + \tilde\epsilon\int_0^{1-a} F'(0)\mathcal{Z}^0\,d\xi + \tilde\epsilon\int_{1-a}^1 F'(1)\mathcal{Z}^0\,d\xi + O(\tilde\epsilon^2).$$

Proof. See Lemma 8.4 in [11]. □

LEMMA A.5. *Let* $\mathcal{W} = \frac{\partial \mathcal{U}}{\partial l}$. *Then*

$$\mathcal{W}(\xi) = H'\left(\frac{l(\xi-\chi)}{\epsilon}\right)\frac{\xi-\chi}{\epsilon} - \overline{H'\left(\frac{l(\xi-\chi)}{\epsilon}\right)\frac{\xi-\chi}{\epsilon}} + \varphi,$$

with $\|\varphi\|_2 = O(1)$. *And* $\varphi = c(h-\overline{h}) + \psi$, *with* $h = H'(\frac{l(\xi-\chi)}{\epsilon})$, $h - \overline{h} \perp \psi$, $c = O(\epsilon^{-1/2})$, *and* $\|\psi\|_2 = O(\epsilon)$.

Proof. The brief argument here summarizes the more elaborate proof of the similar Proposition 9.3 in [11]. Differentiate the Euler–Lagrange equation of (A.2) with respect to l to obtain

$$-\left(\frac{\epsilon}{l}\right)^2 \mathcal{W}'' + \gamma\epsilon l^2 G_0[\mathcal{W}] + f'(\mathcal{U})\mathcal{W} + 4\gamma\epsilon l G_0[u-a] + \frac{2}{l}f(\mathcal{U}) - \frac{2}{l}\overline{f(\mathcal{U})} = \frac{d\overline{f(\mathcal{U})}}{dl} \tag{A.3}$$

for $\mathcal{W}$. Define $g(\xi) = H'(\frac{l(\xi-\chi)}{\epsilon})\frac{\xi-\chi}{\epsilon}$ + e.s. so that g and its derivative vanish at 0 and 1, and $\varphi = \mathcal{W} - (g - \overline{g})$. Here g satisfies the equation

$$-\left(\frac{\epsilon}{l}\right)^2 g'' + f'(H)g + \frac{2}{l}f(H) = \text{e.s.}$$

Subtract this from (A.3) and use the facts $\|g\|_2 = O(\epsilon^{1/2})$, $\overline{g} = O(\epsilon)$, and $(f'(\mathcal{U}) - f'(H))(g - \overline{g}) = O(\epsilon^2)$, where the last one follows from Lemma A.3, to deduce the equation for φ:

$$-\left(\frac{\epsilon}{l}\right)^2 \varphi'' + \epsilon\gamma l^2 G_0[\varphi] + f'(\mathcal{U})\varphi + O(\epsilon) = \text{Const.}, \tag{A.4}$$

where we simply write Const. for a constant since its exact value is not needed in this proof. We multiply this equation by φ and integrate:

$$\int_0^1 \left\{\left(\frac{\epsilon}{l}\right)^2 |\varphi'|^2 + \epsilon\gamma l^2 G_0[\varphi]\varphi + f'(\mathcal{U})\varphi^2\right\} dz = O(\epsilon)\|\varphi\|_2.$$

By Proposition 9.1 in [11] we find $\|\varphi\|_2 = O(1)$.

Decompose $\varphi = c(h-\overline{h}) + \psi$, where $h = H'(\frac{l(\xi-\chi)}{\epsilon})$ + e.s. and $h - \overline{h} \perp \psi$. The exponentially small correction term e.s. is added so that h and h' vanish at 0 and 1. Then

$$c = \frac{\int_0^1 \varphi(h-\overline{h})\,dz}{\|h-\overline{h}\|_2^2} \le \frac{\|\varphi\|_2}{\|h-\overline{h}\|_2} = O(\epsilon^{-1/2}).$$

The equation satisfied by ψ is

$$-\left(\frac{\epsilon}{l}\right)^2 \psi'' + f'(\mathcal{U})\psi + O(\epsilon) = \text{Const.},$$

where we have used the fact $(f(\mathcal{U}) - f'(H))h = O(\epsilon^2)$, again a consequence of Lemma A.3. Argue as in Lemma 4.1 to deduce $\int_0^1 \{-(\frac{\epsilon}{l})^2\psi'' + f'(\mathcal{U})\psi\}\psi\,d\xi \ge C\|\psi\|_2^2$, which implies $\|\psi\|_2 = O(\epsilon)$. □

LEMMA A.6. *Let* $E(l) = lJ_l(\mathcal{U})$. *Then* $E(l)$ *is strictly convex in* l *in any compact subset of* $(0,1)$.

Proof. This lemma is similar to Proposition 10.1 in [11]. Differentiating E with respect to l yields

$$\frac{\partial E}{\partial l} = \int_0^1 \left\{ -\frac{\epsilon^2}{2l^2}|\mathcal{U}'|^2 + W(\mathcal{U}) + \frac{3\epsilon\gamma l^2}{2}|\mathcal{V}'|^2 \right\} d\xi, \tag{A.5}$$

where $\mathcal{V} = G_0[\mathcal{U} - a]$. We have used the fact that $\mathcal{U}$ is a critical point of J_l. Differentiate (A.5) with respect to l:

$$\frac{\partial^2 E}{\partial l^2} = \int_0^1 \left\{ \frac{\epsilon^2}{l^3}|\mathcal{U}'|^2 + 3\epsilon\gamma l|\mathcal{V}'|^2 \right\} d\xi + \int_0^1 \{2f(\mathcal{U})\mathcal{W} + 4\epsilon\gamma l^2 \mathcal{V}\mathcal{W}\}\, d\xi.$$

Call the first integral on the right side T_1 and the second integral T_2. Multiplying the Euler–Lagrange equation of $\mathcal{U}$ by $\mathcal{U} - a$ and integrating by parts, we find the useful integral identity

$$\int_0^1 \left\{ \left(\frac{\epsilon}{l}\right)^2 |\mathcal{U}'|^2 + f(\mathcal{U})(\mathcal{U} - a) + \epsilon\gamma l^2 |\mathcal{V}'|^2 \right\} d\xi = 0.$$

Using this identity and Lemma A.4, we obtain

$$\begin{aligned} T_1 &= \frac{1}{l}\int_0^1 \{-f(\mathcal{U})(\mathcal{U} - a) + 2\epsilon\gamma l^2 |\mathcal{V}'|^2\}\, d\xi \\ &= \frac{\epsilon}{l^2}\int_{-\infty}^{\infty} -f(H)(H - a)\, dt + \frac{\epsilon\gamma l a^2 b^2}{3} + \frac{2\epsilon\gamma l a^2 b^2}{3} + O(\epsilon^2) \\ &= \frac{\epsilon}{l^2}\int_{-\infty}^{\infty} -f(H)H\, dt + \epsilon\gamma l a^2 b^2 + O(\epsilon^2). \end{aligned}$$

Here we have used

$$\int_0^1 |\mathcal{V}'|^2\, d\xi = \int_0^1 |\mathcal{V}_0'|^2\, d\xi + O(\epsilon) = \frac{a^2 b^2}{3} + O(\epsilon),$$

which follows from (8.17) in [11]. By Lemmas A.3 and A.5

$$\begin{aligned} T_2 &= \int_0^1 (2f(H) + O(\epsilon)) \left(\frac{\xi - \chi}{\epsilon} H' + + cH' - c\overline{H'} + \psi \right) d\xi \\ &= \frac{\epsilon}{l^2}\int_{-l\chi/\epsilon}^{l(1-\chi)/\epsilon} 2f(H(t))H'(t)t\, dt + O(\epsilon^{1.5}) \;=\; \frac{\epsilon}{l^2}\int_{-\infty}^{\infty} -2W(H)\, dt + O(\epsilon^{1.5}). \end{aligned}$$

We have used the estimates

$$\begin{aligned} \int_0^1 \left| \frac{\xi - \chi}{\epsilon} H' \right| d\xi &= \frac{\epsilon}{l^2}\int_{-l\chi/\epsilon}^{l(1-\chi)/\epsilon} |H'(t)t|\, dt = O(\epsilon), \\ \int_0^1 |f(H)|\, d\xi &= \frac{\epsilon}{l}\int_{-l\chi/\epsilon}^{l(1-\chi)/\epsilon} |f(H(t))|\, dt = O(\epsilon), \\ \|2f(H) + O(\epsilon)\|_2 &= O(\epsilon^{1/2}), \\ \int_0^1 f(H)H'\, d\xi &= \text{e.s.} \end{aligned}$$

Adding T_1 and T_2, since $\int_{-\infty}^{\infty}(f(H)H+2W(H))\,dt=0$ (a consequence of the integral identity $\int_{-\infty}^{\infty}\{(H')^2+f(H)H\}\,dt=0$ and the first integral of H), we arrive at

$$\frac{\partial^2 E}{\partial l^2}=\epsilon l\gamma a^2b^2+O(\epsilon^{1.5}), \tag{A.6}$$

proving the lemma. □

Proof of Theorem 2.3. We construct a particular periodic solution u_* with K transition layers and show that $u=u_*$. Let $\mathcal{U}$ be the unique minimum of J_l in a δ neighborhood of $\mathcal{U}^0$, with $l=1/K$, and let $\mathcal{U}^R=\mathcal{U}(1-\cdot)$ be its reversal. Set $u_*(x)=\mathcal{U}^R(Kx)$ for $x\in(0,1/K)$. Extend u_* antiperiodically to $(0,1)$, i.e., $u_*(x)=\mathcal{U}(Kx-1)$ for $x\in(1/K,2/K)$, $u_*(x)=\mathcal{U}^R(Kx-2)$ for $x\in(2/K,3/K),\dots$. Clearly u_* is periodic with $K/2$ periods.

For small ϵ, u and u_* belong to the same small L^2 neighborhood in which u is a minimizer. Using the strict convexity of E in Lemma A.6 and (A.1), we find

$$I(u_*)\ge I(u)=\sum_{i=1}^{K}l_iJ_{l_i}(u(l_i\cdot+y_{i-1}))\ge\sum_{i=1}^{K}E(l_i)\ge KE\left(\frac{1}{K}\right)=I(u_*).$$

All the inequalities above must be equalities. Therefore $l_i=1/K$, $y_i=i/K$ for all i, and $l_iJ_{l_i}(u(l_i\cdot+y_{i-1}))=E(l_i)$. Moreover $u((1/K)\cdot+y_{i-1})=\mathcal{U}$ when i is even or $=\mathcal{U}^R$ when i is odd by the local uniqueness of $\mathcal{U}$ and $\mathcal{U}^R$ [11, Proposition 9.2]. Thus $u=u_*$. □

Appendix B. The matrix Q. Consider a matrix Q like those in sections 5 and 7 with $\alpha,\beta>0$. In this appendix Q, whose size is at least 3 by 3, acts on the complex vector space $\mathbf{C}^K$. Let $\vec{q}=(z,tz^2,z^3,tz^4,\dots)^T$, where $z,t\in\mathbf{C}$ and $|z|=1$. Suppose the eigenvalue problem $Q\vec{q}=q\vec{q}$ holds for the second through the next-to-last equations, excluding the first and the last. In these $K-2$ equations

$$\begin{array}{ll}\alpha tz^{l-1}+\beta tz^{l+1}=qz^l & \text{if } l \text{ is odd,}\\ \beta z^{l-1}+\alpha z^{l+1}=qtz^l & \text{if } l \text{ is even.}\end{array} \tag{B.1}$$

They imply

$$t=\pm\frac{\alpha z+\beta\overline{z}}{|\alpha z+\beta\overline{z}|}. \tag{B.2}$$

In particular $|t|=1$. In order to have the first and the last equations satisfied, we let $\vec{h}=A\vec{q}+B\bar{\vec{q}}$ and study $Q\vec{h}=q\vec{h}$.

If the vector $\vec{q}$ is extended by t as the 0th entry and by tz^{K+1} as the $(K+1)$th entry if K is odd, or by z^{K+1} as the $(K+1)$th entry if K is even, then the first and the last equations of $Q\vec{h}=q\vec{h}$ are satisfied if the 0th entry is equal to the first entry and the Kth entry is equal to the $(K+1)$th entry. That is,

$$\begin{cases} Az+B\overline{z}=At+B\bar{t}\\ Az^K+B\overline{z}^K=Atz^{K+1}+B\bar{t}\overline{z}^{K+1}\end{cases}\quad\text{if } K \text{ is odd;} \tag{B.3}$$

$$\begin{cases} Az+B\overline{z}=At+B\bar{t}\\ Atz^K+B\bar{t}\overline{z}^K=Az^{K+1}+B\overline{z}^{K+}\end{cases}\quad\text{if } K \text{ is even.} \tag{B.4}$$

They should have nontrivial solutions for A and B. In the case of (B.3) this means

$$(z-t)(1-\overline{t}\overline{z})\overline{z}^K = (\overline{z}-\overline{t})(1-tz)z^K$$

or, since $|z| = |t| = 1$, $z^{2K} = 1$. The case (B.4) gives the same condition. Define $\theta = \frac{2\pi(j-1)}{K}$, $j = 1, 2, \ldots, 2K$. Then $z = e^{i\theta/2}$. From (B.1) we find

$$q = \alpha t\overline{z} + \beta tz = \pm\sqrt{\alpha^2+\beta^2+2\alpha\beta\cos\theta}.$$

Here θ ranges from 0 to $4\pi - (2\pi/K)$, which is too wide a range. We restrict j to $1, 2, \ldots, (K+1)/2$ if K is odd and $j = 1, 2, \ldots, K/2+1$ if K is even. Even then we have some extra values. When $z = 1$ and $t = -1$, which occur if $\theta = 0$ and $q = -(\alpha+\beta)$, we find $A + B = 0$ and $\vec{h} = \vec{0}$, which is not an eigenvector. Also when K is even, $z = i$, and $t = -i$, which occur if $\theta = \pi$, and $q = \beta - \alpha$, we find $A - B = 0$ and again $\vec{h} = \vec{0}$. In summary the K distinct eigenvalues of Q are

$$\alpha+\beta,\ \pm\sqrt{\alpha^2+\beta^2+2\alpha\beta\cos\theta}\quad\left(\theta = \frac{2\pi(j-1)}{K},\ j = 2, 3, \ldots, \frac{K+1}{2}\right)\quad \text{if } K \text{ is odd;}$$

(B.5)

$$\alpha+\beta,\ \pm\sqrt{\alpha^2+\beta^2+2\alpha\beta\cos\theta}\quad\left(\theta = \frac{2\pi(j-1)}{K},\ j = 2, 3, \ldots, \frac{K}{2}\right),\ \alpha-\beta\ \text{ if } K \text{ is even.}$$

Acknowledgments. We thank Professors R. Kohn and C. Muratov for several stimulating conversations on the subject of block copolymers during the SIAM 50th Anniversary and 2002 Meeting in Philadelphia.

REFERENCES

[1] J.W. Cahn and J.E. Hilliard, *Free energy of a nonuniform system.* I. *Interfacial free energy,* J. Chem. Phys., 28 (1958), pp. 258–267.

[2] R. Choksi, *Scaling laws in microphase separation of diblock copolymers,* J. Nonlinear Sci., 11 (2001), pp. 223–236.

[3] R. Choksi and X. Ren, *On the derivation of a density functional theory for microphase separation of diblock copolymers,* J. Statist. Phys., to appear.

[4] P.C. Fife and D. Hilhorst, *The Nishiura–Ohnishi free boundary problem in the* 1*D case,* SIAM J. Math. Anal., 33 (2001), pp. 589–606.

[5] B. Helffer and J. Sjöstrand, *Multiple wells in the semiclassical limit.* I, Comm. Partial Differential Equations, 9 (1984), pp. 337–408.

[6] M. Henry, *Singular limit of a fourth order problem arising in the micro-phase separation of diblock copolymers,* Adv. Differential Equations, 6 (2001), pp. 1049–1114.

[7] C.B. Muratov, *Theory of domain patterns in systems with long-range interactions of Coulomb type,* Phys. Rev. E, 66 (2002), paper 066108.

[8] Y. Nishiura and I. Ohnishi, *Some mathematical aspects of the microphase separation in diblock copolymers,* Phys. D, 84 (1995), pp. 31–39.

[9] I. Ohnishi, Y. Nishiura, M. Imai, and Y. Matsushita, *Analytical solutions describing the phase separation driven by a free energy functional containing a long-range interaction term,* Chaos, 9 (1999), pp. 329–341.

[10] T. Ohta and K. Kawasaki, *Equilibrium morphology of block copolymer melts,* Macromolecules, 19 (1986), pp. 2621–2632.

[11] X. Ren and J. Wei, *On energy minimizers of the di-block copolymer problem,* Interfaces Free Bound., to appear.

[12] X. Ren and J. Wei, *Wriggled Lamellar Solutions and Their Stability in the Diblock Copolymer Problem,* preprint.

[13] X. Ren and J. Wei, *On the multiplicity of solutions of two nonlocal variational problems,* SIAM J. Math. Anal., 31 (2000), pp. 909–924.

[14] X. Ren and J. Wei, *Concentrically layered energy equilibria of the di-block copolymer problem,* European J. Appl. Math., 13 (2002), pp. 479–496.
[15] X. Ren and J. Wei, *Triblock copolymer theory: Free energy, disordered phase and weak segregation,* Phys. D, 178 (2003), pp. 103–117.
[16] X. Ren and J. Wei, *Triblock copolymer theory: Ordered ABC lamellar phase,* J. Nonlinear Sci., to appear.

SIAM J. MATH. ANAL.
Vol. 35, No. 1, pp. 33–60

HOMOGENIZATION OF HIGH-CONDUCTIVITY PERIODIC PROBLEMS: APPLICATION TO A GENERAL DISTRIBUTION OF ONE-DIRECTIONAL FIBERS*

MARC BRIANE†

Abstract. This article is devoted to the asymptotic study, as $\varepsilon \to 0$, of the Dirichlet problem

$$\begin{cases} -\operatorname{div}\left(A_\varepsilon(\frac{x}{\varepsilon})\nabla u_\varepsilon\right) = f & \text{in } \Omega, \\ u_\varepsilon = 0 & \text{on } \partial\Omega, \end{cases}$$

where Ω is an x_3-axis bounded open cylinder of $\mathbb{R}^3$, and A_ε is a positive measurable function which does not depend on the variable x_3, periodic with respect to the two-dimensional torus Y_2. The conductivity A_ε is not uniformly bounded in an open set of small measure $Q_\varepsilon \subset Y_2$ and is equal to 1 elsewhere.

We propose a new approach to solving this high-conductivity homogenization problem. It is based on the study of the asymptotic behavior of the periodic spectral problem weighted by the conductivity function A_ε:

$$-\operatorname{div}\left(A_\varepsilon \nabla V_{k,\varepsilon}\right) = \Lambda_k(\varepsilon)\, A_\varepsilon\, V_{k,\varepsilon} \quad \text{in } Y_2, \quad k \in \mathbb{N},$$

where the eigenfunctions $V_{k,\varepsilon}$ are Y_2-periodic.

On the one hand, under suitable conditions on Q_ε we prove that nonlocal effects appear through a coupling in the limit problem if and only if the sequence $(\frac{\Lambda_1(\varepsilon)}{\varepsilon^2})_{\varepsilon>0}$ is bounded, where $\Lambda_1(\varepsilon)$ is the first nonzero eigenvalue of the previous spectral problem.

On the other hand, when Q_ε is composed of N smooth connected open subsets of small diameter, we prove that the limit problem is a coupled system of second order linear PDEs whose size is $n \leq N+1$. The number n is equal to the smallest integer such that the sequence $(\frac{\Lambda_n(\varepsilon)}{\varepsilon^2})_{\varepsilon>0}$ tends to $+\infty$ as ε tends to 0. We illustrate this result by studying the case of $N = 2$ highly conducting cylinders in the period cell of the same radius $r_\varepsilon \ll 1$ and separated by distance $d_\varepsilon > 0$.

Key words. homogenization, periodic microstructure, high conductivity, nonlocal effects, spectrum

AMS subject classifications. 35B27, 35J25, 74Q15, 76M50

DOI. 10.1137/S0036141001398666

1. Introduction.

In this paper we study the asymptotic behavior of a class of conduction problems with nonuniformly bounded coefficients. This class is defined as follows:

Let Ω_2 be a bounded domain of $\mathbb{R}^2$ and let Ω be the open cylinder of $\mathbb{R}^3$ defined by $\Omega := \Omega_2 \times\,]0,1[$ along the x_3-axis. We consider the conduction problem in Ω:

$$\begin{cases} -\operatorname{div}\left(a_\varepsilon \nabla u_\varepsilon\right) = f & \text{in } \Omega, \\ u_\varepsilon = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where f is a given function in $L^2(\Omega)$. The conductivity a_ε is independent of x_3 and is assumed to be a highly oscillating function defined by

$$a_\varepsilon(x) := A_\varepsilon\left(\frac{x}{\varepsilon}\right) \quad \text{for almost every } x \in \Omega, \tag{1.2}$$

*Received by the editors November 22, 2001; accepted for publication (in revised form) November 27, 2002; published electronically June 10, 2003.
http://www.siam.org/journals/sima/35-1/39866.html
†Centre de Mathématiques, I.N.S.A. de Rennes & I.R.M.A.R., 20, avenue des Buttes de Coësmes, CS 14315, 35043 Rennes Cedex, France (mbriane@insa-rennes.fr).

where A_ε is a Y_2-periodic ($Y_2 := \mathbb{R}^2/\mathbb{Z}^2$ is the two-dimensional torus) measurable positive function in $L^\infty(\mathbb{R}^2)$. We assume that there exists an open subset Q_ε of Y_2, of small measure such that

(1.3)
$$A_\varepsilon := 1 \text{ a.e. in } Y_2 \setminus Q_\varepsilon, \quad A_\varepsilon \geq 1 \text{ a.e. in } Q_\varepsilon, \quad \text{and} \quad \int_{Q_\varepsilon} A_\varepsilon \underset{\varepsilon\to 0}{\longrightarrow} \kappa \in]0,+\infty[.$$

Therefore A_ε is uniformly bounded from below by a positive constant but nonuniformly bounded from above.

Problem (1.1) is a conduction problem in an ε-periodic medium composed of a moderately conducting material and a highly conducting one. The model composite medium is a matrix periodically reinforced by highly conducting one-directional fibers. We study the asymptotic behavior of problem (1.1) when the period ε tends to zero. Our aim is to determine the homogenized problem satisfied by the limit u of the solution u_ε of (1.1). In general the limit problem is not a conduction problem like (1.1). It may be a coupled system of linear second order PDEs, which expresses nonlocal effects in the homogenization process. Such nonlocal effects are induced by the interaction between the conduction in the matrix and the conduction in the highly conducting fibers.

Fenchenko and Khruslov [11] (see also [12]) studied similar conduction problems with more general nonuniformly bounded coefficients but for Neumann boundary conditions. Under rather complicated assumptions satisfied by the coefficients, they show that the limit of the conduction problem is a coupled system of linear second order PDEs of size 2 satisfied by the limit u of u_ε in $H^1(\Omega)$ and the limit of the rescaled potential in the small size region of high conductivity.

In terms of the strongly local energy

$$E_\varepsilon(u) := \int_\Omega a_\varepsilon \left|\nabla u\right|^2, \tag{1.4}$$

this coupled system corresponds to the appearance of a nonlocal term in the limit energy. From a theoretical point of view Mosco [13] showed, thanks to the Beurling–Deny [2] representation formula of the Dirichlet forms, that the Γ-limit E_0 of strongly local forms of type (1.4) can always be written, for any $u \in C_0^1(\Omega)$,

$$E_0(u) = \int_\Omega a(dx)\nabla u \cdot \nabla u + \int_\Omega u^2\, k(dx) + \int_{\Omega\times\Omega\setminus\mathrm{diag}} (u(x)-u(y))^2\, j(dx,dy), \tag{1.5}$$

where a, k are Radon measures on Ω and the so-called *jumping measure* j is a Radon measure on $\Omega\times\Omega\setminus\mathrm{diag}$ which represents the nonlocal term. The work of Fenchenko and Khruslov [11] thus gives sufficient conditions in order to obtain nonlocal effects. More recently Camar-Eddine and Seppecher [8] proved that a large class of measures k, j can be attained from the convergence of strongly local energies (1.4). An explicit construction is given by Tchou and the author [5] in order to obtain the product Lebesgue measure $dx \otimes dy$ as a jumping measure j in (1.5)

Revisiting the work of Fenchenko and Khruslov, Bellieud and Bouchitté [1] studied in a complete way (for a nonlinear conduction problem with mixed boundary conditions) an example from [11] by using the Γ-convergence of functionals. This example corresponds to the case where the highly conducting set Q_ε is a disk of radius $r_\varepsilon \ll 1$. They prove that nonlocal effects arise if and only if the limit of $\varepsilon^2 \left|\ln r_\varepsilon\right|$

is positive. Moreover they extend the example of [11] by considering an overlapping of $(n-1)$ highly conducting cylinders of different radius and conductivity. For this geometry they obtain a limit coupled system of size n satisfied by the limit u of u_ε and the limit of the rescaled potentials associated with each cylinder.

In this article we propose a completely different approach to the homogenization in periodic media which are reinforced by highly conducting one-directional fibers. This work extends the results of [11] since we obtain limit problems of arbitrary large size, as well as that of [1] since the distribution of the fibers is quite general. Our approach is based on the asymptotic behavior of the "cross-sectional" period cell spectral problem

$$-\operatorname{div}(A_\varepsilon \nabla V_{k,\varepsilon}) = \Lambda_k(\varepsilon)\, A_\varepsilon\, V_{k,\varepsilon} \quad \text{in } Y_2. \tag{1.6}$$

The solutions of (1.6) make an orthonormal basis $(V_{k,\varepsilon})_{k\geq 0}$ of $L^2_\#(Y_2)$ provided with the weighted Hilbert norm

$$\|V\|_\varepsilon := \left(\int_{Y_2} A_\varepsilon\, V^2\right)^{\frac{1}{2}}, \quad V \in L^2_\#(Y_2), \tag{1.7}$$

and are associated with the sequence of eigenvalues

$$\Lambda_0(\varepsilon) = 0 < \Lambda_1(\varepsilon) \leq \cdots \leq \Lambda_k(\varepsilon) \leq \Lambda_{k+1}(\varepsilon) \leq \cdots. \tag{1.8}$$

In the first section of the paper we prove the following result (see Theorem 2.1 for precise statements).

THEOREM 1.1. *Assume that the eigenfunction $V_{1,\varepsilon}$ is not concentrated in the set Q_ε defined by* (1.3)*, that the capacity of Q_ε tends to zero, and that there exists, for any $\mu \in \mathbb{R}^2$, a smooth Y_2-periodic function whose gradient is equal to μ in Q_ε. Then if the eigenvalues $\Lambda_1(\varepsilon), \Lambda_2(\varepsilon)$ of* (1.6) *satisfy*

$$\lim_{\varepsilon\to 0} \frac{\Lambda_1(\varepsilon)}{\varepsilon^2} \in]0, +\infty[\quad \textit{and} \quad \lim_{\varepsilon\to 0} \frac{\Lambda_2(\varepsilon)}{\varepsilon^2} = +\infty, \tag{1.9}$$

the limit problem of (1.1) *is a coupled system of size* 2.

More precisely the limit coupled system is satisfied by the averaged potential u in the matrix and the averaged (rescaled) potential v in the fibers. Following [1] the function v can be expressed in an integral form of u. The coupled system is thus equivalent to a nonlocal equation satisfied by the limit u (see Remark 2.2).

In [3] we studied the case where $\frac{\Lambda_1(\varepsilon)}{\varepsilon^2}$ tends to $+\infty$ in a more general framework. We then obtain a classical conduction problem without coupling or, equivalently, a strongly local limit energy. Therefore (1.9) can be considered as a necessary and sufficient condition of appearance of nonlocal effects under suitable assumptions on Q_ε.

In the second section of the paper we prove a more general result from the point of view of the asymptotic behavior of the spectrum but more restrictive at the level of the geometry of the highly conducting set Q_ε (see Theorem 2.4 and Figure 1).

THEOREM 1.2. *Assume that Q_ε is composed of an arbitrary number N of two-by-two disjoint open subsets $Q_{i,\varepsilon}$ with small diameter and which satisfy a uniform (with respect to ε) Poincaré–Wirtinger inequality. Then there exists a smallest integer n with $1 \leq n \leq N+1$ such that the sequence $(\frac{\Lambda_n(\varepsilon)}{\varepsilon^2})_{\varepsilon>0}$ is not bounded. Moreover if we assume that, for this particular integer n, the eigenvalues of problem* (1.6) *satisfy*

$$\forall\, k = 1, \ldots, n-1, \ \lim_{\varepsilon\to 0} \frac{\Lambda_k(\varepsilon)}{\varepsilon^2} \in]0, +\infty[\quad \textit{and} \quad \lim_{\varepsilon\to 0} \frac{\Lambda_n(\varepsilon)}{\varepsilon^2} = +\infty, \tag{1.10}$$

the limit problem of (1.1) *is then a coupled system of size n.*

Contrary to Theorem 1.1 the averaged potential in the fibers has several $(n-1)$ components which depend on the number N of fibers by period cell. In general the number of components is not equal to N, but it is equal to the number of eigenvalues of order ε^2 of the spectral problem (1.6) according to (1.10). In fact the spectrum connected with (1.6) contains the geometrical information of the distribution of the fibers, which controls the degree of coupling of the limit problem (see Remark 2.5).

In Example 2.6 we illustrate the previous result by considering the case of $N=2$ disks $Q_{1,\varepsilon}, Q_{2,\varepsilon}$ of the same radius r_ε and separated by distance $d_\varepsilon \in]0,1[$. According to the asymptotic behavior of $\varepsilon^2 \left|\ln r_\varepsilon\right|$ and $\frac{\ln d_\varepsilon}{\ln r_\varepsilon}$, we obtain a limit system of size $n = 1, 2,$ or 3 (see Proposition 2.7).

Theorem 1.2 extends the result of [1] to a general periodic distribution of highly conducting one-directional fibers whose sections are the sets $Q_{i,\varepsilon}$ in the period cell. This result shows that the limit problem and in particular its size are completely determined by the asymptotic behavior of the spectrum (1.6) and not by the number of fibers by period cell.

We have already used a similar spectral approach in [4] for low-conductivity problems with isolating regions. In the present context the geometrical assumptions are quite different, and contrary to [4] the reference spectral problem (1.6) is weighted by the conductivity function A_ε. However the conclusion concerning the spectral approach is the same. In the periodic framework the asymptotic behavior of the spectrum (1.6) is a good tool to measure the nonlocal effects arising in fiber-reinforced media.

2. Statement of the results.

2.1. Appearance of nonlocal effects. We denote by (e_1, e_2, e_3) the canonic basis of $\mathbb{R}^3$ and by $|E|$ the Lebesgue measure of any measurable subset E of $\mathbb{R}^2$ or $\mathbb{R}^3$.

Let Ω_2 be a smooth bounded open subset of $\mathbb{R}^2$ and let $\Omega := \Omega \times]0,1[$ be the cylinder of $\mathbb{R}^2$, parallel to e_3, of bottom Ω and height 1.

Let Y_2 be the two-dimensional torus identified as $[0,1[^2$. We denote by $L^2_\#(Y_2)$, resp., $H^1_\#(Y_2)$, the set of the Y_2-periodic functions which are locally in $L^2(\mathbb{R}^2)$, resp., $H^1(\mathbb{R}^2)$.

For any $\varepsilon > 0$ let $A_\varepsilon : \mathbb{R}^2 \longrightarrow \mathbb{R}^*_+$ be a positive measurable function Y_2-periodic, i.e.,

$$A_\varepsilon(y_1+1, y_2) = A_\varepsilon(y_1, y_2+1) = A_\varepsilon(y) \quad \text{for almost every } y = (y_1, y_2) \in \mathbb{R}^2, \tag{2.1}$$

and bounded from above and below by positive constants

$$\operatorname*{infess}_{\mathbb{R}^2} A_\varepsilon = 1 \quad \text{and} \quad \operatorname*{supess}_{\mathbb{R}^2} A_\varepsilon = \beta_\varepsilon < +\infty. \tag{2.2}$$

We assume that there exists an open subset Q_ε of Y_2 such that

$$A_\varepsilon := 1 \text{ a.e. in } Y_2 \setminus Q_\varepsilon, \quad \lim_{\varepsilon\to 0} |Q_\varepsilon| = 0, \quad \text{and} \quad \lim_{\varepsilon\to 0} \int_{Q_\varepsilon} A_\varepsilon = \kappa \in]0, +\infty[. \tag{2.3}$$

We extend A_ε to $\mathbb{R}^3$ by setting $A_\varepsilon(y_1, y_2, y_3) := A_\varepsilon(y_1, y_2)$, and we define the rescaled function

$$a_\varepsilon(x) := A_\varepsilon\left(\frac{x}{\varepsilon}\right) \quad \text{for almost every } x \in \Omega. \tag{2.4}$$

Our aim is to study the asymptotic behavior of the high-conductivity problem

$$\begin{cases} -\operatorname{div}(a_\varepsilon \nabla u_\varepsilon) = f & \text{in } \Omega, \\ u_\varepsilon = 0 & \text{on } \partial\Omega, \end{cases} \tag{2.5}$$

where f is a given function in $L^2(\Omega)$.

We will prove that the asymptotic behavior of the Dirichlet problem (2.5) is completely determined by the asymptotic behavior of the spectrum

$$\Lambda_0(\varepsilon) = 0 < \Lambda_1(\varepsilon) \le \Lambda_2(\varepsilon) \le \cdots \tag{2.6}$$

associated with the periodic eigenfunctions $V_{k,\varepsilon} \in H^1_\#(Y_2)$, $k \in \mathbb{N}$, solutions of

$$-\operatorname{div}(A_\varepsilon \nabla V_{k,\varepsilon}) = \Lambda_k(\varepsilon)\, A_\varepsilon\, V_{k,\varepsilon} \text{ in } Y_2 \quad \text{and} \quad \int_{Y_2} A_\varepsilon\, V^2_{k,\varepsilon} = 1. \tag{2.7}$$

In this section we are only interested in the appearance of nonlocal effects. In [3] we proved, in a more general way, that the limit of $\frac{\Lambda_1(\varepsilon)}{\varepsilon^2}$ is a critical barrier for nonlocal effects. In the present context we will prove that nonlocal effects appear if and only if this limit is finite.

The main result of the section is the following.

THEOREM 2.1. *Assume that*

$$\lim_{\varepsilon\to 0} \frac{\Lambda_1(\varepsilon)}{\varepsilon^2} = \lambda_1 \in]0, +\infty[\quad \textit{and} \quad \lim_{\varepsilon\to 0} \frac{\Lambda_2(\varepsilon)}{\varepsilon^2} = +\infty, \tag{2.8}$$

and that the eigenfunction $V_{1,\varepsilon}$ associated with $\Lambda_1(\varepsilon)$ by (2.7) satisfies

$$\lim_{\varepsilon\to 0} \int_{Y_2} V^2_{1,\varepsilon} > 0. \tag{2.9}$$

Assume that the function A_ε and the set Q_ε satisfy conditions (2.1) to (2.3). Also assume that Q_ε has a vanishing capacity in the torus, i.e.,

$$\operatorname{cap}(Q_\varepsilon) := \inf\left\{ \int_{Y_2} |\nabla\Phi|^2 \;\Big|\; \Phi \in H^1_\#(Y_2),\ \Phi = 0 \text{ in } Q_\varepsilon,\ \textit{and} \int_{Y_2} \Phi = 1 \right\} \underset{\varepsilon\to 0}{\longrightarrow} 0, \tag{2.10}$$

and that there exists, for any $\mu \in \mathbb{R}^2$, a periodic function Ψ_μ such that

$$\Psi_\mu \in C^1_\#(Y_2) \quad \textit{and} \quad \nabla\Psi_\mu = \mu \text{ in } Q_\varepsilon. \tag{2.11}$$

Then the solution u_ε of problem (2.5) weakly converges in $H^1(\Omega)$ to the solution u of the coupled system

$$\begin{cases} -\Delta u - a\dfrac{\partial^2 u}{\partial x_3^2} + \gamma(u - v) = f & \text{in } \Omega, \\ -b\dfrac{\partial^2 v}{\partial x_3^2} + \gamma(v - u) = 0 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \\ v(x_1, x_2, 0) = v(x_1, x_2, 1) = 0 & \text{in } \Omega_2, \end{cases} \tag{2.12}$$

where

$$c_1 := \lim_{\varepsilon\to 0} \int_{Y_2} V_{1,\varepsilon}, \tag{2.13}$$

$$a := \frac{\kappa - (1+\kappa)\, c_1^2}{1 + (1+\kappa)\, c_1^2}, \quad b := \frac{(1+\kappa)^2\, c_1^2}{1 + (1+\kappa)\, c_1^2}, \quad \gamma := \lambda_1 \left(\frac{(1+\kappa)\, c_1}{1 + (1+\kappa)\, c_1^2} \right)^2 . \tag{2.14}$$

Remark 2.2. In the homogenized system (2.12) the function u represents the averaged potential in the matrix and v the averaged (rescaled) potential in the fibers. The two potentials are coupled by the second equation of (2.12) thanks to the capacitary parameter γ. If $\gamma < +\infty$, by following Bellieud and Bouchitté [1] this equation also reads as

$$v(x_1, x_2, x_3) = \int_0^1 G(x_3, t)\, u(x_1, x_2, t)\, dt, \tag{2.15}$$

where the kernel G can be explicitly computed. By substituting v by (2.15) in the first equation of (2.12) one obtains the nonlocal limit equation satisfied by u:

$$-\Delta u - a\, \frac{\partial^2 u}{\partial x_3^2} + \gamma \left(u - \int_0^1 G(x_3, t)\, u(x_1, x_2, t)\, dt \right) = f. \tag{2.16}$$

We proved in [3] that if $\gamma = +\infty$, the averaged potentials u and v are equal. The limit equation then reads as

$$-\Delta u - (a+b)\, \frac{\partial^2 u}{\partial x_3^2} = f, \tag{2.17}$$

in which only the residual conductivity $(a+b)$ attests to the reinforcement by fibers.

For example (see [11], [1], or [3]), when Q_ε is a disk of small radius r_ε the coupling and thus the nonlocal effect appears if and only if r_ε is sufficiently small, namely, $|\ln r_\varepsilon| \geq c\, \varepsilon^{-2}$, where c is a positive constant.

In a more general framework we proved in [3] that there is no coupling and hence no nonlocal effect if the limit of $\frac{\Lambda_1(\varepsilon)}{\varepsilon^2}$ is infinite. Under assumptions (2.10) and (2.11) we can thus claim that a nonlocal effect appears if and only if the sequence $\frac{\Lambda_1(\varepsilon)}{\varepsilon^2}$ is bounded.

Under assumption (2.8) we obtain a coupled system of two equations. In the next section we will obtain a coupled system of arbitrary size under a more general assumption on the spectrum.

2.2. A general distribution of one-directional fibers. In this section we consider a general distribution of $N \geq 1$ highly conducting x_3-directional fibers in the period cell. More precisely the fibers' cross sections are represented in the two-dimensional torus Y_2 by the set

$$Q_\varepsilon := \bigcup_{i=1}^{N} Q_{i,\varepsilon}, \tag{2.18}$$

where $Q_{i,\varepsilon}$ are disjoint regular connected open subsets of Y_2. The closures of two sets $Q_{i,\varepsilon}, Q_{j,\varepsilon}$ may have a nonempty intersection as shown in Figure 1.

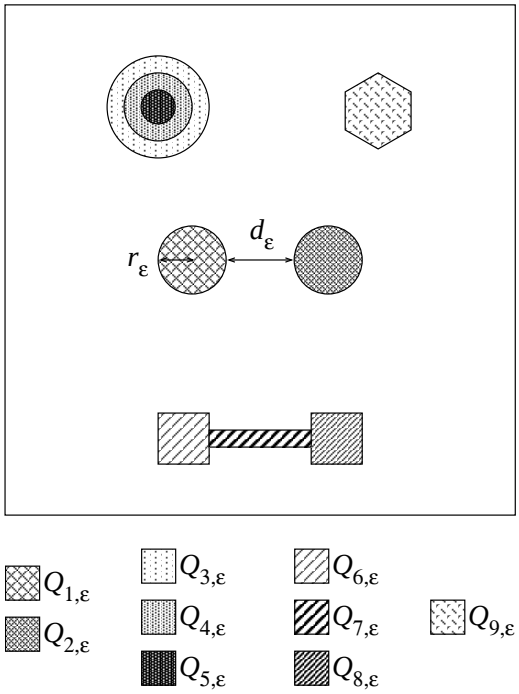


FIG. 1. *The cross section of the period cell with $N = 9$ fibers.*

Moreover we assume that, for any $i = 1, \dots, N$,

$$\lim_{\varepsilon \to 0} \operatorname{diam}(Q_{i,\varepsilon}) = 0, \tag{2.19}$$

and that there exists a positive constant C_i such that the following uniform Poincaré–Wirtinger inequality holds true:

$$\forall\, V \in H^1_{\#}(Q_{i,\varepsilon}), \quad \left\| V - \fint_{Q_{i,\varepsilon}} V \right\|_{L^2(Q_{i,\varepsilon})} \leq C_i \, \|\nabla V\|_{L^2(Q_{i,\varepsilon})}. \tag{2.20}$$

The conductivity of the fiber-reinforced medium is defined by

$$A_\varepsilon := \begin{cases} 1 & \text{in } Y_2 \setminus Q_\varepsilon, \\ \alpha_{i,\varepsilon} & \text{in } Q_{i,\varepsilon}, \end{cases} \qquad \text{where} \quad \lim_{\varepsilon \to 0} \alpha_{i,\varepsilon} \, |Q_{i,\varepsilon}| = \kappa_i \in \,]0, +\infty[. \tag{2.21}$$

We also set

$$\kappa := \kappa_1 + \cdots + \kappa_N, \quad \text{where} \quad \kappa_i := \lim_{\varepsilon \to 0} \alpha_{i,\varepsilon} \, |Q_{i,\varepsilon}|. \tag{2.22}$$

This geometry is more restrictive than in the first section, but we will consider a more general assumption for the asymptotic behavior of the spectrum (2.7).

Before stating the result we need some notation.

Notation 2.3. Let us define the following weighted scalar product and norm in $L^2_\#(Y_2)$:

$$\forall\, V, W \in L^2_\#(Y_2), \quad \langle V, W\rangle_\varepsilon := \int_{Y_2} A_\varepsilon\, V\, W \quad \text{and} \quad \|V\|_\varepsilon := \sqrt{\langle V, V\rangle_\varepsilon}. \tag{2.23}$$

We extend in $L^2(\Omega)$ by

$$\forall\, v, w \in L^2(\Omega), \quad \langle v, w\rangle_\varepsilon := \int_\Omega a_\varepsilon\, v\, w \quad \text{and} \quad \|v\|_\varepsilon := \sqrt{\langle v, v\rangle_\varepsilon}. \tag{2.24}$$

Let us denote by $\approx_\varepsilon$ the following approximation in $L^2_\#(Y_2)$:

$$\forall\, V_\varepsilon, W_\varepsilon \in L^2_\#(Y_2), \quad V_\varepsilon \approx_\varepsilon W_\varepsilon \ \text{ if } \ \lim_{\varepsilon\to 0} \|V_\varepsilon - W_\varepsilon\|_\varepsilon = 0. \tag{2.25}$$

We extend in $L^2(\Omega)$ by

$$\forall\, v_\varepsilon, w_\varepsilon \in L^2(\Omega), \quad v_\varepsilon \approx_\varepsilon w_\varepsilon \ \text{ if } \ \lim_{\varepsilon\to 0} \int_\Omega a_\varepsilon\, (v_\varepsilon - w_\varepsilon)^2 = 0. \tag{2.26}$$

Let $T_{0,1}$ be the truncature defined by

$$T_{0,1}(t) := \frac{1}{2}\,(1 + |t| - |t-1|) = \begin{cases} t & \text{if } t \in [0,1], \\ 0 & \text{if } t < 0, \\ 1 & \text{if } t > 1. \end{cases} \tag{2.27}$$

In what follows we will denote in the same way any function v defined on $\mathbb{R}^2$ and its extension to $\mathbb{R}^3$, i.e., $v(x_1, x_2, x_3) = v(x_1, x_2)$.

Let us now state the main result of this section.

THEOREM 2.4. *Let Q_ε be the set defined by* (2.18) *and conditions* (2.19), (2.20), *and let A_ε be the conductivity defined by* (2.21).

(i) *Then there exists a smallest integer n with*

$$1 \le n \le N+1 \tag{2.28}$$

such that the sequence $(\frac{\Lambda_n(\varepsilon)}{\varepsilon^2})_{\varepsilon>0}$ is not bounded.

We assume that, for this particular integer n, the eigenvalues of problem (2.7) *satisfy*

$$\forall\, k = 1, \dots, n-1, \ \lim_{\varepsilon\to 0} \frac{\Lambda_k(\varepsilon)}{\varepsilon^2} = \lambda_k \in\,]0, +\infty[\quad \textit{and} \quad \lim_{\varepsilon\to 0} \frac{\Lambda_n(\varepsilon)}{\varepsilon^2} = +\infty. \tag{2.29}$$

Let $(V_{0,\varepsilon}, \dots, V_{n-1,\varepsilon})$ ($V_{0,\varepsilon} := \|1\|_\varepsilon^{-1}$) be a family of associated eigenfunctions which is orthonormal with respect to the scalar product $\langle\cdot,\cdot\rangle_\varepsilon$ defined by (2.23).

Then there exists an "asymptotic" partition of the unity $(\hat V_{1,\varepsilon}, \dots, \hat V_{n,\varepsilon})$ composed by n functions in $H^1_\#(Y_2)$ which satisfy the following properties:

$$\forall\, k := 1, \dots, n, \ \ \hat V_{k,\varepsilon} \rightharpoonup \delta_{k,1} \ \textit{ weakly in } H^1_\#(Y_2) \quad \textit{and} \quad \hat V_{1,\varepsilon} + \cdots + \hat V_{n,\varepsilon} \ \approx_\varepsilon \ 1, \tag{2.30}$$

where $\delta_{k,1}$ *is the Kronecker symbol and* $\approx_\varepsilon$ *is defined by* (2.25),

$$\forall\, h \neq k,\ \lim_{\varepsilon\to 0} \langle \hat{V}_{h,\varepsilon}, \hat{V}_{k,\varepsilon}\rangle_\varepsilon = 0 \quad \textit{and} \quad \lim_{\varepsilon\to 0} \|\hat{V}_{k,\varepsilon}\|_\varepsilon > 0, \tag{2.31}$$

$$\forall\, k = 1,\dots,n, \quad T_{0,1}(\hat{V}_{k,\varepsilon}) \approx_\varepsilon \hat{V}_{k,\varepsilon}, \tag{2.32}$$

where $T_{0,1}$ *is defined by* (2.27), *and*

$$\forall\, i = 0,\dots,n-1, \quad V_{i,\varepsilon} = \sum_{k=1}^{n} \hat{c}_{i,k}\hat{V}_{k,\varepsilon}, \quad \textit{where } \hat{c}_{i,k} \in \mathbb{R}. \tag{2.33}$$

(ii) *The solution* u_ε *of the Dirichlet problem* (2.5), *defined with the conductivity* A_ε *of* (2.21), *weakly converges in* $H^1(\Omega)$ *to the solution* u *of the coupled system of size* n*:*

(2.34)
$$\begin{cases} -\Delta u - (\hat{\kappa}_1 - 1)\,\dfrac{\partial^2 u}{\partial x_3^2} + \displaystyle\sum_{i=1}^{n-1} \hat{\kappa}_1\hat{c}_{i,1}\lambda_i \left(\hat{\kappa}_1\hat{c}_{i,1}\,u + \sum_{h=2}^{n} \hat{\kappa}_h\hat{c}_{i,h}\,v_h\right) = f & \textit{in } \Omega, \\ -\hat{\kappa}_k\,\dfrac{\partial^2 v_k}{\partial x_3^2} + \displaystyle\sum_{i=1}^{n-1} \hat{\kappa}_k\hat{c}_{i,k}\lambda_i \left(\hat{\kappa}_k\hat{c}_{i,k}\,u + \sum_{h=2}^{n} \hat{\kappa}_h\hat{c}_{i,h}\,v_h\right) = 0 & \textit{in } \Omega, \\ u = 0 & \textit{on } \Omega, \\ v_k(x_1,x_2,0) = v_k(x_1,x_2,1) = 0 & \textit{in } \Omega_2, \end{cases}$$

$$\textit{where} \quad \forall\, k = 1,\dots,n, \ \ \hat{\kappa}_k := \lim_{\varepsilon\to 0} \langle \hat{V}_{k,\varepsilon}, 1\rangle_\varepsilon = \lim_{\varepsilon\to 0} \|\hat{V}_{k,\varepsilon}\|_\varepsilon^2. \tag{2.35}$$

Remark 2.5. In the homogenized system (2.34) and contrary to the case of Theorem 2.1, the averaged potential in the fibers is divided into several potentials $v_2,\dots,v_{n-1}$. Each function v_k corresponds to the averaged contribution of a group of fibers among the N fibers in the period cell. The number $(n-1)$ of groups is not necessarily equal to the number N of fibers by cell, but it is controlled by the rescaled spectrum $(\frac{\Lambda_k(\varepsilon)}{\varepsilon^2})_{k\in\mathbb{N}}$ of problem (2.7) according to condition (2.29). Of course the spectrum (2.7) strongly depends on the geometry of the distribution of the fibers but, inversely, it contains implicitly the information of geometrical nature. That is the main interest of this spectral approach.

For instance, Example 2.6 below shows that the degree of coupling in the limit problem depends in particular on the closeness of the fibers in the period cell. So two fibers in the period cell ($N = 2$) will induce the same averaged potential in the fibers ($n = 1$) if they are sufficiently neighboring, and two different averaged potentials ($n = 2$) elsewhere.

Example 2.6. The case of $N = 2$ neighboring fibers.

We consider the case where the set Q_ε is composed of two disjoint open disks $Q_{1,\varepsilon}, Q_{2,\varepsilon}$ of the same radius $r_\varepsilon \ll 1$ and separated by distance $d_\varepsilon \in\,]0,1[$, as shown in Figure 1. Then the size n of the limit system is determined by the following result.

Proposition 2.7. (i) *Assume that*

$$\lim_{\varepsilon\to 0} \left(\frac{2\,\pi}{\varepsilon^2\,|\ln r_\varepsilon|}\right) = \delta \in\,]0,+\infty] \quad \textit{and} \quad \lim_{\varepsilon\to 0} \left|\frac{\ln d_\varepsilon}{\ln r_\varepsilon}\right| = \gamma \in [0,+\infty]. \tag{2.36}$$

Then we have

$$\begin{cases} \delta = +\infty & \Leftrightarrow \quad n = 1, \\ \delta < +\infty \text{ and } \gamma \geq 1 & \Rightarrow \quad n = 2, \\ \delta < +\infty \text{ and } \gamma < 1 & \Rightarrow \quad n = 3. \end{cases} \tag{2.37}$$

(ii) *In particular assume that* $\delta < +\infty$, $\gamma = 0$, *and* $\kappa_1 = \kappa_2$ *defined in* (2.22). *Then the asymptotic behavior of the eigenvalues* $\Lambda_1(\varepsilon), \Lambda_2(\varepsilon)$ *of problem* (2.7) *is given by*

$$\lambda_1 = \lim_{\varepsilon \to 0} \frac{\Lambda_1(\varepsilon)}{\varepsilon^2} = \frac{\delta}{\kappa_1} \quad \text{and} \quad \lambda_2 = \lim_{\varepsilon \to 0} \frac{\Lambda_2(\varepsilon)}{\varepsilon^2} = \frac{\delta}{\kappa_1}\,(1 + 2\,\kappa_1), \tag{2.38}$$

and the limit system (2.34) *reads as*

$$\begin{cases} -\,\Delta u + 2\,\delta\,u - \delta\,v_2 - \delta\,v_3 \;= f & \text{in } \Omega, \\ -\,\kappa_1\,\dfrac{\partial^2 v_2}{\partial x_3^2} - \delta\,u + \delta\,v_2 \;= 0 & \text{in } \Omega, \\ -\,\kappa_1\,\dfrac{\partial^2 v_3}{\partial x_3^2} - \delta\,u + \delta\,v_3 \;= 0 & \text{in } \Omega. \end{cases} \tag{2.39}$$

3. Proof of the results.

3.1. Proof of Theorem 2.1. The proof of Theorem 2.1 is divided into three steps. In the first step we build an asymptotic partition of the unity in $H^1_\#(Y_2)$, $(\hat{V}_{1,\varepsilon}, 1 - \hat{V}_{1,\varepsilon})$, which is characterized by Lemma 3.1 below. This partition is the key ingredient of the proof and induces the coupling (2.12). It is also related to the spectral problem (2.7) since both functions $\hat{V}_{1,\varepsilon}$ and $1-\hat{V}_{1,\varepsilon}$ span the same space as the first eigenfunctions 1 and $V_{1,\varepsilon}$ associated with the eigenvalues $\Lambda_0(\varepsilon) = 0$ and $\Lambda_1(\varepsilon)$. We then define the rescaled functions

$$\hat{v}_{1,\varepsilon}(x) := \hat{V}_{1,\varepsilon}\left(\frac{x}{\varepsilon}\right) \quad \text{and} \quad \hat{v}_{2,\varepsilon}(x) := 1 - \hat{V}_{1,\varepsilon}\left(\frac{x}{\varepsilon}\right) \quad \text{for almost every } x \in \Omega. \tag{3.1}$$

In the second step we give the weak $*$ limits of $a_\varepsilon\,\hat{v}_{i,\varepsilon}\,u_\varepsilon$ and $a_\varepsilon\,\hat{v}_{i,\varepsilon}\,\nabla u_\varepsilon$, $i = 1, 2$, in terms of the solutions u and v of (2.12). In the third step we determine the limit problem (2.12) using $\hat{v}_{1,\varepsilon}$ (3.1) as a test function in problem (2.5). In particular we prove in Lemma 3.2 below that $a_\varepsilon\,\nabla\hat{v}_{1,\varepsilon}\,u_\varepsilon$ weakly $*$ tends to 0 in the distributions sense thanks to the assumptions (2.10) and (2.11) satisfied by the set Q_ε.

In the following all the limits hold true up to a subsequence of ε still denoted ε for the sake of simplicity. We will also denote by c an arbitrary positive constant.

First step. Partition of the unity $(\hat{V}_{1,\varepsilon}, 1-\hat{V}_{1,\varepsilon})$. This step is contained in the following result.

LEMMA 3.1. *Let* $\hat{V}_{1,\varepsilon}$ *be the function defined by*

$$\hat{V}_{1,\varepsilon} := \alpha + \beta\,V_{1,\varepsilon} \quad \textit{with} \quad \alpha := \frac{1}{1 + (1+\kappa)\,c_1^2} \quad \textit{and} \quad \beta := \frac{(1+\kappa)\,c_1}{1 + (1+\kappa)\,c_1^2}, \tag{3.2}$$

where $V_{1,\varepsilon}$ *is defined by* (2.7), κ *by* (2.3), *and* c_1 *by* (2.13). *Then the function* $\hat{V}_{1,\varepsilon}$ *satisfies the following properties (see Notation* 2.3.*):*

$$\hat{V}_{1,\varepsilon} \rightharpoonup 1 \;\; H^1_\#(Y_2) \textit{ weak } *, \quad \hat{V}_{1,\varepsilon} \not\approx_\varepsilon 1, \quad T_{0,1}(\hat{V}_{1,\varepsilon}) \approx_\varepsilon \hat{V}_{1,\varepsilon}, \quad \langle \hat{V}_{1,\varepsilon}, 1-\hat{V}_{1,\varepsilon} \rangle_\varepsilon \to 0. \tag{3.3}$$

Proof of Lemma 3.1. By assumption (2.8) the eigenfunction $V_{1,\varepsilon}$ associated with the eigenvalue $\Lambda_1(\varepsilon)$ by (2.7) satisfies

$$\int_{Y_2} |\nabla V_{1,\varepsilon}|^2 \le \int_{Y_2} A_\varepsilon \, |\nabla V_{1,\varepsilon}|^2 = \Lambda_1(\varepsilon) = O(\varepsilon^2) \ \underset{\varepsilon\to 0}{\longrightarrow}\ 0.$$

Then by the Poincaré–Wirtinger inequality in Y_2, the sequence $V_{1,\varepsilon}$ strongly converges in $H^1_\#(Y_2)$ to a constant which is equal to the limit of its averaged value, i.e., c_1 by (2.13). Since $V^2_{1,\varepsilon}$ strongly converges to c_1^2 in $L^1(Y_2)$, assumption (2.9) implies that $c_1 \neq 0$. We can assume that $c_1 > 0$ even if it means replacing $V_{1,\varepsilon}$ by $-V_{1,\varepsilon}$. This choice uniquely determines the eigenfunction $V_{1,\varepsilon}$ since its multiplicity is equal to 1 by (2.8). Therefore we obtain the following by the definitions of (3.2) and the orthogonality of the eigenfunctions 1, $V_{1,\varepsilon}$ with respect to the scalar product $\langle\cdot,\cdot\rangle_\varepsilon$ of (2.23):

$$\hat V_{1,\varepsilon} \rightharpoonup \alpha + \beta\, c_1 = 1 \ \text{ weakly in } H^1_\#(Y_2),$$

$$\|\hat V_{1,\varepsilon} - 1\|_\varepsilon = (\alpha-1)^2 \int_{Y_2} A_\varepsilon + \beta^2 \ \underset{\varepsilon\to 0}{\longrightarrow}\ (1{+}\kappa)\,(\alpha-1)^2 + \beta^2 > 0,$$

$$\langle \hat V_{1,\varepsilon}, 1{-}\hat V_{1,\varepsilon}\rangle_\varepsilon = \alpha\,(1-\alpha) \int_{Y_2} A_\varepsilon - \beta^2 \ \underset{\varepsilon\to 0}{\longrightarrow}\ (1{+}\kappa)\,(\alpha-\alpha^2) - \beta^2 = 0.$$

It remains to prove that $T_{0,1}(\hat V_{1,\varepsilon}) \approx_\varepsilon \hat V_{1,\varepsilon}$. Let $V \in H^1_\#(Y_2)$. By the Courant–Fisher formulas we have

$$\left\| V - \frac{\langle V,1\rangle_\varepsilon}{\langle 1,1\rangle_\varepsilon}\, 1 - \langle V, V_{1,\varepsilon}\rangle_\varepsilon\, V_{1,\varepsilon} \right\|^2_\varepsilon \le \frac{1}{\Lambda_2(\varepsilon)}\ \|\nabla V - \langle V, V_{1,\varepsilon}\rangle_\varepsilon\, \nabla V_{1,\varepsilon}\|^2_\varepsilon . \tag{3.4}$$

Then for any Lipschitz function $T:\mathbb{R}\longrightarrow\mathbb{R}$ and for any function $V_\varepsilon \in \mathrm{Span}\,(1, V_{1,\varepsilon})$ we have $|\nabla T(V_\varepsilon)| \le c_T\,|\nabla V_\varepsilon|$ and $\|\nabla V_{1,\varepsilon}\|^2_\varepsilon \le \Lambda_1(\varepsilon)$, whence by (2.8)

$$\left\| T(V_\varepsilon) - \frac{\langle T(V_\varepsilon),1\rangle_\varepsilon}{\langle 1,1\rangle_\varepsilon}\, 1 - \langle T(V_\varepsilon), V_{1,\varepsilon}\rangle_\varepsilon\, V_{1,\varepsilon} \right\|^2_\varepsilon \le c\,\frac{\Lambda_1(\varepsilon)}{\Lambda_2(\varepsilon)} \ \underset{\varepsilon\to 0}{\longrightarrow}\ 0.$$

Therefore we obtain (up to a subsequence)

$$T(V_\varepsilon) \approx_\varepsilon a + b\, V_{1,\varepsilon}, \quad \text{where} \quad a := \lim_{\varepsilon\to 0} \frac{\langle T(V_\varepsilon),1\rangle_\varepsilon}{\langle 1,1\rangle_\varepsilon} \ \text{ and } \ b := \lim_{\varepsilon\to 0} \langle T(V_\varepsilon), V_{1,\varepsilon}\rangle_\varepsilon. \tag{3.5}$$

In particular there exist two constants $a_1, b_1 \in \mathbb{R}$ such that $|\hat V_{1,\varepsilon}| \approx_\varepsilon a_1 + b_1\, V_{1,\varepsilon}$. Since $\hat V_{1,\varepsilon}$ strongly converges to 1 in $L^2_\#(Y_2)$ we have

$$1 = a_1 + b_1\, c_1 = \alpha + \beta\, c_1.$$

Moreover the equality $|\hat V_{1,\varepsilon}|^2 = \hat V^2_{1,\varepsilon}$ yields

$$\int_{Y_2} A_\varepsilon\,(a_1 + b_1\, V_{1,\varepsilon})^2 + o(1) = \int_{Y_2} A_\varepsilon\,(\alpha + \beta\, V_{1,\varepsilon})^2;$$

then by passing to the limit we obtain

$$(1{+}\kappa)\, a_1^2 + b_1^2 = (1{+}\kappa)\,\alpha^2 + \beta^2 = \frac{1{+}\kappa}{1 + (1{+}\kappa)\, c_1^2}.$$

We deduce from both previous equalities that b_1 is a solution of

$$\big(1+(1+\kappa)\,c_1^2\big)\,b_1^2-2\,(1+\kappa)\,c_1\,b_1+\frac{(1+\kappa)^2\,c_1^2}{1+(1+\kappa)\,c_1^2}=0,$$

which has a unique solution β, whence $b_1=\beta$ and $a_1=\alpha$. We thus obtain $|\hat V_{1,\varepsilon}|\approx_\varepsilon \hat V_{1,\varepsilon}$.

Similarly there exist two constants $a_2,b_2\in\mathbb{R}$ such that $|1-\hat V_{1,\varepsilon}|\approx_\varepsilon a_2+b_2\,V_{1,\varepsilon}$ with

$$a_2+b_2\,c_1=0\quad\text{and}\quad(1+\kappa)\,a_2^2+b_2^2=(1+\kappa)\,(\alpha-1)^2+\beta^2=\frac{(1+\kappa)^2\,c_1^2}{1+(1+\kappa)\,c_1^2},$$

whence $b_2=\pm\beta$. Moreover by applying (3.5) with $V_\varepsilon:=|1-\hat V_{1,\varepsilon}|$, we obtain $a_2\geq 0$ and thus $b_2\leq 0$, whence $b_2=-\beta$ and $a_2=c_1\,\beta=1-\alpha$. Therefore we have $|1-\hat V_{1,\varepsilon}|\approx_\varepsilon 1-\hat V_{1,\varepsilon}$ and $|\hat V_{1,\varepsilon}|\approx_\varepsilon \hat V_{1,\varepsilon}$, which imply $T_{0,1}(\hat V_{1,\varepsilon})\approx_\varepsilon \hat V_{1,\varepsilon}$ and (3.3). This concludes the proof of Lemma 3.1 and the first step.

Second step. Limits of $a_\varepsilon\,\hat v_{i,\varepsilon}\,u_\varepsilon$ and $a_\varepsilon\,\hat v_{i,\varepsilon}\,\nabla u_\varepsilon$, $i=1,2$.

Since the function a_ε defined by (2.4) does not depend on the variable x_3 and u_ε is equal to 0 on $\Omega_2\times\{0,1\}$ the following estimate holds:

$$\int_\Omega a_\varepsilon\,u_\varepsilon^2\leq c\int_\Omega a_\varepsilon\left(\frac{\partial u_\varepsilon}{\partial x_3}\right)^2\leq c\int_\Omega a_\varepsilon\,|\nabla u_\varepsilon|^2.$$

By the Poincaré inequality

$$\int_\Omega a_\varepsilon\,|\nabla u_\varepsilon|^2=\int_\Omega f\,u_\varepsilon\leq\|f\|_{L^2(\Omega)}\,\|u_\varepsilon\|_{L^2(\Omega)}\leq c\,\|\nabla u_\varepsilon\|_{L^2(\Omega)}\leq c\left(\int_\Omega a_\varepsilon\,|\nabla u_\varepsilon|^2\right)^{\frac12},$$

whence the following a priori estimate:

$$\int_\Omega a_\varepsilon\,u_\varepsilon^2+\int_\Omega a_\varepsilon\,|\nabla u_\varepsilon|^2\leq c. \tag{3.6}$$

By definitions (3.1), (2.26) and by the third property of (3.3) we have $T_{0,1}(\hat v_{i,\varepsilon})\approx_\varepsilon \hat v_{i,\varepsilon}$ for $i=1,2$, which combined with estimate (3.6) and assumption (2.3) implies that the sequence $a_\varepsilon\,\hat v_{i,\varepsilon}\,u_\varepsilon$ is bounded in $L^1(\Omega)$. In the beginning of the proof of Lemma 3.1 we showed that the constant c_1 of (2.13) is not equal to 0, whence the constant b in (2.14) is positive. We can then define a Radon measure v by the weak convergence

$$a_\varepsilon\,\hat v_{2,\varepsilon}\,u_\varepsilon\rightharpoonup b\,v\quad\text{weakly in }\mathcal{M}(\overline\Omega). \tag{3.7}$$

In the following we will prove that the weak limits u of u_ε in $H^1(\Omega)$ and v satisfy

$$v\in H_0^1(0,1;L^2(\Omega_2)), \tag{3.8}$$

$$a_\varepsilon\,\hat v_{1,\varepsilon}\,u_\varepsilon\rightharpoonup(1+\kappa)\,\alpha\,u\quad\text{weakly in }\mathcal{D}'(\Omega), \tag{3.9}$$

$$a_\varepsilon\,\hat v_{1,\varepsilon}\,\nabla u_\varepsilon\rightharpoonup\nabla u+a\,\frac{\partial u}{\partial x_3}\,e_3\quad\text{weakly in }\mathcal{D}'(\Omega;\mathbb{R}^3), \tag{3.10}$$

$$a_\varepsilon\,\hat v_{2,\varepsilon}\,\nabla u_\varepsilon\rightharpoonup b\,\frac{\partial v}{\partial x_3}\,e_3\quad\text{weakly in }\mathcal{D}'(\Omega;\mathbb{R}^3), \tag{3.11}$$

where a,b are defined by (2.14) and α by (3.2).

Proof of (3.8). Let $\varphi \in C^0(\overline{\Omega})$. By the Cauchy–Schwarz inequality and estimate (3.6) we have

$$\left| \int_\Omega a_\varepsilon \, T_{0,1}(\hat{v}_{2,\varepsilon}) \, u_\varepsilon \, \varphi \right| \leq c \left(\int_\Omega a_\varepsilon \, \varphi^2 \right)^{\frac{1}{2}},$$

and since $T_{0,1}(\hat{v}_{2,\varepsilon}) \approx_\varepsilon \hat{v}_{2,\varepsilon}$,

$$\left| \int_\Omega a_\varepsilon \, \hat{v}_{2,\varepsilon} \, u_\varepsilon \, \varphi + o(1) \right| \leq c \left(\int_\Omega a_\varepsilon \, \varphi^2 \right)^{\frac{1}{2}}.$$

Moreover the periodicity of A_ε combined with the second limit of (2.3) implies that a_ε weakly $*$ converges to $1+\kappa$ in $\mathcal{M}(\overline{\Omega})$. Then passing to the limit in the previous inequality, thanks to (3.7), yields

$$\left| \int_\Omega v \, \varphi \right| \leq c \, \|\varphi\|_{L^2(\Omega)} \quad \text{for any } \varphi \in C^0(\overline{\Omega}),$$

which implies that $v \in L^2(\Omega)$. Similarly, since $a_\varepsilon \, \hat{v}_{2,\varepsilon}$ does not depend on x_3 and $u_\varepsilon = 0$ on $\partial\Omega$, we obtain, thanks to an integration by parts, for any $\varphi \in C^1(\overline{\Omega})$,

$$\left| \int_\Omega a_\varepsilon \, T_{0,1}(\hat{v}_{2,\varepsilon}) \, u_\varepsilon \, \frac{\partial \varphi}{\partial x_3} \right| = \left| \int_\Omega a_\varepsilon \, T_{0,1}(\hat{v}_{2,\varepsilon}) \, \frac{\partial u_\varepsilon}{\partial x_3} \, \varphi \right| \leq c \left(\int_\Omega a_\varepsilon \, \varphi^2 \right)^{\frac{1}{2}} \quad \text{by (3.6)},$$

whence by passing to the limit

$$\left| \int_\Omega v \, \frac{\partial \varphi}{\partial x_3} \right| \leq c \, \|\varphi\|_{L^2(\Omega)} \quad \text{for any } \varphi \in C^1(\overline{\Omega}),$$

which implies that $\frac{\partial v}{\partial x_3} \in L^2(\Omega)$ and $v = 0$ on $\Omega_2 \times \{0,1\}$. Therefore (3.8) holds.

Proof of (3.9). By estimate (3.4) we have for any $V \in H^1_\#(Y_2)$

$$\left\| V - \frac{\langle V, 1 \rangle_\varepsilon}{\langle 1, 1 \rangle_\varepsilon} \, 1 - \langle V, V_{1,\varepsilon} \rangle_\varepsilon \, V_{1,\varepsilon} \right\|_\varepsilon \leq c \left(\frac{\Lambda_1(\varepsilon)}{\Lambda_2(\varepsilon)} \right)^{\frac{1}{2}} \|V\|_\varepsilon + \Lambda_2(\varepsilon)^{-\frac{1}{2}} \, \|\nabla V\|_\varepsilon,$$

which implies the estimate

$$\left| \langle \hat{V}_{1,\varepsilon}, V \rangle_\varepsilon - \langle \hat{V}_{1,\varepsilon}, 1 \rangle_\varepsilon \int_{Y_2} V \right| \leq c \, \eta_\varepsilon \, \|V\|_\varepsilon + c \, \Lambda_2(\varepsilon)^{-\frac{1}{2}} \, \|\nabla V\|_\varepsilon, \tag{3.12}$$

where by (3.2) and (2.8)

$$\eta_\varepsilon := \left| \langle \hat{V}_{1,\varepsilon}, V_{1,\varepsilon} \rangle_\varepsilon - \langle \hat{V}_{1,\varepsilon}, 1 \rangle_\varepsilon \int_{Y_2} V_{1,\varepsilon} \right| + \left(\frac{\Lambda_1(\varepsilon)}{\Lambda_2(\varepsilon)} \right)^{\frac{1}{2}} \underset{\varepsilon \to 0}{\longrightarrow} |\beta - (1+\kappa) \, \alpha \, c_1| = 0. \tag{3.13}$$

On the other hand, since $A_\varepsilon \, \hat{V}_{1,\varepsilon} - \langle \hat{V}_{1,\varepsilon}, 1 \rangle_\varepsilon$ has a zero Y_2-averaged value, by the Lax–Milgram theorem there exists $W_\varepsilon \in H^1_\#(Y_2)$, defined up to an additive constant, solution of the problem

$$-\operatorname{div}(A_\varepsilon \nabla W_\varepsilon) = A_\varepsilon \, \hat{V}_{1,\varepsilon} - \langle \hat{V}_{1,\varepsilon}, 1 \rangle_\varepsilon \quad \text{in } Y_2, \tag{3.14}$$

whose variational formulation is

$$\forall\, V \in H^1_\#(Y_2), \quad \langle W_\varepsilon, V\rangle_\varepsilon = \langle \hat V_{1,\varepsilon}, V\rangle_\varepsilon - \langle \hat V_{1,\varepsilon}, 1\rangle_\varepsilon \int_{Y_2} V. \tag{3.15}$$

We choose W_ε such that $\langle W_\varepsilon, 1\rangle_\varepsilon = 0$, which implies $\|W_\varepsilon\|^2_\varepsilon \le \Lambda_1(\varepsilon)^{-1}\,\|\nabla W_\varepsilon\|^2_\varepsilon$ by the definition of the eigenvalue $\Lambda_1(\varepsilon)$. Then by putting W_ε in (3.15) we obtain, thanks to estimate (3.12),

$$\|\nabla W_\varepsilon\|_\varepsilon \le c\left(\eta_\varepsilon\,\Lambda_1(\varepsilon)^{-\frac12} + \Lambda_2(\varepsilon)^{-\frac12}\right). \tag{3.16}$$

Let $\varphi \in \mathcal{D}(\Omega)$. Putting φu_ε in the rescaled equation (3.14) yields

$$\varepsilon \int_\Omega a_\varepsilon \nabla_y W_\varepsilon(\tfrac{x}{\varepsilon}) \cdot \nabla(\varphi u_\varepsilon) = \int_\Omega a_\varepsilon\,\hat v_{1,\varepsilon}\,\varphi\,u_\varepsilon - \langle \hat V_{1,\varepsilon}, 1\rangle_\varepsilon \int_\Omega \varphi\,u_\varepsilon. \tag{3.17}$$

Moreover, by estimates (3.6) and (3.16) the left-hand side of (3.17) is bounded by the sequence $\eta_\varepsilon\,\varepsilon\,\Lambda_1(\varepsilon)^{-\frac12} + \varepsilon\,\Lambda_2(\varepsilon)^{-\frac12}$, which tends to zero by (3.13) and (2.8). Therefore, passing to the limit in (3.17) yields

$$\int_\Omega a_\varepsilon\,\hat v_{1,\varepsilon}\,\varphi\,u_\varepsilon \;\underset{\varepsilon\to0}{\longrightarrow}\; (1+\kappa)\,\alpha \int_\Omega \varphi\,u \quad \text{for any } \varphi \in \mathcal{D}(\Omega),$$

whence (3.9).

Proof of (3.10). Let ω_ε be the set of high conductivity whose period is $\varepsilon\,Q_\varepsilon$. Since u_ε is bounded in $H^1(\Omega)$, $\hat v_{1,\varepsilon}$ strongly converges to 1 in $L^2(\Omega)$, and $|\omega_\varepsilon| \to 0$, we have

$$\mathbf{1}_{\Omega\setminus\omega_\varepsilon} a_\varepsilon\,T_{0,1}(\hat v_{1,\varepsilon})\,\nabla u_\varepsilon \rightharpoonup \nabla u \quad \text{weakly in } L^2(\Omega;\mathbb{R}^3),$$

and since $T_{0,1}(\hat v_{1,\varepsilon}) \approx_\varepsilon \hat v_{1,\varepsilon}$ in the sense of definition (2.26) we also have

$$\mathbf{1}_{\Omega\setminus\omega_\varepsilon} a_\varepsilon\,\hat v_{1,\varepsilon}\,\nabla u_\varepsilon - \mathbf{1}_{\Omega\setminus\omega_\varepsilon} a_\varepsilon\,T_{0,1}(\hat v_{1,\varepsilon})\,\nabla u_\varepsilon \longrightarrow \nabla u \quad \text{strongly in } L^2(\Omega;\mathbb{R}^3),$$

whence the weak convergence

$$\mathbf{1}_{\Omega\setminus\omega_\varepsilon} a_\varepsilon\,\hat v_{1,\varepsilon}\,\nabla u_\varepsilon \rightharpoonup \nabla u \quad \text{weakly in } \mathcal{D}'(\Omega;\mathbb{R}^3).$$

It thus remains to prove the following convergence:

$$\mathbf{1}_{\omega_\varepsilon} a_\varepsilon\,\hat v_{1,\varepsilon}\,\nabla u_\varepsilon \rightharpoonup a\,\frac{\partial u}{\partial x_3}\,e_3 \quad \text{weakly in } \mathcal{D}'(\Omega;\mathbb{R}^3). \tag{3.18}$$

Consider $\mu \in (e_3)^\perp$ to be a vector of $\mathbb{R}^2$. By combining assumptions (2.10) and (2.11) the function $\Psi_{\mu,\varepsilon} := (1-\Phi_\varepsilon)\,\Psi_\mu$, where $\mathrm{cap}\,(Q_\varepsilon) = \|\nabla\Phi_\varepsilon\|^2_{L^2(Y_2)}$, satisfies

$$\Psi_{\mu,\varepsilon} \longrightarrow 0 \ \text{ strongly in } H^1_\#(Y_2) \quad \text{and} \quad \nabla\Psi_{\mu,\varepsilon} = \mu \text{ in } Q_\varepsilon. \tag{3.19}$$

We can also assume that $\Psi_{\mu,\varepsilon}$ is uniformly bounded by using a truncature of Φ_ε. By the definition of problem (2.5) we have

$$\varepsilon\,a_\varepsilon \nabla u_\varepsilon \cdot \nabla\left(T_{0,1}(\hat v_{1,\varepsilon})\,\Psi_{\mu,\varepsilon}(\tfrac{x}{\varepsilon})\right) \rightharpoonup 0 \quad \text{weakly in } \mathcal{D}'(\Omega),$$

and since $a_\varepsilon\nabla u_\varepsilon \cdot \nabla T_{0,1}(\hat v_{1,\varepsilon})$ is bounded in $L^1(\Omega)$ by (3.6) and by (2.8) (which implies $\|\nabla \hat V_{1,\varepsilon}\|_\varepsilon = O(\varepsilon)$), we also have

$$\varepsilon\,a_\varepsilon \nabla u_\varepsilon \cdot \nabla T_{0,1}(\hat v_{1,\varepsilon})\,\Psi_{\mu,\varepsilon}(\tfrac{x}{\varepsilon}) \longrightarrow 0 \quad \text{strongly in } L^1(\Omega),$$

whence the convergence

$$a_\varepsilon \nabla u_\varepsilon \cdot \nabla_y \Psi_{\mu,\varepsilon}(\tfrac{x}{\varepsilon})\, T_{0,1}(\hat{v}_{1,\varepsilon}) \rightharpoonup 0 \quad \text{weakly in } \mathcal{D}'(\Omega). \tag{3.20}$$

However, by the strong convergence of (3.19) we have

$$\mathbf{1}_{\Omega\setminus\omega_\varepsilon} a_\varepsilon \nabla u_\varepsilon \cdot \nabla_y \Psi_{\mu,\varepsilon}(\tfrac{x}{\varepsilon})\, T_{0,1}(\hat{v}_{1,\varepsilon}) = \mathbf{1}_{\Omega\setminus\omega_\varepsilon} \nabla u_\varepsilon \cdot \nabla_y \Psi_{\mu,\varepsilon}(\tfrac{x}{\varepsilon})\, T_{0,1}(\hat{v}_{1,\varepsilon}) \longrightarrow 0 \quad \text{in } L^1(\Omega).$$

This, combined with convergence (3.20) and the equality of (3.19), yields

$$\mathbf{1}_{\omega_\varepsilon} a_\varepsilon \nabla u_\varepsilon \cdot \nabla_y \Psi_{\mu,\varepsilon}(\tfrac{x}{\varepsilon})\, T_{0,1}(\hat{v}_{1,\varepsilon}) = \mathbf{1}_{\omega_\varepsilon} a_\varepsilon\, T_{0,1}(\hat{v}_{1,\varepsilon})\, \nabla u_\varepsilon \cdot \mu \rightharpoonup 0 \quad \text{weakly in } \mathcal{D}'(\Omega),$$

and since $T_{0,1}(\hat{v}_{1,\varepsilon}) \approx_\varepsilon \hat{v}_{1,\varepsilon}$ in the sense of (2.26), we obtain

$$\mathbf{1}_{\omega_\varepsilon} a_\varepsilon\, \hat{v}_{1,\varepsilon}\, \nabla u_\varepsilon \cdot \mu \rightharpoonup 0 \quad \text{weakly in } \mathcal{D}'(\Omega). \tag{3.21}$$

On the other hand, since a_ε and $\hat{v}_{1,\varepsilon}$ are independent of x_3 we have

$$\mathbf{1}_{\omega_\varepsilon} a_\varepsilon\, \hat{v}_{1,\varepsilon}\, \frac{\partial u_\varepsilon}{\partial x_3} = \frac{\partial}{\partial x_3}\left(\mathbf{1}_{\omega_\varepsilon} a_\varepsilon\, \hat{v}_{1,\varepsilon}\, u_\varepsilon\right),$$

and by (3.9), (3.3) we have

$$\mathbf{1}_{\omega_\varepsilon} a_\varepsilon\, \hat{v}_{1,\varepsilon}\, u_\varepsilon = a_\varepsilon\, \hat{v}_{1,\varepsilon}\, u_\varepsilon - \mathbf{1}_{\Omega\setminus\omega_\varepsilon} \hat{v}_{1,\varepsilon}\, u_\varepsilon \rightharpoonup (1+\kappa)\, \alpha\, u - u = a\, u \quad \text{weakly in } \mathcal{D}'(\Omega),$$

whence the convergence

$$\mathbf{1}_{\omega_\varepsilon} a_\varepsilon\, \hat{v}_{1,\varepsilon}\, \frac{\partial u_\varepsilon}{\partial x_3} \rightharpoonup a\, \frac{\partial u}{\partial x_3} \quad \text{weakly in } \mathcal{D}'(\Omega).$$

This, combined with (3.21), yields the desired convergence (3.18) and thus (3.10).

Proof of (3.11). It is quite similar to the proof of (3.10) by using the definition (3.7) of the function v.

Third step. Determination of the limit problem.

Let $\varphi \in \mathcal{D}(\Omega)$. We put the function $\varphi\, \hat{v}_{1,\varepsilon}$ defined by (3.1) as a test function in problem (2.5), whence

$$\int_\Omega a_\varepsilon \nabla u_\varepsilon \cdot \nabla \hat{v}_{1,\varepsilon}\, \varphi + \int_\Omega a_\varepsilon \nabla u_\varepsilon \cdot \nabla \varphi\, \hat{v}_{1,\varepsilon} = \int_\Omega f\, \varphi\, \hat{v}_{1,\varepsilon}.$$

Then by the convergence (3.10) of the second step we obtain

$$\int_\Omega a_\varepsilon \nabla u_\varepsilon \cdot \nabla \hat{v}_{1,\varepsilon}\, \varphi \;\underset{\varepsilon\to 0}{\longrightarrow}\; \int_\Omega f\, \varphi - \int_\Omega \nabla u \cdot \nabla \varphi - \int_\Omega a\, \frac{\partial u}{\partial x_3}\, \frac{\partial \varphi}{\partial x_3}. \tag{3.22}$$

For the limit of the left-hand side of (3.22) we need the following result.

LEMMA 3.2. *Let v_ε be a function in $H^1_0(\Omega)$ such that the sequence $\|v_\varepsilon\|_\varepsilon + \|\nabla v_\varepsilon\|_\varepsilon$ (see notation* (2.24)*) is bounded. Then the rescaled function $v_{1,\varepsilon}(x) := V_{1,\varepsilon}(\frac{x}{\varepsilon})$, where $V_{1,\varepsilon}$ is the eigenfunction of* (2.7)*, satisfies*

$$\int_\Omega a_\varepsilon \nabla v_{1,\varepsilon}\, v_\varepsilon \;\underset{\varepsilon\to 0}{\longrightarrow}\; 0. \tag{3.23}$$

By the definition (3.2) of $\hat{V}_{1,\varepsilon}$ and by rescaling the spectral problem (2.7) we have

$$\int_\Omega a_\varepsilon \nabla u_\varepsilon \cdot \nabla \hat{v}_{1,\varepsilon}\, \varphi = \beta \int_\Omega a_\varepsilon \nabla v_{1,\varepsilon} \cdot \nabla(\varphi\, u_\varepsilon) - \beta \int_\Omega a_\varepsilon \nabla v_{1,\varepsilon} \cdot \nabla \varphi\, u_\varepsilon$$
$$= \frac{\Lambda_1(\varepsilon)}{\varepsilon^2} \int_\Omega a_\varepsilon\, u_\varepsilon\, (\hat{v}_{1,\varepsilon} - \alpha)\, \varphi - \beta \int_\Omega a_\varepsilon \nabla v_{1,\varepsilon} \cdot \nabla \varphi\, u_\varepsilon,$$

which, combined with the limit (3.23) of Lemma 3.2, yields

$$\int_\Omega a_\varepsilon \nabla u_\varepsilon \cdot \nabla \hat{v}_{1,\varepsilon}\, \varphi - \frac{\Lambda_1(\varepsilon)}{\varepsilon^2} \int_\Omega a_\varepsilon\, u_\varepsilon\, (\hat{v}_{1,\varepsilon} - \alpha)\, \varphi \ \underset{\varepsilon\to 0}{\longrightarrow}\ 0.$$

Moreover by assumption (2.8), by the convergences (3.9), (3.7) combined with the estimate $\hat{v}_{1,\varepsilon} + \hat{v}_{2,\varepsilon} \approx_\varepsilon 1$, and by the definitions (2.14) we have

$$\frac{\Lambda_1(\varepsilon)}{\varepsilon^2}\, a_\varepsilon\, u_\varepsilon\, (\hat{v}_{1,\varepsilon} - \alpha) \ \underset{\varepsilon\to 0}{\longrightarrow}\ \lambda_1\, (1+\kappa)\, (\alpha - \alpha^2)\, u - \lambda_1\, \alpha\, b\, v = \gamma\, (u - v) \quad \text{weakly in } \mathcal{D}'(\Omega).$$

From both previous limits we deduce that

$$\int_\Omega a_\varepsilon \nabla u_\varepsilon \cdot \nabla \hat{v}_{1,\varepsilon}\, \varphi \ \underset{\varepsilon\to 0}{\longrightarrow}\ \int_\Omega \gamma\, (u - v), \tag{3.24}$$

which, combined with (3.22), yields

$$\int_\Omega \nabla u \cdot \nabla \varphi + \int_\Omega a\, \frac{\partial u}{\partial x_3}\, \frac{\partial \varphi}{\partial x_3} + \int_\Omega \gamma\, (u - v)\, \varphi = \int_\Omega f\, \varphi.$$

This equality is the variational formulation of the first equation of (2.12).

Similarly by using the test function $(1 - \hat{v}_{1,\varepsilon})\, \varphi$ and convergences (3.11), (3.24) we obtain

$$\int_\Omega a\, \frac{\partial v}{\partial x_3}\, \frac{\partial \varphi}{\partial x_3} + \int_\Omega \gamma\, (v - u) = 0,$$

which is the variational formulation of the second equation of problem (2.12). Theorem 2.1 is thus proved.

Proof of Lemma 3.2. Let $(1, V_{1,\varepsilon})^{\perp_\varepsilon}$ be the space of the functions in $L^2_\#(Y_2)$ which are orthogonal to the functions $1, V_{1,\varepsilon}$ in the sense of the scalar product $\langle \cdot, \cdot \rangle_\varepsilon$ of (2.23). The idea is to consider, for $\mu \in \mathbb{R}^2$, the orthogonal projection of $\nabla V_{1,\varepsilon} \cdot \mu$ on the space $(1, V_{1,\varepsilon})^{\perp_\varepsilon}$. So for $\mu \in \mathbb{R}^2$, let us define the function Σ_ε by

$$\varepsilon\, \Sigma_\varepsilon = \nabla V_{1,\varepsilon} \cdot \mu - \frac{\langle \nabla V_{1,\varepsilon} \cdot \mu, 1 \rangle_\varepsilon}{\langle 1, 1 \rangle_\varepsilon} - \langle \nabla V_{1,\varepsilon} \cdot \mu, V_{1,\varepsilon} \rangle_\varepsilon\, V_{1,\varepsilon}. \tag{3.25}$$

By the Lax–Milgram theorem there exists a unique solution X_ε in $H^1_\#(Y_2) \cap (1, V_{1,\varepsilon})^{\perp_\varepsilon}$ of the problem

$$\forall\, V \in H^1_\#(Y_2) \cap (1, V_{1,\varepsilon})^{\perp_\varepsilon}, \quad \langle \nabla X_\varepsilon, \nabla V \rangle_\varepsilon = \langle \Sigma_\varepsilon, V \rangle_\varepsilon. \tag{3.26}$$

Since $X_\varepsilon, \Sigma_\varepsilon \in (1, V_{1,\varepsilon})^{\perp_\varepsilon}$, we have $\langle \nabla X_\varepsilon, \nabla V_{1,\varepsilon} \rangle_\varepsilon = \Lambda_1(\varepsilon)\, \langle X_\varepsilon, V_{1,\varepsilon} \rangle_\varepsilon = 0$ and

$$\forall\, V \in H^1_\#(Y_2), \quad \langle \Sigma_\varepsilon, V \rangle_\varepsilon = \langle \nabla X_\varepsilon, \nabla V \rangle_\varepsilon - \langle \nabla X_\varepsilon, \nabla V_{1,\varepsilon} \rangle_\varepsilon\, \langle V_{1,\varepsilon}, V \rangle_\varepsilon = \langle \nabla X_\varepsilon, \nabla V \rangle_\varepsilon,$$
$$\text{whence} \quad A_\varepsilon\, \Sigma_\varepsilon = -\operatorname{div}(A_\varepsilon \nabla X_\varepsilon) \quad \text{in } Y_2.$$

Set $\sigma_\varepsilon := \Sigma_\varepsilon(\frac{x}{\varepsilon})$ and $\chi_\varepsilon(x) := X_\varepsilon(\frac{x}{\varepsilon})$. Then by rescaling the previous equation we obtain

$$a_\varepsilon\,\sigma_\varepsilon = -\,\varepsilon^2\,\mathrm{div}\,(a_\varepsilon \nabla\chi_\varepsilon) \quad \text{in } \mathcal{D}'(\Omega). \tag{3.27}$$

Moreover, by putting the function X_ε in problem (3.26) and applying estimate (3.4) with $V := X_\varepsilon \in (1, V_{1,\varepsilon})^{\perp_\varepsilon}$ we have

$$\|\nabla X_\varepsilon\|_\varepsilon^2 \le \|\Sigma_\varepsilon\|_\varepsilon\,\|X_\varepsilon\|_\varepsilon \le \frac{1}{\Lambda_2(\varepsilon)}\,\|\Sigma_\varepsilon\|_\varepsilon\,\|\nabla X_\varepsilon\|_\varepsilon,$$

whence by the definitions (3.25) of Σ_ε and (2.7) of $V_{1,\varepsilon}$ the following estimate:

$$\|\nabla X_\varepsilon\|_\varepsilon \le \frac{c}{\varepsilon\,\Lambda_2(\varepsilon)}\,\|\nabla V_{1,\varepsilon}\|_\varepsilon \le \frac{c}{\Lambda_2(\varepsilon)^{\frac{1}{2}}}. \tag{3.28}$$

Now let v_ε be a function such that the sequence $\|v_\varepsilon\|_\varepsilon + \|\nabla v_\varepsilon\|_\varepsilon$ is bounded. We have, by using successively the Cauchy–Schwarz inequality, estimate (3.28), and assumption (2.8),

$$\left|\,\varepsilon^2 \int_\Omega a_\varepsilon \nabla\chi_\varepsilon \cdot \nabla v_\varepsilon\,\right| \le \varepsilon^2\,\|\nabla\chi_\varepsilon\|_\varepsilon\,\|\nabla v_\varepsilon\|_\varepsilon \le c\,\varepsilon\,\|\nabla X_\varepsilon\|_\varepsilon \le \frac{c\,\varepsilon}{\Lambda_2(\varepsilon)^{\frac{1}{2}}} \underset{\varepsilon\to 0}{\longrightarrow} 0.$$

By putting v_ε as a test function in (3.27) we deduce from both previous limits that

$$\int_\Omega a_\varepsilon\,\sigma_\varepsilon\,v_\varepsilon \underset{\varepsilon\to 0}{\longrightarrow} 0. \tag{3.29}$$

Finally by the definition (3.25) of Σ_ε we have

$$a_\varepsilon\,\nabla v_{1,\varepsilon}\cdot\mu\,v_\varepsilon = a_\varepsilon\,\sigma_\varepsilon\,v_\varepsilon + \frac{1}{\varepsilon}\,\frac{\langle \nabla V_{1,\varepsilon}\cdot\mu, 1\rangle_\varepsilon}{\langle 1,1\rangle_\varepsilon}\,a_\varepsilon\,v_\varepsilon + \frac{1}{\varepsilon}\,\langle \nabla V_{1,\varepsilon}\cdot\mu, V_{1,\varepsilon}\rangle_\varepsilon\,a_\varepsilon\,v_{1,\varepsilon}\,v_\varepsilon.$$

Then by taking into account limit (3.29) it remains to prove that

$$\lim_{\varepsilon\to 0}\,\frac{1}{\varepsilon}\,\langle \nabla V_{1,\varepsilon}\cdot\mu, 1\rangle_\varepsilon = \lim_{\varepsilon\to 0}\,\frac{1}{\varepsilon}\,\langle \nabla V_{1,\varepsilon}\cdot\mu, V_{1,\varepsilon}\rangle_\varepsilon = 0 \tag{3.30}$$

in order to establish Lemma 3.2.

Set $W_{1,\varepsilon} := \frac{1}{\varepsilon}(V_{1,\varepsilon} - \int_{Y_2} V_{1,\varepsilon})$. The sequence $W_{1,\varepsilon}$ is bounded in $H^1_\#(Y_2)$ thanks to the Poincaré–Wirtinger inequality combined with assumption (2.8). Therefore $W_{1,\varepsilon}$ weakly converges to some W_1 in $H^1_\#(Y_2)$ up to a subsequence. By assumption (2.10) there exists a function Φ_ε in $H^1_\#(Y_2)$ which strongly converges to 1 in $H^1_\#(Y_2)$ and which is equal to zero in the set Q_ε. Let $V \in C^1_\#(Y_2)$. We have, by the definition (2.7) of the eigenfunction $V_{1,\varepsilon}$ and by assumption (2.8),

$$\frac{1}{\varepsilon}\,\langle \nabla V_{1,\varepsilon}, \nabla(\Phi_\varepsilon\,V)\rangle_\varepsilon = \frac{\Lambda_1(\varepsilon)}{\varepsilon}\,\langle V_{1,\varepsilon}, \Phi_\varepsilon\,V\rangle_\varepsilon = O(\varepsilon) \underset{\varepsilon\to 0}{\longrightarrow} 0,$$

and we also have

$$\begin{aligned}\frac{1}{\varepsilon}\,\langle \nabla V_{1,\varepsilon}, \nabla(\Phi_\varepsilon\,V)\rangle_\varepsilon &= \int_{Y_2} \nabla W_{1,\varepsilon}\cdot\nabla V\,\Phi_\varepsilon + \int_{Y_2} \nabla W_{1,\varepsilon}\cdot\nabla\Phi_\varepsilon\,V \\ &= \int_{Y_2} \nabla W_{1,\varepsilon}\cdot\nabla V + o(1),\end{aligned}$$

$$\text{whence} \quad \int_{Y_2} \nabla W_1\cdot\nabla V = 0 \quad \text{for any } V \in H^1_\#(Y_2).$$

Therefore $W_1 = 0$ and $W_{1,\varepsilon}$ weakly converges to 0 in $H^1_\#(Y_2)$. On the other hand we have by assumption (2.11) and for any $\mu \in \mathbb{R}^2$,

$$\begin{aligned}\frac{1}{\varepsilon}\,\langle \nabla V_{1,\varepsilon}, \nabla \Psi_\mu \rangle_\varepsilon &= \int_{Y\setminus Q_\varepsilon} \nabla W_{1,\varepsilon}\cdot\nabla\Psi_\mu + \frac{1}{\varepsilon}\int_{Q_\varepsilon} A_\varepsilon \nabla V_{1,\varepsilon}\cdot\mu \\ &= \frac{\Lambda_1(\varepsilon)}{\varepsilon}\,\langle V_{1,\varepsilon}, \Psi_\mu\, V\rangle_\varepsilon = O(\varepsilon) \underset{\varepsilon\to 0}{\longrightarrow} 0,\end{aligned}$$

whence $\displaystyle\frac{1}{\varepsilon}\int_{Q_\varepsilon} A_\varepsilon\nabla V_{1,\varepsilon}\cdot\mu = -\int_{Y\setminus Q_\varepsilon}\nabla W_{1,\varepsilon}\cdot\nabla\Psi_\mu + o(1) \underset{\varepsilon\to 0}{\longrightarrow} -\int_{Y_2}\nabla W_1\cdot\nabla\Psi_\mu = 0.$

Finally we obtain

$$\frac{1}{\varepsilon}\,\langle \nabla V_{1,\varepsilon}\cdot\mu, 1\rangle_\varepsilon = \int_{Y\setminus Q_\varepsilon}\nabla W_{1,\varepsilon}\cdot\mu + \frac{1}{\varepsilon}\int_{Q_\varepsilon} A_\varepsilon\nabla V_{1,\varepsilon}\cdot\mu \underset{\varepsilon\to 0}{\longrightarrow} \int_{Y_2}\nabla W_1\cdot\mu = 0,$$

which yields the first limit of (3.30).

Similarly we obtain the second limit of (3.30) by starting from the equality

$$\frac{1}{\varepsilon}\,\langle \nabla V_{1,\varepsilon}, \nabla(V_{1,\varepsilon}\,\Psi_\mu)\rangle_\varepsilon = \frac{\Lambda_1(\varepsilon)}{\varepsilon}\,\langle V_{1,\varepsilon}, V_{1,\varepsilon}\,\Psi_\mu\rangle_\varepsilon = O(\varepsilon)$$

and by using the strong convergence of $V_{1,\varepsilon}$ to c_1 in $L^2_\#(Y_2)$. Lemma 3.2 is proved, which also concludes the proof of Theorem 2.1.

3.2. Proof of Theorem 2.4.

Proof of part (i) *of Theorem* 2.4. Let $n \geq 1$ be an integer such that the eigenvalues $\Lambda_1(\varepsilon), \ldots, \Lambda_{n-1}(\varepsilon)$ of problem (2.7) are $O(\varepsilon^2)$. Let $i = 0, \ldots, n-1$. Since the eigenfunction $V_{i,\varepsilon}$ of (2.7) satisfies $\|V_{i,\varepsilon}\|_\varepsilon = 1$ and $\|\nabla V_{i,\varepsilon}\|^2_\varepsilon = \Lambda_i(\varepsilon) = O(\varepsilon^2)$ by assumption, there exists a constant c_i such that (up to a subsequence)

$$V_{i,\varepsilon} \rightharpoonup c_i \quad \text{weakly in } H^1_\#(Y_2). \tag{3.31}$$

Moreover, for any $j = 1, \ldots, N$, since the conductivity $\alpha_{j,\varepsilon}$ in $Q_{j,\varepsilon}$ satisfies the limit $\alpha_{j,\varepsilon}\,|Q_{j,\varepsilon}| \to \kappa_j > 0$, we have by the Cauchy–Schwarz inequality

$$\left| \fint_{Q_{j,\varepsilon}} V_{i,\varepsilon} \right| \leq \frac{1}{|Q_{j,\varepsilon}|^{\frac{1}{2}}}\,\|V_{i,\varepsilon}\|_{L^2(Q_{j,\varepsilon})} \leq c\,\|V_{i,\varepsilon}\|_\varepsilon = c.$$

Then we have (up to a subsequence)

$$\fint_{Q_{j,\varepsilon}} V_{i,\varepsilon} \underset{\varepsilon\to 0}{\longrightarrow} c_{i,j} \quad \text{for } j = 1, \ldots, N. \tag{3.32}$$

The limit of $V_{i,\varepsilon}$ is thus characterized by the sequence of $\mathbb{R}^{N+1}$:

$$\dot{c}_i := (c_i, c_{i,1}, \ldots, c_{i,N}) \quad \text{for } i = 0, \ldots, n-1. \tag{3.33}$$

Moreover by the orthonormality of the family $(V_{0,\varepsilon}, \ldots, V_{n-1,\varepsilon})$ with respect to the scalar product $\langle\cdot,\cdot\rangle_\varepsilon$ of (2.23), the limit family $(\dot{c}_0, \ldots, \dot{c}_{n-1})$ is also orthonormal with respect to the following scalar product of $\mathbb{R}^{N+1}$:

$$\langle \dot{a}, \dot{b}\rangle := a_0\,b_0 + \sum_{j=1}^N \kappa_j\,a_j\,b_j, \quad \dot{a}, \dot{b} \in \mathbb{R}^{N+1}, \tag{3.34}$$

which implies $n \le N+1$. Hence there exists a largest integer n satisfying (2.28) such that all the eigenvalues $\Lambda_1(\varepsilon), \dots, \Lambda_{n-1}(\varepsilon)$ are $O(\varepsilon^2)$; n is thus the smallest integer such that the sequence $\frac{\Lambda_n(\varepsilon)}{\varepsilon^2}$ is not bounded.

Now we can assume that condition (2.29) is satisfied with this integer n.

We will use as in [3] an argument based on truncatures of the orthonormal family of eigenfunctions $(V_{0,\varepsilon}, \dots, V_{n-1,\varepsilon})$ of (2.7), where $V_{0,\varepsilon} := \|1\|_\varepsilon^{-1}$, in order to build the desired asymptotic partition.

Let $i = 0, \dots, n-1$ and let T be a Lipschitz function in $C^1(\mathbb{R};\mathbb{R})$. By the Courant–Fisher formula applied to the eigenvalue $\Lambda_n(\varepsilon)$ we have by assumption (2.29)

$$\begin{aligned}
&\left\| T(V_{i,\varepsilon}) - \sum_{h=0}^{n-1} \langle T(V_{i,\varepsilon}), V_{h,\varepsilon}\rangle_\varepsilon \, V_{h,\varepsilon} \right\|_\varepsilon \\
&\le \Lambda_n(\varepsilon)^{-\frac12} \left\| \nabla T(V_{i,\varepsilon}) - \sum_{h=1}^{n-1} \langle T(V_{i,\varepsilon}), V_{h,\varepsilon}\rangle_\varepsilon \, \nabla V_{h,\varepsilon} \right\|_\varepsilon \le c\,\varepsilon\,\Lambda_n(\varepsilon)^{-\frac12} \underset{\varepsilon\to 0}{\longrightarrow} 0,
\end{aligned}$$

whence the existence of constants $a_{i,h}$ such that we have according to notation (2.25)

$$T(V_{i,\varepsilon}) \approx_\varepsilon \sum_{h=0}^{n-1} a_{i,h}\, V_{h,\varepsilon} \quad \text{for } i = 0, \dots, n-1. \tag{3.35}$$

On the other hand by using successively the Jensen inequality, the Cauchy–Schwarz inequality, the Poincaré–Wirtinger inequality (2.20), the assumption $\alpha_{j,\varepsilon}\,|Q_{j,\varepsilon}| \to \kappa_j > 0$, and estimate (2.29) we have for any $j = 1, \dots, N$

$$\begin{aligned}
\left| T\left(\fint_{Q_{j,\varepsilon}} V_{i,\varepsilon} \right) - \fint_{Q_{j,\varepsilon}} T(V_{i,\varepsilon}) \right| &\le \fint_{Q_{j,\varepsilon}} \left| T\left(\fint_{Q_{j,\varepsilon}} V_{i,\varepsilon} \right) - T(V_{i,\varepsilon}) \right| \\
&\le |Q_{j,\varepsilon}|^{-\frac12} \left\| T\left(\fint_{Q_{j,\varepsilon}} V_{i,\varepsilon} \right) - T(V_{i,\varepsilon}) \right\|_{L^2(Q_{j,\varepsilon})} \\
&\le \|T'\|_\infty\, |Q_{j,\varepsilon}|^{-\frac12} \left\| V_{i,\varepsilon} - \fint_{Q_{j,\varepsilon}} V_{i,\varepsilon} \right\|_{L^2(Q_{j,\varepsilon})} \\
&\le c\,\|\nabla V_{i,\varepsilon}\|_\varepsilon \le c\,\Lambda_i(\varepsilon)^{\frac12} \underset{\varepsilon\to 0}{\longrightarrow} 0,
\end{aligned}$$

whence by convergence (3.32)

$$\fint_{Q_{j,\varepsilon}} T(V_{i,\varepsilon}) \underset{\varepsilon\to 0}{\longrightarrow} T(c_{i,j}) \quad \text{for } j = 1, \dots, N.$$

Moreover by convergence (3.31) $T(V_{i,\varepsilon})$ weakly converges to $T(c_i)$ in $H^1_\#(Y_2)$. These convergences combined with estimate (3.35) and definition (3.33) yield

$$T(\dot c_i) := (T(c_i), \dots, T(c_{i,N})) = \sum_{h=0}^{n-1} a_{i,h}\, \dot c_h \quad \text{for } i = 0, \dots, n-1.$$

Therefore the subspace $F := \operatorname{Span}(\dot c_0, \dots, \dot c_{n-1})$ of $\mathbb{R}^{N+1}$ satisfies the following truncature property:

$$\text{for any Lipschitz function } T \in C^1(\mathbb{R},\mathbb{R}), \quad T(F) \subset F. \tag{3.36}$$

Moreover the family $(\dot c_0,\dots,\dot c_{n-1})$ is an orthonormal basis of the space F provided with the scalar product (3.34) and the vector $(1,\dots,1)=\frac{1}{\sqrt{1+\kappa}}\,\dot c_0$ belongs to F.

Then, by virtue of Lemma 3.1 of [4], the properties of the space F imply the existence of a partition $(\hat I_1,\dots,\hat I_n)$ of the set $\{0,\dots,N\}$ composed of n nonempty sets such that $0\in\hat I_1$ and for any $i=0,\dots,n-1$,

$$\begin{cases}\forall\, j\in\hat I_1\setminus\{0\}, & \hat c_{i,1}:=c_i=c_{i,j},\\ \forall\, j_1,j_2\in\hat I_k, & \hat c_{i,k}:=c_{i,j_1}=c_{i,j_2}\quad\text{for } k=2,\dots,n.\end{cases}\tag{3.37}$$

This means that for any $i=0,\dots,n-1$, the limits $c_{i,j}$ of (3.32) are equal in a given set $\hat I_k$. We also set from definition (2.22)

$$\begin{cases}\hat\kappa_1:=1+\displaystyle\sum_{h\in\hat I_1\setminus\{0\}}\kappa_h,\\ \hat\kappa_k:=\displaystyle\sum_{h\in\hat I_k}\kappa_h & \text{for } k=2,\dots,n.\end{cases}\tag{3.38}$$

On the first side the existence of such a partition implies inequality (2.28): $n\le N+1$. On the other side we will prove that the functions

$$\hat V_{k,\varepsilon}:=\sum_{i=0}^{n-1}\hat\kappa_k\,\hat c_{i,k}\,V_{i,\varepsilon}\quad\text{for } k=1,\dots,n\tag{3.39}$$

define the desired asymptotic partition of the unity.

Since the family $(V_{0,\varepsilon},\dots,V_{n-1,\varepsilon})$ is orthonormal with respect to the scalar product $\langle\cdot,\cdot\rangle_\varepsilon$ of (2.23) we have for any $i,j=0,\dots,n-1$

$$\delta_{i,j}=\langle V_{i,\varepsilon},V_{j,\varepsilon}\rangle_\varepsilon\ \underset{\varepsilon\to0}{\longrightarrow}\ c_i\,c_j+\sum_{h=1}^{N}\kappa_h\,c_{i,h}\,c_{j,h}=\delta_{i,j},\tag{3.40}$$

where $\delta_{i,j}$ denotes the Kronecker symbol, whence by definition (3.37)

$$\sum_{k=1}^{n}\hat\kappa_k\,\hat c_{i,k}\,\hat c_{j,k}=\delta_{i,j}\quad\text{for any } i,j=0,\dots,n-1;\tag{3.41}$$

i.e., the $(n\times n)$ matrix $\hat C:=\big[\sqrt{\hat\kappa_k}\,c_{i,k}\big]_{0\le i\le n-1,1\le k\le n}$ is orthogonal. From the orthonormality of the columns of $\hat C$ we deduce that

$$\sum_{i=0}^{n-1}\sqrt{\hat\kappa_h\,\hat\kappa_k}\,\hat c_{i,h}\,\hat c_{i,k}=\delta_{h,k}=\sum_{i=0}^{n-1}\hat\kappa_k\,\hat c_{i,h}\,\hat c_{i,k}\quad\text{for any } h,k=1,\dots,n.\tag{3.42}$$

On the other hand, thanks to the Poincaré–Wirtinger inequalities in Y_2 and (2.20) in $Q_{j,\varepsilon}$, any sequence V_ε in $H^1_\#(Y_2)$ such that the sequence $\|V\|_\varepsilon$ is bounded and $\|\nabla V_\varepsilon\|_\varepsilon\to0$ (according to notation (2.23)) can be replaced by its averaged values in Y_2 and in any set $Q_{j,\varepsilon}$ in such a way that (according to notation (2.25))

$$V_\varepsilon\approx_\varepsilon\left(\int_{Y_2}V_\varepsilon\right)\mathbf{1}_{Y_2\setminus Q_\varepsilon}+\sum_{j=1}^{N}\left(\fint_{Q_{j,\varepsilon}}V_\varepsilon\right)\mathbf{1}_{Q_{j,\varepsilon}}.\tag{3.43}$$

Let us now prove the properties (2.30) to (2.33) which define the asymptotic partition of the unity.

Proof of (2.30). By definitions (3.37), (3.39) and by equalities (3.42) we have for any $h,k=1,\dots,n$

$$\text{(3.44)}\qquad \begin{cases} \displaystyle\int_{Y_2} \hat{V}_{k,\varepsilon} \;\underset{\varepsilon\to 0}{\longrightarrow}\; \sum_{i=0}^{n-1} \hat{\kappa}_k\,\hat{c}_{i,k}\,\hat{c}_{i,1} = \delta_{k,1}, \\ \displaystyle\fint_{Q_{j,\varepsilon}} \hat{V}_{k,\varepsilon} \;\underset{\varepsilon\to 0}{\longrightarrow}\; \sum_{i=0}^{n-1} \hat{\kappa}_k\,\hat{c}_{i,k}\,\hat{c}_{i,h} = \delta_{k,h} \quad \text{for any } j\in\hat{I}_h. \end{cases}$$

Then by applying estimate (3.43) with $V_\varepsilon := V_{k,\varepsilon}$ and by taking into account limits (3.44) we obtain estimate (2.30).

Proof of (2.31). By applying again estimate (3.43) with the functions $\hat{V}_{k,\varepsilon}, \hat{V}_{h,\varepsilon}$ we obtain for any $h,k\in\{1,\dots,n\}$

$$\langle \hat{V}_{h,\varepsilon}, \hat{V}_{k,\varepsilon}\rangle_\varepsilon = \left(\int_{Y_2} \hat{V}_{h,\varepsilon}\right)\left(\int_{Y_2} \hat{V}_{k,\varepsilon}\right) + \sum_{j=1}^{N} \alpha_{j,\varepsilon}\,|Q_{j,\varepsilon}| \left(\fint_{Q_{j,\varepsilon}} \hat{V}_{h,\varepsilon}\right)\left(\fint_{Q_{j,\varepsilon}} \hat{V}_{k,\varepsilon}\right) + o(1),$$

which, combined with limits (3.44), (2.21) and definitions (3.38), yields

$$\langle \hat{V}_{h,\varepsilon}, \hat{V}_{k,\varepsilon}\rangle_\varepsilon \;\underset{\varepsilon\to 0}{\longrightarrow}\; \delta_{1,k}\,\delta_{1,h} + \sum_{j\in\hat{I}_k\cap\hat{I}_h\setminus\{0\}} \kappa_j = \delta_{k,h}\hat{\kappa}_k,$$

whence (2.31) and (2.35).

Proof of (2.32). Let $k=1,\dots,n$. Since $\hat{V}_{k,\varepsilon}$ weakly converges to 0 or 1 in $H^1_\#(Y_2)$ we have by the definition of the truncature (2.27)

$$\lim_{\varepsilon\to 0}\left(\int_{Y_2} T_{0,1}(\hat{V}_{k,\varepsilon})\right) = \lim_{\varepsilon\to 0}\left(\int_{Y_2} \hat{V}_{k,\varepsilon}\right),$$

and by the Poincaré–Wirtinger inequality (2.20) combined with $\|\nabla\hat{V}_{k,\varepsilon}\|_\varepsilon\to 0$ we also have for any $j=1,\dots,N$

$$\lim_{\varepsilon\to 0}\left(\fint_{Q_{j,\varepsilon}} T_{0,1}(\hat{V}_{k,\varepsilon})\right) = \lim_{\varepsilon\to 0}\; T_{0,1}\left(\fint_{Q_{j,\varepsilon}} \hat{V}_{k,\varepsilon}\right) = \lim_{\varepsilon\to 0}\left(\fint_{Q_{j,\varepsilon}} \hat{V}_{k,\varepsilon}\right)$$

since the last limit is equal to 0 or 1 by (3.44). We then deduce (2.32) from the previous equalities and from the estimate (3.43) with $V_\varepsilon := T_{0,1}(\hat{V}_{k,\varepsilon}) - \hat{V}_{k,\varepsilon}$.

Proof of (2.33). Let $i=0,\dots,n-1$. We have by the definition (3.39) of $\hat{V}_{k,\varepsilon}$ and by equalities (3.41)

$$\sum_{k=1}^{n} \hat{c}_{i,k}\,\hat{V}_{k,\varepsilon} = \sum_{k=1}^{n} \hat{c}_{i,k}\left(\sum_{j=0}^{n-1} \hat{\kappa}_k\,\hat{c}_{j,k}\,V_{j,\varepsilon}\right) = \sum_{j=0}^{n-1}\left(\sum_{k=1}^{n} \hat{\kappa}_k\,\hat{c}_{i,k}\,\hat{c}_{j,k}\right) V_{j,\varepsilon} = V_{i,\varepsilon},$$

which implies (2.33) and concludes the proof of part (i) of Theorem 2.4.

Proof of part (ii) *of Theorem* 2.4. The proof is quite similar to the second and third steps of the proof of Theorem 2.1 in that it uses the asymptotic partition $(\hat{V}_{1,\varepsilon},\dots,\hat{V}_{n,\varepsilon})$ defined in part (i). Therefore we recall only the main steps of the proof without details.

First we define the rescaled test functions

$$\hat{v}_{k,\varepsilon}(x) := \hat{V}_{k,\varepsilon}\left(\frac{x}{\varepsilon}\right) \quad \text{for almost every } x \in \Omega, \quad k = 1, \dots, n. \tag{3.45}$$

From the conditions (2.30) and (2.32) satisfied by the partition $(\hat{V}_{1,\varepsilon}, \dots, \hat{V}_{n,\varepsilon})$ we easily deduce the following strong convergences:

$$\hat{v}_{1,\varepsilon} \to 1 \text{ strongly in } L^2(\Omega) \quad \text{and} \quad \hat{v}_{k,\varepsilon} \to 0 \text{ strongly in } L^2(\Omega) \quad \text{for } k = 2, \dots, n. \tag{3.46}$$

Moreover by the truncature property (2.32) and the positivity of $\hat{\kappa}_k$ in (3.38) there exists, for any $k = 2, \dots, n$, a Radon measure v_k such that

$$a_\varepsilon \, \hat{v}_{k,\varepsilon} \, u_\varepsilon \rightharpoonup \hat{\kappa}_k \, v_k \quad \text{weakly} * \text{ in } \mathcal{M}(\overline{\Omega}). \tag{3.47}$$

Proceeding as in the second step of the proof of (2.1) we obtain that $v_k \in H^1_0(0,1; L^2(\Omega_2))$.

Then the following convergences hold true:

$$\begin{cases} a_\varepsilon \, \hat{v}_{1,\varepsilon} \, u_\varepsilon \rightharpoonup \hat{\kappa}_1 \, u & \text{weakly in } \mathcal{D}'(\Omega), \\ a_\varepsilon \, \hat{v}_{1,\varepsilon} \, \nabla u_\varepsilon \rightharpoonup \nabla u + (\hat{\kappa}_1 - 1) \, \dfrac{\partial u}{\partial x_3} & \text{weakly in } \mathcal{D}'(\Omega; \mathbb{R}^3), \\ a_\varepsilon \, \hat{v}_{k,\varepsilon} \, \nabla u_\varepsilon \rightharpoonup \hat{\kappa}_k \, \dfrac{\partial v_k}{\partial x_3} & \text{weakly in } \mathcal{D}'(\Omega; \mathbb{R}^3) \quad \text{for } k = 2, \dots, n, \end{cases} \tag{3.48}$$

where u is the weak limit of u_ε in $H^1_0(\Omega)$. The proof of (3.48) is quite similar to the second step of Theorem 2.1 in that it takes into account the following result, the proof of which is given at the end of the section.

LEMMA 3.3. *Under the geometrical assumptions* (2.18) *and* (2.19) *of the distribution of the fibers, the conditions* (2.10) *and* (2.11) *of* Theorem 2.1 *are satisfied.*

Finally for any $k = 1, \dots, n$ and $\varphi \in \mathcal{D}(\Omega)$, we put $\varphi \, \hat{v}_{k,\varepsilon}$ as test function in the conduction problem (2.5) in order to obtain the desired limit system (2.34) of size n. The proof is quite similar to the third step of Theorem 2.1 in that it uses the strong convergences (3.46), the expression (3.39) of the function $\hat{v}_{k,\varepsilon}$ (3.45) in terms of the rescaled eigenfunctions $v_{i,\varepsilon}(x) := V_{i,\varepsilon}(\frac{x}{\varepsilon})$, the assumption (2.29) on the eigenvalues, as well as the weak convergences

$$a_\varepsilon \, \nabla v_{i,\varepsilon} \, u_\varepsilon \rightharpoonup 0 \quad \text{weakly in } \mathcal{D}'(\Omega; \mathbb{R}^3) \quad \text{for } i = 0, \dots, n-1. \tag{3.49}$$

The proof of convergences (3.49) is quite similar to the proof of Lemma 3.2. Indeed by considering the orthogonal projection of $\frac{1}{\varepsilon} \, A_\varepsilon \nabla V_{i,\varepsilon}$ in the space $(V_{0,\varepsilon}, \dots, V_{n-1,\varepsilon})^{\perp_\varepsilon}$ (where $\perp_\varepsilon$ is the orthogonality according to (2.23)) we obtain for any $i = 0, \dots, n-1$ the following convergence:

$$a_\varepsilon \, \nabla v_{i,\varepsilon} \, u_\varepsilon - \sum_{j=0}^{n-1} \frac{1}{\varepsilon} \, \langle \nabla V_{i,\varepsilon}, V_{j,\varepsilon} \rangle_\varepsilon \, a_\varepsilon \, v_{j,\varepsilon} \, u_\varepsilon \rightharpoonup 0 \quad \text{weakly in } \mathcal{D}'(\Omega; \mathbb{R}^3). \tag{3.50}$$

We then deduce (3.49) from (3.50) combined with the limits

$$\frac{1}{\varepsilon} \, \langle \nabla V_{i,\varepsilon}, V_{j,\varepsilon} \rangle_\varepsilon \underset{\varepsilon \to 0}{\longrightarrow} 0 \quad \text{for } i, j = 0, \dots, n-1. \tag{3.51}$$

Proceeding as in the proof of Lemma 3.2, we see that the limits (3.51) are a consequence of conditions (2.10) and (2.11), which are given by Lemma 3.3. It thus remains to prove this lemma in order to conclude the proof of Theorem 2.4.

Proof of Lemma 3.3.

Proof of (2.10). We can regroup the neighboring sets $Q_{i,\varepsilon}$ of (2.18), $i = 1, \dots, N$, in such a way that the distance between two sets $Q_{i,\varepsilon}, Q_{j,\varepsilon}$ of the same group tends to zero and the distance between two different groups is greater than a positive constant. Let us denote by $\widehat{Q}_{1,\varepsilon}, \dots, \widehat{Q}_{p,\varepsilon}$ the p groups defined in this way. By construction for any $k = 1, \dots, p$, $\widehat{Q}_{k,\varepsilon}$ is contained in an open ball $B(y_{k,\varepsilon}, r_{k,\varepsilon})$ in the torus Y_2, of center $y_{k,\varepsilon}$ and of radius $r_{k,\varepsilon} \to 0$. Let Φ_ε be the function of $H^1_\#(Y_2)$ defined by

(3.52)
$$\begin{cases} \Phi_\varepsilon(y) := 0 & \text{if } y \in \displaystyle\bigcup_{k=1}^{p} B(y_{k,\varepsilon}, r_{k,\varepsilon}), \\ \Phi_\varepsilon(y) := 1 & \text{if } y \notin \displaystyle\bigcup_{k=1}^{p} B(y_{k,\varepsilon}, \sqrt{r_{k,\varepsilon}}), \\ \Phi_\varepsilon(y) := \dfrac{\ln|y - y_{k,\varepsilon}| - \ln r_{k,\varepsilon}}{\ln\sqrt{r_{k,\varepsilon}} - \ln r_{k,\varepsilon}} & \text{if } r_{k,\varepsilon} \le |y - y_{k,\varepsilon}| < \sqrt{r_{k,\varepsilon}}, \quad k = 1, \dots, p. \end{cases}$$

We have $\Phi_\varepsilon = 0$ in Q_ε since by definition

$$Q_\varepsilon \subset \bigcup_{k=1}^{p} \widehat{Q}_{k,\varepsilon} \subset \bigcup_{k=1}^{p} B(y_{k,\varepsilon}, r_{k,\varepsilon}).$$

The sequence Φ_ε strongly converges to 1 in $L^2_\#(Y_2)$ since

$$0 \le \Phi_\varepsilon \le 1 \quad \text{and} \quad \left| \bigcup_{k=1}^{p} B(y_{k,\varepsilon}, \sqrt{r_{k,\varepsilon}}) \right| \underset{\varepsilon \to 0}{\longrightarrow} 0.$$

We also have

$$\int_{Y_2} |\nabla \Phi_\varepsilon|^2 = \sum_{k=1}^{p} \frac{4\pi}{|\ln r_{k,\varepsilon}|} \underset{\varepsilon \to 0}{\longrightarrow} 0,$$

whence $\nabla \Phi_\varepsilon$ strongly converges to 0 in $L^2_\#(Y_2)$. Therefore the capacity condition (2.10) is satisfied with the sequence $\frac{1}{\int_{Y_2} \Phi_\varepsilon} \Phi_\varepsilon$.

Proof of (2.11). Let $\mu \in \mathbb{R}^2$. Since the distance between two sets $\widehat{Q}_{h,\varepsilon}, \widehat{Q}_{k,\varepsilon}$ is greater than a positive constant and the diameter of $\widehat{Q}_{k,\varepsilon}$ tends to zero, there exist fixed open balls $B(y_k, r_k), B(y_k, r'_k)$ in the torus Y_2, with radius $r_k < r'_k < \frac{1}{2}$, such that $\widehat{Q}_{k,\varepsilon} \subset B(y_k, r_k)$ for any $k = 1, \dots, p$ and the balls $B(y_k, r'_k)$ are two-by-two disjoint. We then define a smooth Y_2-periodic function Ψ_μ by

$$\begin{cases} \Psi_\mu(y) := \mu \cdot (y - \tau) & \text{if } y \in \tau + B(y_k, r_k) \quad \text{for any } \tau \in \mathbb{Z}^2 \text{ and } k := 1, \dots, p, \\ \Psi_\mu(y) := 0 & \text{if } y \notin \displaystyle\bigcup_{\tau \in \mathbb{Z}^2} \bigcup_{k=1}^{p} (\tau + B(y_k, r'_k)). \end{cases}$$

This definition is consistent since all the balls $\tau + B(y_k, r_k)$ are two-by-two disjoint closures in $\mathbb{R}^2$ by construction. Therefore condition (2.11) is also satisfied, which concludes the proof of Lemma 3.3.

3.3. Proof of Proposition 2.7. *Proof of part* (i) *of Proposition* 2.7.
Case $n=1$. Let $A_{i,\varepsilon}$, $i=1,2$, be the function defined by

$$A_{i,\varepsilon} := \begin{cases} 1 & \text{in } Y_2 \setminus Q_{i,\varepsilon}, \\ \alpha_{i,\varepsilon} & \text{in } Q_{i,\varepsilon}. \end{cases} \tag{3.53}$$

By the definitions (3.31) and (3.32), the eigenfunction $V_{1,\varepsilon}$ satisfies for $i=1,2$

$$\begin{aligned} &\int_{Y_2} A_{i,\varepsilon} \left(V_{1,\varepsilon} - \frac{\int_{Y_2} A_{i,\varepsilon} V_{1,\varepsilon}}{\int_{Y_2} A_{i,\varepsilon}} \right)^2 \\ &\underset{\varepsilon\to 0}{\longrightarrow} \left(c_1 - \frac{c_1 + \kappa_i c_{1,i}}{1+\kappa_i} \right)^2 + \kappa_i \left(c_{1,i} - \frac{c_1 + \kappa_i c_{1,i}}{1+\kappa_i} \right)^2 = d\,(c_1 - c_{1,i})^2, \end{aligned}$$

where d is a positive constant. We cannot have $c_1 = c_{1,1} = c_{1,2}$ since the orthonormality equalities (3.40) for $N=2$ are the following:

$$c_i + \kappa_1\, c_{i,1} + \kappa_2\, c_{i,2} = 0 \quad \text{and} \quad c_i^2 + \kappa_1\, c_{i,1}^2 + \kappa_2\, c_{i,2}^2 = 1 \quad \text{for } i=1,2. \tag{3.54}$$

Hence there exist $i=1,2$ and a positive constant c such that

$$I_{i,\varepsilon} := \int_{Y_2} A_{i,\varepsilon} \left(V_{1,\varepsilon} - \frac{\int_{Y_2} A_{i,\varepsilon} V_{1,\varepsilon}}{\int_{Y_2} A_{i,\varepsilon}} \right)^2 > c.$$

Then the first nonzero eigenvalue $\Lambda_{1,i}(\varepsilon)$ of the problem (2.7) defined with the function $A_{i,\varepsilon}$ satisfies the estimate

$$\Lambda_1(\varepsilon) = \int_{Y_2} A_\varepsilon\, |\nabla V_{1,\varepsilon}|^2 \geq \int_{Y_2} A_{i,\varepsilon}\, |\nabla V_{1,\varepsilon}|^2 = I_{i,\varepsilon}\, \frac{\int_{Y_2} A_{i,\varepsilon}\, |\nabla V_{1,\varepsilon}|^2}{I_{i,\varepsilon}} \geq c\, \Lambda_{1,i}(\varepsilon). \tag{3.55}$$

Now assume that $\delta = +\infty$, i.e., $\varepsilon^2\, |\ln r_\varepsilon| \to 0$. Proposition 2.4 of [3] for one fiber by cell implies that $\frac{\Lambda_{1,i}(\varepsilon)}{\varepsilon^2} \to +\infty$, whence by (3.55) $\frac{\Lambda_1(\varepsilon)}{\varepsilon^2} \to +\infty$. Therefore $n=1$ by the definition of n in part (i) of Theorem 2.4.

Inversely assume that $\delta < +\infty$, i.e., $\lim\limits_{\varepsilon\to 0} \varepsilon^2\, |\ln r_\varepsilon| > 0$. Let $\hat{W}_{i,\varepsilon}$, $i=1,2$, be the Y_2-periodic function defined for $y \in Y_2$ by

$$\hat{W}_{i,\varepsilon}(y) := \begin{cases} 0 & \text{if } r < r_\varepsilon \text{ or equivalently } y \in Q_{i,\varepsilon}, \\ 1 & \text{if } r > \rho_\varepsilon \text{ where } |\ln \rho_\varepsilon| \ll |\ln r_\varepsilon|, \\ \dfrac{\ln r - \ln r_\varepsilon}{\ln \rho_\varepsilon - \ln r_\varepsilon} & \text{if } r_\varepsilon \leq r \leq \rho_\varepsilon, \end{cases} \tag{3.56}$$

where r represents the distance between y and the center of the disk $Q_{i,\varepsilon}$.

It is easy to check that

$$\|\nabla(\hat{W}_{1,\varepsilon}\hat{W}_{2,\varepsilon})\|_\varepsilon^2 \leq \frac{c}{|\ln r_\varepsilon|} = O(\varepsilon^2),$$

and since $\hat{W}_{1,\varepsilon}\hat{W}_{2,\varepsilon} = 0$ in Q_ε we also have, in the sense of definition (2.25),

$$\hat{W}_{1,\varepsilon}\hat{W}_{2,\varepsilon} \not\approx_\varepsilon \frac{\langle \hat{W}_{1,\varepsilon}\hat{W}_{2,\varepsilon}, 1\rangle_\varepsilon}{\langle 1,1\rangle_\varepsilon} \underset{\varepsilon\to 0}{\longrightarrow} \frac{1}{1+\kappa_1+\kappa_2} > 0.$$

From both previous estimates we deduce that

$$\Lambda_1(\varepsilon) \le \frac{\|\nabla(\hat{W}_{1,\varepsilon}\hat{W}_{2,\varepsilon})\|_\varepsilon^2}{\left\|\hat{W}_{1,\varepsilon}\hat{W}_{2,\varepsilon} - \frac{\langle \hat{W}_{1,\varepsilon}\hat{W}_{2,\varepsilon},1\rangle_\varepsilon}{\langle 1,1\rangle_\varepsilon}\right\|_\varepsilon^2} = O(\varepsilon^2). \tag{3.57}$$

Moreover by (3.55) we have $\Lambda_1(\varepsilon) \ge c\,\Lambda_{1,1}(\varepsilon)$. By comparing the result of [1] to Theorem 2.1 (see also the proof of part (ii) of Proposition 2.7) we also have $\frac{\Lambda_{1,1}(\varepsilon)}{\varepsilon^2} \to \delta$. From these results we deduce the estimate $\Lambda_1(\varepsilon) \ge c\,\varepsilon^2$, which, combined with (3.57), implies that (up to a subsequence) $\frac{\Lambda_1(\varepsilon)}{\varepsilon^2} \to \lambda_1 \in]0,+\infty[$. We then obtain $n \ge 2$ thanks to part (i) of Theorem 2.4. Therefore the first equivalence of (2.37) has been proved.

Case $n = 2$. Assume that the left-hand side of the second implication of (2.37) holds true. By the first equivalence of (2.37) and part (i) of Theorem 2.4 the integer n is equal to 2 or 3. Then assume by contradiction that $n = 3$.

We will use the following result.

LEMMA 3.4. *There exists a positive constant* c *such that*

$$\forall\, V \in H^1_\#(Y_2), \quad \left| \fint_{Q_{1,\varepsilon}} V - \fint_{Q_{2,\varepsilon}} V \right| \le c\left(\sqrt{\ln(d_\varepsilon + r_\varepsilon) - \ln r_\varepsilon} + 1\right) \|\nabla V\|_{L^2(Y_2)}. \tag{3.58}$$

By studying the cases where the limit of $\frac{d_\varepsilon}{r_\varepsilon}$ is finite or is not, it is easy to prove the implication

$$\lim_{\varepsilon\to 0}\left|\frac{\ln d_\varepsilon}{\ln r_\varepsilon}\right| = \gamma \ge 1 \quad\Rightarrow\quad \lim_{\varepsilon\to 0}\left|\frac{\ln(d_\varepsilon + r_\varepsilon)}{\ln r_\varepsilon}\right| = 1,$$

and since the first limit is satisfied by assumption, the second one thus holds true. Then, thanks to estimate (3.58) applied to the eigenfunction $V_{j,\varepsilon}$, $j = 1, 2$, we obtain

$$\left| \fint_{Q_{1,\varepsilon}} V_{j,\varepsilon} - \fint_{Q_{2,\varepsilon}} V_{j,\varepsilon} \right| = o\left(\sqrt{|\ln r_\varepsilon|}\right) \|\nabla V_{j,\varepsilon}\|_{L^2(Y_2)} = o\left(\varepsilon\sqrt{|\ln r_\varepsilon|}\right) \underset{\varepsilon\to 0}{\longrightarrow} 0,$$

whence $c_{j,1} = c_{j,1}$ for $j = 1, 2$. These equalities, combined with the orthonormality equalities (3.54), yield

$$c_1 + (\kappa_1 + \kappa_2)\,c_{1,1} = c_2 + (\kappa_1 + \kappa_2)\,c_{2,1} = c_1\,c_2 + (\kappa_1 + \kappa_2)\,c_{1,1}\,c_{2,1} = 0,$$

which implies $c_{1,1}\,c_{2,1} = 0$. For instance, $c_{1,1} = 0$, whence $c_1 = c_{1,1} = c_{1,2} = 0$, which contradicts the second equality of (3.54). The second implication of (2.37) has thus been proved.

Case $n = 3$. Assume that the left-hand side of the third implication of (2.37) holds true. The integer n is still equal to 2 or 3. Assume by contradiction that $n = 2$, whence by Theorem 2.4 we have $\frac{\Lambda_2(\varepsilon)}{\varepsilon^2} \to +\infty$. Let $i = 1, 2$. Since by definition (3.56) we have $\|\nabla\hat{W}_{i,\varepsilon}\|_\varepsilon = O(\varepsilon)$, the Courant–Fisher formula for $\Lambda_2(\varepsilon)$ yields

$$\left\| \hat{W}_{i,\varepsilon} - \frac{\langle \hat{W}_{i,\varepsilon},1\rangle_\varepsilon}{\langle 1,1\rangle_\varepsilon} - \langle \hat{W}_{i,\varepsilon}, V_{1,\varepsilon}\rangle_\varepsilon\, V_{1,\varepsilon} \right\|_\varepsilon \le c\,\Lambda_2(\varepsilon)^{-\frac{1}{2}} \underset{\varepsilon\to 0}{\longrightarrow} 0,$$

which implies (according to notation (2.25)) that

$$\hat{W}_{i,\varepsilon} \approx_\varepsilon a_i + b_i\, V_{1,\varepsilon}, \quad \text{where } a_i, b_i \text{ are two constants.}$$

Since by hypothesis $d_\varepsilon \gg r_\varepsilon$ we also have for $i \neq j$

$$\fint_{Q_{j,\varepsilon}} \hat{W}_{i,\varepsilon} = \frac{\ln(d_\varepsilon + r_\varepsilon) - \ln r_\varepsilon}{\ln \rho_\varepsilon - \ln r_\varepsilon} + o(1) \underset{\varepsilon\to 0}{\longrightarrow} 1 - \gamma.$$

Therefore, for $j \neq i$, a_i, b_i satisfy

$$\begin{cases} 1 = a_i + b_i\, c_1 & \text{by the limit in } Y_2 \setminus Q_\varepsilon, \\ 0 = a_i + b_i\, c_{1,i} & \text{by the limit in } Q_{i,\varepsilon}, \\ 1-\gamma = a_i + b_i\, c_{1,j} & \text{by the limit in } Q_{j,\varepsilon}, \end{cases}$$

whence $c_{1,j} = \gamma\, c_{1,i} + (1-\gamma)\, c_1$. Similarly with $\hat{W}_{j,\varepsilon}$ we obtain $c_{1,i} = \gamma\, c_{1,j} + (1-\gamma)\, c_1$. Since $\gamma \neq 1$ we then deduce from both previous equalities $c_{1,i} = c_{1,j}$ and $c_1 = c_{1,1} = c_{1,2} = 0$, which again contradicts the second equality of (3.54).

Proof of Lemma 3.4. Let $R_{i,\varepsilon}$, $i = 1, 2$, be the disk of radius $d_\varepsilon + r_\varepsilon$ and of the same center as $Q_{i,\varepsilon}$. On the first side by using polar coordinates it is easy to check that for any $V \in H^1_\#(Y_2)$ we have

$$\left| \fint_{\partial Q_{i,\varepsilon}} V - \fint_{\partial R_{i,\varepsilon}} V \right| \leq c \sqrt{\ln(d_\varepsilon + r_\varepsilon) - \ln r_\varepsilon}\, \|\nabla V\|_{L^2(R_{i,\varepsilon})}.$$

We also have

$$\left| \fint_{Q_{i,\varepsilon}} V - \fint_{\partial Q_{i,\varepsilon}} V \right| \leq c\, \|\nabla V\|_{L^2(Q_{i,\varepsilon})},$$

and the same estimate for the disk $R_{i,\varepsilon}$. From these estimates we easily deduce the new estimate,

$$\left| \fint_{Q_{i,\varepsilon}} V - \fint_{R_{i,\varepsilon}} V \right| \leq c \left(\sqrt{\ln(d_\varepsilon + r_\varepsilon) - \ln r_\varepsilon} + 1 \right) \|\nabla V\|_{L^2(R_{i,\varepsilon})}. \tag{3.59}$$

On the other side let R_ε be the disk of radius $3\,(d_\varepsilon + r_\varepsilon)$ whose center is the middle of both centers of $Q_{1,\varepsilon}, Q_{2,\varepsilon}$. In particular we have $R_{1,\varepsilon} \cup R_{2,\varepsilon} \subset R_\varepsilon$. By rescaling with the scale $(d_\varepsilon + r_\varepsilon)$ we obtain that there exists a positive constant c such that

$$\left| \fint_{R_{1,\varepsilon}} V - \fint_{R_{2,\varepsilon}} V \right| \leq c\, \|\nabla V\|_{L^2(R_\varepsilon)}.$$

This, combined with (3.59), implies the desired estimate (3.58), which concludes the proof of Lemma 3.4.

Proof of part (ii) *of Proposition* 2.7. We will use the result of Lemma 1 in [5], which consists of the following estimate satisfied by the function $\hat{W}_{i,\varepsilon}$, $i = 1, 2$, defined by (3.56): for any $V \in H^1_\#(Y_2)$,

$$\begin{aligned} &\left| \int_{Y_2} \nabla \hat{W}_{i,\varepsilon} \cdot \nabla V - \frac{2\,\pi}{|\ln r_\varepsilon|} \left(\fint_{Y_2 \setminus Q_{i,\varepsilon}} V - \fint_{Q_{i,\varepsilon}} V \right) \right| \\ &\leq \frac{C}{|\ln r_\varepsilon|} \left(\sqrt{|\ln \rho_\varepsilon|}\, \|\nabla V\|_{L^2(Y_2)} + \frac{1}{r_\varepsilon}\, \|\nabla V\|_{L^2(Q_{i,\varepsilon})} \right). \end{aligned} \tag{3.60}$$

Since $\gamma = 0$ in (2.36) we can choose ρ_ε in (3.56) such that $\rho_\varepsilon \ll d_\varepsilon$. By this choice we have $\nabla \hat{W}_{i,\varepsilon} = 0$, $i = 1, 2$, in Q_ε, whence by the definition (2.7) of the eigenfunctions $V_{j,\varepsilon}$, $j = 1, 2$, we have

$$\int_{Y_2} \nabla \hat{W}_{i,\varepsilon} \cdot \nabla V_{j,\varepsilon} = \langle \nabla \hat{W}_{i,\varepsilon}, \nabla V_{j,\varepsilon} \rangle_\varepsilon = \Lambda_j(\varepsilon) \, \langle \hat{W}_{i,\varepsilon}, V_{j,\varepsilon} \rangle_\varepsilon.$$

Then by taking $V := V_{j,\varepsilon}$ in estimate (3.60) the condition $\delta < +\infty$, combined with the assumption of (2.21) $\alpha_{i,\varepsilon} \, \pi r_\varepsilon^2 \to \kappa_i > 0$ and the assumption (2.29) satisfied by the eigenvalues, implies that for any $i, j = 1, 2$,

$$\begin{aligned} &\frac{\Lambda_j(\varepsilon)}{\varepsilon^2} \, \langle \hat{W}_{i,\varepsilon}, V_{j,\varepsilon} \rangle_\varepsilon - \frac{2\,\pi}{\varepsilon^2 \, |\ln r_\varepsilon|} \left(\fint_{Y_2 \setminus Q_{i,\varepsilon}} V_{j,\varepsilon} - \fint_{Q_{i,\varepsilon}} V_{j,\varepsilon} \right) \\ &= O\left(\varepsilon \sqrt{|\ln \rho_\varepsilon|} + \varepsilon \right) \underset{\varepsilon \to 0}{\longrightarrow} 0. \end{aligned} \tag{3.61}$$

Since $\hat{W}_{i,\varepsilon} = 0$ in $Q_{i,\varepsilon}$ and $\hat{W}_{3-i,\varepsilon} = 1$ in $Q_{3-i,\varepsilon}$ by the choice of ρ_ε, by passing to the limit in (3.61) with definitions (3.31), (3.32) we thus have

$$\lambda_j \, (c_j + \kappa_{3-i} \, c_{j,3-i}) - \delta \, (c_j - c_{j,i}) = 0,$$

which, combined with the first equality of (3.54), yields

$$\lambda_j \, \kappa_i \, c_{j,i} = \delta \, (c_{j,i} - c_j) \quad \text{for } i, j = 1, 2. \tag{3.62}$$

We have $c_{j,i} \neq 0$. Otherwise (3.62), combined with the first equality of (3.54), leads to a contradiction. Then equalities (3.62) for $i = 1, 2$ and $\kappa_1 = \kappa_2$ imply that $c_j = 0$ or $c_{j,1} = c_{j,2}$. Since $\lambda_1 \leq \lambda_2$ we thus obtain with (3.54)

$$c_1 = 0, \; c_{1,1} = -c_{1,2} = \frac{\pm 1}{\sqrt{2\kappa_1}} \;\text{ and }\; c_2 = - \pm \sqrt{\frac{2\kappa_1}{1+2\kappa_1}}, \; c_{2,1} = -c_{2,2} = \frac{\pm 1}{\sqrt{2\kappa_1 + 4\kappa_1^2}},$$

as well as the values (2.38) for λ_1, λ_2. Finally by the definitions (2.35) of the constant $\hat{\kappa}_k$ with respect to κ_i and the definitions (3.37) of the constant $\hat{c}_{k,i}$ with respect to $c_{i,j}$, we easily deduce the desired limit system (2.39) from the general one, (2.34) of Theorem 2.4. Proposition 2.7 has thus been proved.

REFERENCES

[1] M. Bellieud and G. Bouchitté, *Homogenization of elliptic problems in a fiber reinforced structure. Nonlocal effects*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 26 (1998), pp. 407–436.

[2] A. Beurling and J. Deny, *Espaces de Dirichlet*, Acta Math., 99 (1958), pp. 203–224.

[3] M. Briane, *Homogenization of non-uniformly bounded operators: Critical barrier for nonlocal effects*, Arch. Ration. Mech. Anal., 164 (2002), pp. 73–101.

[4] M. Briane, *Homogenization in general periodically perforated domains by a spectral approach*, Calc. Var. Partial Differential Equations, 15 (2002), pp. 1–24.

[5] M. Briane and N. Tchou, *Fibered microstructures for some nonlocal Dirichlet forms*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 30 (2001), pp. 681–711.

[6] G. Buttazzo and G. Dal Maso, *Γ-limits of integral functionals*, J. Anal. Math., 37 (1980), pp. 145–185.

[7] L. Carbone and C. Sbordone, *Some properties of Γ-limits of integral functionals*, Ann. Mat. Pura Appl. (4), 122 (1979), pp. 1–60.

[8] M. Camar-Eddine and P. Seppecher, *Non-local interactions resulting from the homogenization of a linear diffusive medium*, C. R. Acad. Sci. Paris Sér. I Math., 332 (2001), pp. 485–490.

[9] C. Pideri and P. Seppecher, *A second gradient material resulting from the homogenization of an heterogeneous linear elastic medium*, Contin. Mech. Thermodyn., 9 (1997), pp. 241–257.
[10] G. Dal Maso, *An introduction to* Γ*-convergence*, Birkhäuser Boston, Boston, 1993.
[11] V.N. Fenchenko and E.Ya. Khruslov, *Asymptotic behavior of solutions of differential equations with a strongly oscillating coefficient matrix that does not satisfy a uniform boundedness condition*, Dokl. Akad. Nauk Ukrain. SSR Ser. A, 4 (1981), pp. 24–27 (in Russian).
[12] E.Ya. Khruslov, *Homogenized models of composite media*, in Composite Media and Homogenization Theory, G. Dal Maso and G. F. Dell'Antonio, eds., Progr. Nonlinear Differential Equations Appl. 5, Birkhäuser Boston, Boston, 1991, pp. 159–182.
[13] U. Mosco, *Composite media and asymptotic Dirichlet forms*, J. Funct. Anal., 123 (1994), pp. 368–421.
[14] F. Murat and L. Tartar, *H-convergence*, in Topics in the Mathematical Modelling of Composite Materials, L. Cherkaev and R.V. Kohn, eds., Progr. Nonlinear Differential Equations Appl., Birkhäuser Boston, Boston, 1998, pp. 21–43.
[15] L. Tartar, *Cours Peccot*, Collège de France, 1977; partially appears in [14].

SIAM J. MATH. ANAL.
Vol. 35, No. 1, pp. 61–97

MINIMIZING FLOWS FOR THE MONGE–KANTOROVICH PROBLEM*

SIGURD ANGENENT†, STEVEN HAKER‡, AND ALLEN TANNENBAUM§

Abstract. In this work, we formulate a new minimizing flow for the optimal mass transport (Monge–Kantorovich) problem. We study certain properties of the flow, including weak solutions as well as short- and long-term existence. Optimal transport has found a number of applications, including econometrics, fluid dynamics, cosmology, image processing, automatic control, transportation, statistical physics, shape optimization, expert systems, and meteorology.

Key words. optimal transport, gradient flows, weak solutions, image registration, medical imaging

AMS subject classifications. 49Q20, 49J45, 94A08

DOI. 10.1137/S0036141002410927

1. Introduction. In this paper, we derive a novel gradient descent flow for the computation of the optimal transport map (when it exists) in the Monge–Kantorovich framework. Besides being quite useful for the efficient computation of the transport map, we believe that the flow presented here is quite interesting from a theoretical point of view as well. In the present work, we undertake a study of some of its key properties.

The *mass transport problem* was first formulated by Monge in 1781 and concerned finding the optimal way, in the sense of minimal transportation cost, of moving a pile of soil from one site to another. This problem was given a modern formulation in the work of Kantorovich [13] and so is now known as the *Monge–Kantorovich problem.* We recall the formulation of the Monge–Kantorovich problem for smooth densities and domains in Euclidean space. For more general measures, see [1]. Let Ω_0 and Ω_1 be two diffeomorphic connected subdomains of $\mathbb{R}^d$, with smooth boundaries, and let μ_0, μ_1 be Borel measures on Ω_0 and Ω_1, each with a positive density function μ_0 and μ_1, respectively. We assume

$$\mu_0(\Omega_0) = \mu_1(\Omega_1),$$

i.e.,

$$\int_{\Omega_0} \mu_0(x)dx = \int_{\Omega_0} \mu_1(x)dx,$$

so that the same total mass is associated with Ω_0 and Ω_1, and we consider diffeomorphisms $u : \Omega_0 \to \Omega_1$ which map one density to the other in the sense that

$$\mu_0 = \det(\nabla u)\, \mu_1 \circ u, \tag{1}$$

*Received by the editors July 3, 2002; accepted for publication (in revised form) November 20, 2002; published electronically June 10, 2003. This work was supported in part by grants from the National Science Foundation, Air Force Office of Scientific Research, MRI, and a Lady Davis Fellowship through the Technion, Israel Institute of Technology.

http://www.siam.org/journals/sima/35-1/41092.html

†Department of Mathematics, University of Wisconsin, Madison, WI 53706 (angenent@math.wisc).

‡Department of Radiology, Surgical Planning Laboratory, Brigham and Women's Hospital, Boston, MA 02115 (haker@bwh.harvard.edu).

§Departments of Electrical and Computer and Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250 (tannenba@ece.gatech.edu).

where $\det \nabla u$ denotes the determinant of the Jacobian map ∇u. This is the well-known *Jacobian equation*, which constrains the mapping u to be mass preserving with respect to μ_0 and μ_1.

There may be many such mappings, and we want to pick out an optimal one in some sense. Accordingly, we define a generalized Monge–Kantorovich functional as

$$M(u) = \int \Phi(x, u(x))\, \mathrm{d}\mu_0(x),$$

where $\Phi : \bar{\Omega}_0 \times \bar{\Omega}_1 \to \mathbb{R}$ is a positive C^1 cost function. A Φ-*optimal mass preserving map*, when it exists, is a diffeomorphism which satisfies (1) and minimizes this integral.

In particular, the L^2 Monge–Kantorovich problem, corresponding to the cost function $\Phi(x, \xi) = \frac{1}{2}|x-\xi|^2$, has been studied in statistics, functional analysis, atmospheric sciences, automatic control, computer vision, statistical physics, and expert systems. See [3, 5, 8, 15, 16] and the references therein. This functional is seen to place a quadratic penalty on the distance the map u moves each bit of material, weighted by the material's mass. A fundamental theoretical result for the L^2 case [14, 4, 9] is that there is a unique optimal mass preserving u, and that this u is characterized as the gradient of a convex function p, i.e., $u = \nabla p$.

1.1. Reallocation measures. It turns out to be very convenient to use Kantorovich's generalization of the notion of a measure preserving map $u : (\Omega_0, \mu_0) \to (\Omega_1, \mu_1)$. Instead of considering a map u we introduce its graph

$$\{(x, u(x)) \mid x \in \Omega_0\} \subset \Omega_0 \times \Omega_1$$

and, more importantly, the measure

$$\gamma_u = (\mathrm{id} \times u)_{\#}\mu_0 \tag{2}$$

on $\Omega_0 \times \Omega_1$ supported on this graph.

The measures γ that arise in this way all satisfy[1]

$$\big(p_0\big)_{\#}\gamma = \mu_0 \quad \text{and} \quad \big(p_1\big)_{\#}\gamma = \mu_1. \tag{3}$$

We define $\mathfrak{X}$ to be the space of nonnegative Borel measures on $\bar{\Omega}_0 \times \bar{\Omega}_1$ which satisfy (3).

The Monge–Kantorovich cost functional extends in a natural way to the space of measures $\mathfrak{X}$ by

$$M(\gamma) = \int_{\Omega_0 \times \Omega_1} \Phi(x, y)\ \mathrm{d}\gamma(x, y).$$

We may think of a measure $\gamma \in \mathfrak{X}$ as a "multivalued map," which, rather than sending a point $x \in \Omega_0$ to *one* other point $u(x)$, assigns a probability measure P_x on the range space Ω_1 and "smears the point x out over Ω_1 according to the probability measure P_x." The measure γ is reconstructed from the family of probability measures $\{P_x\}$ by specifying

$$\int_{\Omega_0 \times \Omega_1} \phi(x, y)\ \mathrm{d}\gamma(x, y) = \int_{\Omega_0} \left\{ \int_{\Omega_1} \phi(x, y)\ \mathrm{d}P_x(y) \right\}\ \mathrm{d}\mu_0(x). \tag{4}$$

[1] If X and Y are sets with σ-algebras $\mathcal{M}$ and $\mathcal{N}$, and if $f : X \to Y$ is a measurable map, then we write $f_{\#}\mu$ for the pushforward of any measure μ on $(X, \mathcal{M})$, i.e., for any measurable $E \subset Y$ we define $f_{\#}\mu(E) = \mu\big(f^{-1}(E)\big)$.

See [1] for a rigorous measure-theoretic account of this way of decomposing γ. In this paper we will write for any bounded Borel measurable function $\phi : \Omega_0 \times \Omega_1 \to \mathbb{R}$

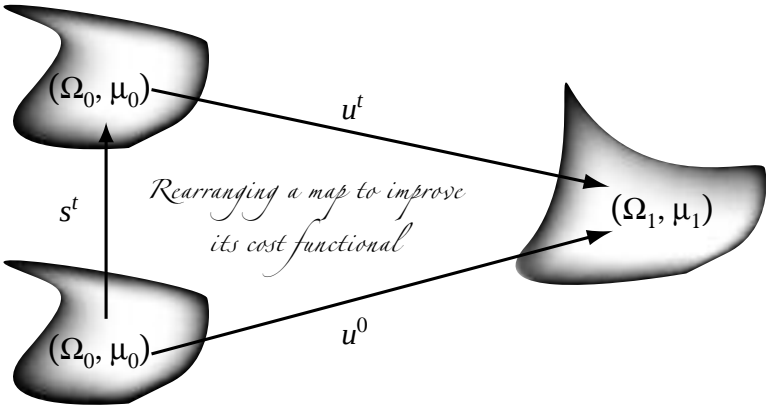
$$\mathbb{E}_\gamma(\phi(x,y) \mid x) = \int_{\Omega_1} \phi(x,y)\, \mathrm{d}P_x(y)$$

for the expectation of $\phi(x,\cdot)$ with respect to the probability measure P_x. See Lemma 3.1, where we define this expectation directly without using the probability measures P_x. The principal role of the expectation $\mathbb{E}_\gamma(\phi(x,y) \mid x)$ is as generalization of the expression $\phi(x,u(x))$. Indeed, when $\gamma = (\mathrm{id} \times u)_\# \mu_0$, then both expressions coincide.

1.2. The gradient flow. To reduce the Monge–Kantorovich cost $M(u)$ of a map $u^0 : \Omega_0 \to \Omega_1$ we "rearrange the points in the domain of the map"; i.e., we replace the map u^0 by a family of maps u^t for which one has $u^t \circ s^t = u^0$ for some family of diffeomorphisms $s^t : \Omega_0 \to \Omega_0$ (see figure below). If the initial map u^0 sends the measure μ_0 to μ_1 (if it satisfies (1)), and if the diffeomorphisms s^t preserve the measure μ_0, then the maps $u^t = u^0 \circ \left(s^t\right)^{-1}$ will also send μ_0 to μ_1. Thus the group $\mathrm{Diff}^1_{\mu_0}(\Omega_0)$ of C^1, μ_0 preserving diffeomorphisms acts on the space of measure preserving maps $u : (\Omega_0, \mu_0) \to (\Omega_1, \mu_1)$. The group action of $\mathrm{Diff}^1_{\mu_0}(\Omega_0)$ can be extended to an action on $\mathfrak{X}$ by

$$s \cdot \gamma = \left(s \times \mathrm{id}_{\Omega_1}\right)_\# \gamma.$$

Any sufficiently smooth family of diffeomorphism $s^t : \Omega_0 \to \Omega_0$ is determined by its velocity field, defined by $\partial_t s^t = v^t \circ s^t$.



In section 3 we compute the change in $M(\gamma^t)$ for measures $\gamma^t = s^t \cdot \gamma^0 \in \mathfrak{X}$ obtained by letting a family of diffeomorphisms $s^t \in \mathrm{Diff}^1_{\mu_0}(\Omega_0)$ act on an initial measure $\gamma^0 \in \mathfrak{X}$. We find that steepest descent is achieved by a family $s^t \in \mathrm{Diff}^1_{\mu_0}(\Omega_0)$, whose velocity is given by

$$v^t = -\frac{1}{\mu_0(x)} \mathcal{P}\left(\mathbb{E}_{\gamma^t}(\Phi_x \mid x)\right). \tag{5}$$

Here $\mathcal{P}$ is the Helmholtz projection, which extracts the divergence-free part of vector fields on Ω_0 (see section 7).

In the special case where the measures γ^t are given by graph measures $\gamma^t = \gamma_{u^t}$ as in (2) we get the following equations for the evolution of the map u^t. From $u^0 = u^t \circ s^t$ we get the transport equation

$$\frac{\partial u^t}{\partial t} + v^t \cdot \nabla u^t = 0. \tag{6}$$

The velocity field is still given by (5) above, but this now simplifies to

$$v^t = \frac{-1}{\mu_0(x)} \mathcal{P}\big\{\Phi_x(x, u^t(x))\big\}.$$

This equation, together with the transport equation (6), determines an initial value problem for the map u^t.

We will show the following.

THEOREM 1.1. *Let $0 < \alpha < 1$. For any $C^{1,\alpha}$, measure preserving initial map u^0 a smooth ($C^{1,\alpha}$) family of maps $\{u^t \mid 0 \le t < T\}$ exists such that the maps s^t generated by the vector field v^t given by* (5) *satisfy $u^0 = u^t \circ s^t$.*

The existence time T of the smooth solution depends on the $C^{1,\alpha}$ norm of the initial map u^0.

See Lemma 11.1 for more detail.

It is not clear if these smooth solutions exist for *all* $t > 0$ (we make no geometric assumptions on Ω_0 or the cost function Φ at all). To construct global solutions we modify the equation by introducing a smoothing operator $\mathcal{A}$. This operator acts on the space $\mathfrak{H}$ of all L^2 vector fields on Ω_0. We choose $\mathcal{A}$ to be an operator which approximates the identity and for which $\mathcal{A}w$ will always be smooth for all $w \in \mathfrak{H}$. The operators $\mathcal{A}$ we use are versions of a parabolic operator $\mathcal{A} = e^{\varepsilon\Delta}$. See section 8 for more detail.

Instead of considering the gradient flow generated by the velocity field (5), we smooth out v^t and consider

$$v^t = -\frac{1}{\mu_0(x)} \mathcal{P}\mathcal{A}^2\mathcal{P}(W^t) = -\frac{1}{\mu_0(x)} \mathcal{P}\mathcal{A}^2\mathcal{P}\left(\mathbb{E}_{\gamma^t}(\Phi_x \mid x)\right). \tag{7}$$

We refer to the corresponding initial value problem as *the regularized problem.* Since the velocity field here is smooth for any possible $\gamma^t \in \mathfrak{X}$, no singularities can occur, and we can prove the following.

THEOREM 1.2. *Under appropriate assumptions on the smoothing operator A solutions to the initial value problem exist for all time $t \ge 0$ and for any initial measure $\gamma^0 \in \mathfrak{X}$.*

See Theorem 9.1 for a more precise statement.

In fact, assuming the smoothing operator is injective, it follows that the initial value problem corresponding to the velocity field (7) generates a continuous semiflow on $\mathfrak{X}$ and that the Monge–Kantorovich functional $M(\gamma)$ acts as a Lyapunov function for this flow. Thus all orbits exist for all $t > 0$, and all orbits have ω-limit sets consisting of critical points only. Here a critical point of the flow is measure $\gamma \in \mathfrak{X}$ whose velocity field defined in (7) vanishes. Injectivity of $\mathcal{A}$ implies that critical points can be characterized independently of the smoothing operator $\mathcal{A}$: *A critical point is a measure $\gamma \in \mathfrak{X}$ for which*

$$\mathbb{E}_\gamma\big(\Phi_x \mid x\big) = \nabla p$$

for some Lipschitz continuous function $p : \Omega_0 \to \mathbb{R}$. These are precisely the measures $\gamma \in \mathfrak{X}$ whose Monge–Kantorovich cost $M(\gamma)$ cannot be reduced by the action of some $s \in \mathrm{Diff}^1_{\mu_0}(\Omega_0)$ infinitesimally close to the identity.

If the measure γ is given by $\gamma = (\mathrm{id} \times u)_{\#}\mu_0$ for some measure preserving $u : \Omega_0 \to \Omega_1$, then $\gamma = \gamma_u$ is a critical point exactly when the map u satisfies

$$\Phi_x(x, u(x)) = \nabla p(x) \quad \text{a.e. on } \Omega_0$$

for some Lipschitz function $p : \Omega_0 \to \mathbb{R}$. This is very important since it motivates our approach for finding a flow which in a certain sense kills the curl of a vector field (see our discussion in section 3.2). In particular, if the cost function is quadratic, $\Phi(x, y) = \frac{1}{2}|x - y|^2$, then a measure preserving map $u : \Omega_0 \to \Omega_1$ whose reallocation measure $\gamma_u \in \mathfrak{X}$ is a critical point also satisfies

$$u(x) = x - \nabla p(x)$$

for some Lipschitz function p.

Our gradient flows (both regularized and unregularized) move measures $\gamma \in \mathfrak{X}$ around on orbits of the group action $\mathrm{Diff}^1_{\mu_0}(\Omega_0) \times \mathfrak{X} \to \mathfrak{X}$.

A pertinent example is the group orbit of a C^1 diffeomorphism $\hat{u} : \bar{\Omega}_0 \to \bar{\Omega}_1$, or, rather, the measure γ_u associated to such a map. This orbit consists of all measures of the form $s \cdot \gamma_u = \gamma_{u \circ s^{-1}}$. Since any other diffeomorphism $\tilde{u} : \bar{\Omega}_0 \to \bar{\Omega}_1$ is of the form $\tilde{u} = \hat{u} \circ s^{-1}$ for some $s \in \mathrm{Diff}^1_{\mu_0}(\Omega_0)$ we see that the set

$$\big\{\gamma_u \mid u : \bar{\Omega}_0 \to \bar{\Omega}_1 \text{ is a } C^1 \text{ measure preserving diffeomorphism}\big\}$$

is exactly one orbit of the group action. So, if we have an initial measure $\gamma = \gamma_u$ which is generated by some map u and solve the initial value problem, then the solution we get will again consist of measures of the form $\gamma^t = \gamma_{u^t}$.

Unfortunately, such group orbits are not always closed, so if $\{\gamma^t = s^t \cdot \gamma^0 \mid t \geq 0\}$ is a trajectory of one of our gradient flows, then its ω-limit set might not be contained in the same orbit of the group action; i.e., if $\hat{\gamma}$ belongs to the ω-limit set, then it is possible that $\hat{\gamma}$ is not of the form $s \cdot \gamma^0$ for any $s \in \mathrm{Diff}^1_{\mu_0}(\Omega_0)$. In particular, if we start with $\gamma^0 = \gamma_{u^0}$, then the corresponding solution γ^t to the regularized flow will be of the form $\gamma^t = \gamma_{u^t}$ for a family of maps $u^t = u^0 \circ (s^t)^{-1}$, but, as $t \nearrow \infty$, the γ^t might converge to a measure $\tilde{\gamma} \in \mathfrak{X}$, which does not correspond to any map. (For example, if $\Omega_0 = \Omega_1$ is the unit disc, $\mu_0 = \mu_1$ is Lebesgue measure, $u^0 = \mathrm{id}_{\Omega_0}$, and s^t is defined by $s^t(z) = e^{it|z|}z$ in complex notation, then the measures γ_{u^t} converge weakly to $\bar{\gamma} \in \mathfrak{X}$. The limiting measure $\bar{\gamma}$ is described as in (4) with P_x the uniform distribution on the circle of radius $|x|$. In other words, instead of corresponding to a map, the limiting measure $\bar{\gamma}$ takes each point $x \in \Omega_0$ and spreads it out uniformly over the circle through x, centered at the origin. See Figure 1.)

We must therefore study the gradient flow(s) on all of $\mathfrak{X}$.

It turns out that there is always one stationary measure, namely,

$$\gamma_\times = \mu_0 \times \mu_1. \tag{8}$$

This measure takes each point in Ω_0 and spreads it out evenly (with a probability measure proportional to μ_1) over Ω_1. This measure must be a critical measure, for it is a fixed point for the group action; i.e., for all $s \in \mathrm{Diff}^1_{\mu_0}$ one has

$$s \cdot \gamma_\times = \big(s \times \mathrm{id}_{\Omega_1}\big)_{\#}\gamma_\times = (s_{\#}\mu_0) \times \mu_1 = \mu_0 \times \mu_1 = \gamma_\infty.$$

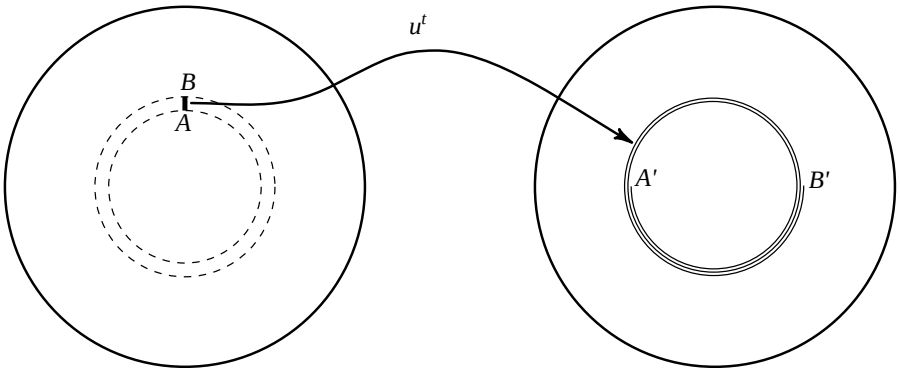


FIG. 1. *The map u^t spreads the short line segment AB out over the spiral $A'B'$.*

Therefore any of the gradient flows $\gamma \mapsto s^t \cdot \gamma$ we construct here act trivially on $\gamma_\times$.

Although we have no global existence result for the unregularized flow, we can choose a family of smoothing operators $\mathcal{A}_\varepsilon = e^{\varepsilon\Delta}$ which approximate the identity operator as $\varepsilon \searrow 0$ and consider the solutions $\{\gamma_\varepsilon^t \mid t \geq 0\}$ of the regularized flows whose existence we have already proved. We then show in section 10.2 that the γ_ε^t converge weakly to a family of measures $\tilde{\gamma}^t$ whose Monge–Kantorovich cost is decreasing and whose ω-limit set consists of critical measures (Proposition 10.1.)

1.3. Computations. Our interest in Monge–Kantorovich arose because of certain problems in computer vision and image processing, including image registration and image warping [2, 11, 12]. Image registration is the process of establishing a common geometric reference frame between two or more data sets possibly taken at different times. In [11, 12], we present a method for computing elastic registration maps based on the Monge–Kantorovich problem of optimal mass transport.

For image registration, it is natural to take $\Phi(x, y) = \frac{1}{2}|x - y|^2$ and $\Omega_0 = \Omega_1$ to be a rectangle. Extensive numerical computations show that the solution to the unregularized flow converges to a limiting map for a large choice of measures and initial maps. Indeed, in this case, we can write the minimizing flow in the following "nonlocal" form:

$$\frac{\partial u^t}{\partial t} = -\frac{1}{\mu_0}\left(u^t - \nabla\Delta^{-1}\operatorname{div}(u^t)\right) \cdot \nabla u^t. \tag{9}$$

In section 12, we give some details on our numerical methods as well as some illustrative examples.

2. Reallocation measures. The search for minimizers of $M(u)$ simplifies greatly if one suitably generalizes the notion of "mapping from Ω_0 to Ω_1." The standard way to do this in the present context is to identify the measure preserving map $u : (\Omega_0, \mu_0) \to (\Omega_1, \mu_1)$ with its graph, or, rather, with the Borel measure γ_u on $\Omega_0 \times \Omega_1$ defined by

$$\gamma_u(E) = \mu_0\big(\{x \in \Omega_0 : (x, u(x)) \in E\}\big).$$

This measure is supported on the graph of the map u; it is the pushforward of μ_0 under the map $\mathrm{id} \times u$, so $\gamma_u = (\mathrm{id} \times u)_\#(\mu_0)$.

The map u is measure preserving if and only if the measure γ_u satisfies $(p_0)_\#(\gamma_u) = \mu_0$ and $(p_1)_\#(\gamma_u) = \mu_1$, where $p_j : \Omega_0 \times \Omega_1 \to \Omega_j$ is the canonical projection. This prompts us to consider the space

$$\mathfrak{X} = \{\text{Borel measures } \gamma \geq 0 \text{ on } \Omega_0 \times \Omega_1 \ \mid (p_j)_\#\gamma = \mu_j \text{ for } j = 0, 1\}.$$

If the measure γ has a density, so that $d\gamma(x, y) = \mu(x, y)\mathrm{d}x\mathrm{d}y$, then $\gamma \in \mathfrak{X}$ exactly when

$$\int_{\Omega_0} \mu(x, y)\mathrm{d}x = \mu_1(y) \quad \text{for } \mu_1 \text{ almost all } y \in \Omega_1 \tag{10}$$

and

$$\int_{\Omega_1} \mu(x, y)\mathrm{d}y = \mu_0(x) \quad \text{for } \mu_0 \text{ almost all } x \in \Omega_0. \tag{11}$$

All measures $\gamma \in \mathfrak{X}$ have total mass

$$\gamma(\Omega_0 \times \Omega_1) = \mu_0(\Omega_0) = \mu_1(\Omega_1). \tag{12}$$

The space $\mathfrak{X}$ with the weak* topology is a compact metrizable space. (It is a closed and convex subset of the dual of $C^0(\Omega_0 \times \Omega_1)$.) The Monge–Kantorovich cost functional is linear on $\mathfrak{X}$. It is simply given by

$$M(\gamma) = \langle \gamma, \Phi \rangle = \int_{\Omega_0 \times \Omega_1} \Phi(x, y) d\gamma(x, y).$$

As such, there is always a minimizer for the cost functional (although in general it is only known to be a measure $\gamma \in \mathfrak{X}$, and it does not follow from general principles that γ is of the form γ_u for some measure preserving map).

3. Steepest descent. The group $\mathcal{G}$ of μ_0 measure preserving transformations on $s : \Omega_0 \to \Omega_0$ acts on $\mathfrak{X}$ by $s \cdot \gamma \mapsto (s \times \mathrm{id}_{\Omega_1})\gamma$. We propose to study a cost reducing flow on $\mathfrak{X}$ which is defined by the group action $\mathcal{G} \times \mathfrak{X} \to \mathfrak{X}$. Rather than applying arbitrary measurable maps $s \in \mathcal{G}$, we restrict ourselves to smooth (C^1) orientation preserving diffeomorphisms $s : \Omega_0 \to \Omega_0$.

3.1. The first variation. If we have a one-parameter family of μ_0 preserving C^1 diffeomorphisms $s^t : \Omega_0 \to \Omega_0$ with velocity field v^t, and we write $\gamma^t = s^t \cdot \gamma$ for some $\gamma \in \mathfrak{X}$, then the first variation of the Monge–Kantorovich cost functional is

$$\begin{aligned} \frac{dM(\gamma^t)}{dt} &= \frac{d}{dt} \int_{\Omega_0 \times \Omega_1} \Phi(x, y)\mathrm{d}(s^t \times \mathrm{id})_\#\gamma(x, y) \\ &= \frac{d}{dt} \int_{\Omega_0 \times \Omega_1} \Phi(s^t(x), y)\mathrm{d}\gamma(x, y) \\ &= \int_{\Omega_0 \times \Omega_1} v^t(s^t(x)) \cdot \Phi_x(s^t(x), y)\mathrm{d}\gamma(x, y) \\ &= \int_{\Omega_0 \times \Omega_1} v^t(x) \cdot \Phi_x(x, y)\mathrm{d}\gamma^t(x, y). \end{aligned} \tag{13}$$

LEMMA 3.1. *For any bounded measurable function $F : \Omega_0 \times \Omega_1 \to \mathbb{R}$ there exists a bounded measurable function $\tilde{F} : \Omega_0 \to \mathbb{R}$ for which*

$$\int_{\Omega_0 \times \Omega_1} \phi(x)F(x, y)\mathrm{d}\gamma(x, y) = \int_{\Omega_0} \phi(x)\tilde{F}(x)\mathrm{d}\mu_0(x)$$

holds for all $\phi \in L^1(\Omega_0; \mathrm{d}\mu_0)$.

Proof. The left-hand side defines a bounded linear functional on $L^1(\Omega_0, \mathrm{d}\mu_0)$, so the existence and uniqueness of $\tilde{F}$ is guaranteed. □

We will denote the function $\tilde{F}$ by

$$\tilde{F}(x) = \mathbb{E}_\gamma(F \mid x) \quad \text{or} \quad \tilde{F}(x) = \mathbb{E}_\gamma(F(x,y) \mid x). \tag{14}$$

If the measure γ has a density $\mu(x,y)$, then $\tilde{F}(x)$ is given by

$$\mathbb{E}_\gamma(F \mid x) = \int_{\Omega_1} F(x,y) \frac{\mu(x,y)}{\mu_0(x)} \mathrm{d}y. \tag{15}$$

Fubini's theorem implies that this integral exists for μ_0 almost all $x \in \Omega_0$. The condition $\mu(x,y)\mathrm{d}x\mathrm{d}y \in \mathfrak{X}$ implies that $\frac{\mu(x,y)}{\mu_0(x)}\mathrm{d}y$ is a probability measure on Ω_1 for every $x \in \Omega_0$, and $\tilde{F}(x)$ is just the expectation of $F(x,y)$ for this probability measure. This justifies the notation in (14).

If the measure is of the form $\gamma = \gamma_u$ for some measure preserving map $u : \Omega_0 \to \Omega_1$, then $\tilde{F}$ is given by

$$\mathbb{E}_\gamma(F \mid x) = F(x, u(x)). \tag{16}$$

One may think of (16) as a special case of (15) in which the "density" $\mu(x,y)$ is given by $\mu(x,y) = \mu_0(x)\delta(y - u(x))$ (δ being the Dirac delta-function). Here the probability measure $\frac{\mu(x,y)}{\mu_0(x)}\mathrm{d}y$ puts probability one at $y = u(x)$, and thus the expectation of $F(x,y)$ for this measure is just $F(x,u(x))$.

With this notation we now complete our computation (13) of the first variation:

$$\frac{dM(\gamma^t)}{dt} = \int_{\Omega_0} v^t(x) \cdot W^t(x) \mathrm{d}\mu_0(x), \tag{17}$$

where

$$W^t(x) = \mathbb{E}_{\gamma^t}(\Phi_x(x,y) \mid x). \tag{18}$$

When the measure $\gamma \in \mathfrak{X}$ is of the form $\gamma = \gamma_u$ for some map $u : \Omega_0 \to \Omega_1$, one has $\gamma^t = \gamma_{u^t}$ with $u^t \circ s^t = u$, and thus (18) reduces to

$$W^t(x) = \Phi_x(x, u^t(x)).$$

In the case of a quadratic cost function $\Phi(x,y) = \frac{1}{2}|x-y|^2$ but general measure $\gamma^t \in \mathfrak{X}$, one has

$$W^t(x) = \mathbb{E}_{\gamma^t}(x - y \mid x) = x - Y^t(x),$$

where

$$Y^t(x) = \mathbb{E}_{\gamma^t}(y|x)$$

is the expected y value to which the measure γ^t reallocates the point x.

If the cost function is quadratic, *and* if the measure γ^t is of the form γ_{u^t}, then we get $Y^t(x) = u^t(x)$, and hence

$$W^t(x) = x - u^t(x).$$

3.2. Steepest descent. To reduce the cost functional we choose the velocity field v^t so as to minimize $\int v^t \cdot W^t \mathrm{d}\mu_0$ subject to a constraint on $\|v^t\|_{L^2}$ (or some similar quadratic norm) and subject to the constraint that v^t preserve the measure μ_0, i.e., $\operatorname{div} \mu_0 v^t = 0$.

To this end we use the Helmholtz projection to split W^t into a gradient and its divergence-free part,

$$W^t = \nabla p^t + \mathcal{P}(W^t),$$

where

$$\operatorname{div} \mathcal{P}(W^t) = 0,$$

and where $\mathcal{P}(W^t)\,|_{\partial\Omega_0}$ is tangential to the boundary of Ω_0. Such a decomposition is always possible, and $\mathcal{P}$ can be interpreted as orthogonal projection in $L^2(\Omega_0) \otimes \mathbb{R}^d$. See section 7, where we discuss $\mathcal{P}$ in more detail.

If the velocity field satisfies $\operatorname{div} \mu_0 v^t = 0$, then we get

$$\begin{aligned} \frac{dM(\gamma^t)}{dt} &= \int_{\Omega_0} \mu_0(x) v^t(x) \cdot W^t \mathrm{d}x \\ &= \int_{\Omega_0} \mu_0(x) v^t(x) \cdot \{\nabla p^t + \mathcal{P}(W^t)\} \mathrm{d}x \\ &= \int_{\Omega_0} \{-p^t \nabla \cdot (\mu_0 v^t) + \mu_0 v^t \cdot \mathcal{P}(W^t)\} \mathrm{d}x \\ &= \int_{\Omega_0} \mu_0(x) v^t(x) \cdot \mathcal{P}(W^t) \mathrm{d}x. \end{aligned} \tag{19}$$

We choose the following velocity field:

$$v^t = -\frac{1}{\mu_0(x)} \mathcal{P}\mathcal{A}^2\mathcal{P}(W^t) = -\frac{1}{\mu_0(x)} \mathcal{P}\mathcal{A}^2\mathcal{P}\left(\mathbb{E}_{\gamma^t}(\Phi_x \mid x)\right). \tag{20}$$

Here $\mathcal{A}$ is an operator on the Hilbert space

$$\mathfrak{H} \stackrel{\text{def}}{=} L^2(\Omega_0) \otimes \mathbb{R}^d.$$

Throughout this paper we will assume that $\mathcal{A}$ satisfies

$$\mathcal{A} \text{ is a } \textit{bounded, symmetric, and injective} \text{ operator on } \mathfrak{H}. \tag{21}$$

Thus $\mathcal{A}^2$ is positive definite, and $\mathcal{P}\mathcal{A}^2\mathcal{P}$ is positive definite on divergence-free vector fields on Ω_0.

The most natural choice for $\mathcal{A}$ would be $\mathcal{A} = I_{\mathfrak{H}}$, the identity operator on $\mathfrak{H}$. In that case $\mathcal{P}\mathcal{A}^2\mathcal{P} = \mathcal{P}$, so that

$$v^t = -\frac{1}{\mu_0(x)} \mathcal{P}\mathbb{E}_{\gamma^t}(\Phi_x \mid x).$$

In what follows we are also interested in the case where the operator $\mathcal{A}$ is an approximate identity, e.g., $\mathcal{A}$ could be defined by running a heat equation for a short time, $\mathcal{A}f = e^{\varepsilon\Delta} f$. In section 8 we specify a class of operators $\mathcal{A}$ to which the theory in this paper is applicable.

3.3. Evolution equation for the measure γ^t. Let $\gamma^t = (s^t \times \mathrm{id})_\# \gamma^0$ for some initial measure $\gamma^0 \in \mathfrak{X}$. Here we compute the distributional time derivative of the γ^t assuming the diffeomorphisms s^t have velocity field v^t given by (20).

Let $\varphi \in C^1(\bar{\Omega}_0 \times \bar{\Omega}_1)$ be a test function. Then $\langle \gamma^t, \varphi \rangle = \big\langle \gamma^0, \varphi \circ (s^t \times \mathrm{id}) \big\rangle$, so that one has

$$\begin{aligned} \frac{d}{dt} \big\langle \gamma^t, \varphi \big\rangle &= \big\langle \gamma^0, (v^t \cdot \nabla_x \varphi) \circ (s^t \times \mathrm{id}) \big\rangle \\ &= \big\langle \gamma^t, v^t \cdot \nabla_x \varphi \big\rangle \\ &= \big\langle -\nabla \cdot (v^t \gamma^t), \varphi \big\rangle \,, \end{aligned} \tag{22}$$

where $\nabla_x f(x,y)$ represents the gradient in the $x \in \Omega_0$ variable for any function $f : \Omega_0 \times \Omega_1 \to \mathbb{R}$.

Thus we have found that the family of measures γ^t satisfies

$$\frac{\partial \gamma^t}{\partial t} + \nabla_x \cdot \big(v^t(x) \gamma^t \big) = 0 \tag{23}$$

in the sense of distributions. This equation, combined with (20), which prescribes v^t in terms of γ^t, gives an initial value problem for γ^t,

$$\boxed{\frac{\partial \gamma^t}{\partial t} = \nabla_x \cdot \left(\frac{1}{\mu_0(x)} \mathcal{P}\mathcal{A}^2\mathcal{P}\big(\mathbb{E}_{\gamma^t}(\Phi_x \mid x)\big)\gamma^t \right).} \tag{24}$$

3.4. A PDE for the map u^t. If the measure γ^t is given by $\gamma^t = \gamma_{u^t}$ for some family of measure preserving maps $u^t : (\Omega_0, \mu_0) \to (\Omega_1, \mu_1)$, then we have $u^0 = u^t \circ s^t$, so that the u^t satisfy the transport equation

$$\frac{\partial u^t}{\partial t} + v^t \cdot \nabla u^t = 0. \tag{25}$$

Since for $\gamma^t = \gamma_{u^t}$ one has

$$\mathbb{E}_{\gamma^t}(\Phi_x \mid x) = \Phi_x(x, u^t(x)),$$

the velocity field is given by

$$v^t = \frac{-1}{\mu_0(x)} \mathcal{P}\mathcal{A}^2\mathcal{P}\big\{\Phi_x(x, u^t(x))\big\}. \tag{26}$$

Together, (25) and (26) determine an evolution equation for the map u^t.

3.5. Evolution of the rearrangement s^t. We return to the case where γ^t is a general measure in $\mathfrak{X}$. Let us assume that the operator $\mathcal{P}\mathcal{A}^2\mathcal{P}$ can be represented as an integral operator with kernel $K(x,\xi)$, so that for any vector field $W \in L^2(\Omega_0; \mathbb{R}^n)$ one has

$$(\mathcal{P}\mathcal{A}^2\mathcal{P}W)(x) = \int_{\Omega_0} K(x,\xi) \cdot W(\xi) \mathrm{d}\xi. \tag{27}$$

Here $\mathrm{d}y$ is the Lebesgue measure, $K(x,y)$ is an $n \times n$ matrix-valued function, and $K(x,y) \cdot W(y)$ is pointwise matrix multiplication.

Self-adjointness of the operator $\mathcal{P}\mathcal{A}^2\mathcal{P}$ implies

$$K(x,\xi) = K(\xi,x)^T. \tag{28}$$

When $\mathcal{A}$ is the identity operator on $\mathfrak{H}$, the kernel $K(x,\xi)$ is a singular integral kernel. When $\mathcal{A}$ is given by solving a heat equation, $\mathcal{A}f = e^{\varepsilon\Delta}f$, then the kernel $K(x,\xi)$ is a $C^{1,\alpha}$ function on $\bar{\Omega}_0 \times \bar{\Omega}_1$. (See section 8 for more details.)

The velocity field is given by

$$\begin{aligned} v^t(x) &= \frac{-1}{\mu_0(x)} \int_{\Omega_0} K(x,\xi)\mathbb{E}_{\gamma^t}(\Phi_x(\xi,\eta) \mid \xi)\,\mathrm{d}\xi \\ &= \frac{-1}{\mu_0(x)} \int_{\Omega_0\times\Omega_1} K(x,\xi)\cdot\Phi_x(\xi,\eta)\,\frac{\mathrm{d}\gamma^t(\xi,\eta)}{\mu_0(\xi)} \\ &= -\int_{\Omega_0\times\Omega_1} \frac{K(x,\xi)}{\mu_0(x)\mu_0(\xi)}\cdot\Phi_x(\xi,\eta)\,\mathrm{d}\gamma^t(\xi,\eta). \end{aligned} \tag{29}$$

Since the rearrangement maps $s^t : \Omega_0 \to \Omega_0$ are related to the velocity field v^t by $\partial_t s^t = v^t \circ s^t$, we find the following integral-differential equation for s^t:

$$\begin{aligned} \frac{\partial s^t}{\partial t} &= -\int_{\Omega_0\times\Omega_1} \frac{K(s^t(x),\xi)}{\mu_0(s^t(x))\mu_0(\xi)}\cdot\Phi_x(\xi,\eta)\,\mathrm{d}\gamma^t(\xi,\eta) \\ &= -\int_{\Omega_0\times\Omega_1} \frac{K(s^t(x),s^t(\xi))}{\mu_0(s^t(x))\mu_0(s^t(\xi))}\cdot\Phi_x(s^t(\xi),\eta)\,\mathrm{d}\gamma^0(\xi,\eta), \end{aligned} \tag{30}$$

where we have used $\gamma^t = (s^t \times \mathrm{id})_{\#}\gamma^0$, with γ^0 the initial measure.

3.6. An alternative steepest descent flow. We can also derive a related flow in the following manner. Instead of using the Helmholtz projection to get a divergence-free vector field out of W^t, as we did in section 3.2, we set

$$\mu_0 v^t = \nabla \mathrm{div} W^t - \Delta W^t. \tag{31}$$

It is straightforward to check that in this case $\mathrm{div}(\mu_0 v^t) = 0$ and

$$\begin{aligned} M_t &= -\int_{\Omega_0} W^t \cdot \mu_0 v^t \,\mathrm{d}x \\ &= -\int_{\Omega_0} W^t \cdot \left(\nabla(\nabla\cdot W^t) - \Delta W^t\right)\,\mathrm{d}x \\ &= -\int_{\Omega_0} (W^t)^k((W^t)^l_{lk} - (W^t)^k_{ll})\,\mathrm{d}x, \end{aligned}$$

where we've used superscripts to denote vector components and subscripts for spatial derivatives, with the standard convention of summation over repeated indices. Integrating by parts, and ignoring the boundary for the sake of exposition, gives

$$\begin{aligned} M_t &= -\int_{\Omega_0} -(W^t)^k_l((W^t)^l_k - (W^t)^k_l)\,\mathrm{d}x \\ &= -\frac{1}{2}\int_{\Omega_0} ((W^t)^l_k - (W^t)^k_l)^2\,\mathrm{d}x \\ &= -\frac{1}{2}\int_{\Omega_0} |\operatorname{curl} W^t|^2\,\mathrm{d}x \\ &\le 0. \end{aligned} \tag{32}$$

If the measures γ^t are of the form $\gamma^t = (\mathrm{id}\times u^t)_\#(\mu_0)$, then we have $W^t = \Phi_x(x, u^t(x))$, resulting in the evolution equation

$$\frac{\partial u^t}{\partial t} = -\frac{1}{\mu_0(x)}\big(\nabla \mathrm{div}\Phi_x(x,u^t) - \Delta\Phi_x(x,u^t)\big)\cdot\nabla u^t \tag{33}$$

for u^t corresponding to (31), and (32) shows that at optimality we must again have curl $W^t = 0$, so $W^t = \nabla p$ for some function p.

For the quadratic cost function $\Phi(x,\xi) = \frac{1}{2}|x-\xi|^2$ we have $\Phi_x(x,\xi) = x - \xi$, so we get the following PDE:

$$\frac{\partial u^t}{\partial t} = \frac{1}{\mu_0(x)}\big(\nabla(\nabla\cdot u^t) - \Delta u^t\big)\cdot\nabla u^t.$$

We plan to study this equation in future work.

4. Weak solutions. Let $\gamma^t = (s^t\times\mathrm{id})_\#\gamma^0$ for some smooth family of diffeomorphisms $s^t : \bar\Omega_0 \to \bar\Omega_0$, whose velocity field satisfies (20).

In section 3.3 we observed that for any test function $\varphi \in C^1(\bar\Omega_0\times\bar\Omega_1)$ one has

$$\frac{d}{dt}\left\langle \gamma^t, \varphi\right\rangle = \left\langle \gamma^t, v^t\cdot\nabla_x\varphi\right\rangle.$$

Using (20) we get

$$\frac{d}{dt}\left\langle \gamma^t, \varphi\right\rangle = \left\langle \gamma^t, \frac{-1}{\mu_0(x)}\varphi_x\cdot\mathcal{P}\mathcal{A}^2\mathcal{P}\big(\mathbb{E}_{\gamma^t}(\Phi_x\mid x)\big)\right\rangle,$$

which implies

$$\begin{aligned}\frac{d}{dt}\left\langle \gamma^t, \varphi\right\rangle &= \big(\mathbb{E}_{\gamma^t}(\varphi_x\mid x), \mathcal{P}\mathcal{A}^2\mathcal{P}\mathbb{E}_{\gamma^t}(\Phi_x\mid x)\big)_{\mathfrak{H}}\\ &= \big(\mathcal{A}\mathcal{P}\mathbb{E}_{\gamma^t}(\varphi_x\mid x), \mathcal{A}\mathcal{P}\mathbb{E}_{\gamma^t}(\Phi_x\mid x)\big)_{\mathfrak{H}}.\end{aligned} \tag{34}$$

Integrate this in time, and you get

$$\int_{t_0}^{t_1}\big(\mathcal{A}\mathcal{P}\mathbb{E}_{\gamma^t}(\varphi_x\mid x), \mathcal{A}\mathcal{P}\mathbb{E}_{\gamma^t}(\Phi_x\mid x)\big)_{\mathfrak{H}}\,\mathrm{d}t = \left\langle\gamma^{t_0},\varphi\right\rangle - \left\langle\gamma^{t_1},\varphi\right\rangle. \tag{35}$$

For any measure $\gamma\in\mathfrak{X}$ and any $\varphi\in C^1(\bar\Omega_0\times\bar\Omega_1)$ the functions $\mathbb{E}_\gamma(\varphi_x\mid x)$ are bounded and measurable, and hence $\mathbb{E}_{\gamma^t}(\varphi_x\mid x)\in\mathfrak{H} = L^2(\Omega_0;\mathbb{R}^d)$ will always hold. Since $\mathcal{P}$ and $\mathcal{A}$ are bounded operators on $\mathfrak{H}$, both sides of the equation in (35) are defined for any weak* continuous family of measures $\gamma^t\in\mathfrak{X}$.

Definition 4.1 (weak solution). *A* weak solution *to the initial value problem* (24) *is a map* $t\in[0,T)\mapsto\gamma^t\in\mathfrak{X}$ *which is weak* continuous, and which satisfies* (35) *for all test functions* $\varphi\in C^1(\bar\Omega_0\times\bar\Omega_1)$ *and for all* $0\le t_0<t_1<T$.

If $\{\gamma^t, 0\le t<T\}$ is a weak solution, then (35) implies that $\langle\gamma^t,\varphi\rangle$ is an absolutely continuous function of t and that (34) holds for almost all t.

We could also introduce the notion of *classical solution* by requiring a classical solution to be a family of measures $\{\gamma^t, t\in[0,T)\}$ which is of the form $\gamma^t = (s^t\times\mathrm{id})_\#\gamma^0$ for some family of C^1 diffeomorphisms $s^t:\Omega_0\to\Omega_0$ whose velocity field $v^t = (\partial_t s^t)\circ(s^t)^{-1}$ satisfies $\mu_0 v^t = -\mathcal{P}\mathcal{A}^2\mathcal{P}\{\mathbb{E}_{\gamma^t}(\Phi_x\mid x)\}$.

LEMMA 4.2. *If the kernel $K(x,\xi)$ of the operator $\mathcal{P}\mathcal{A}^2\mathcal{P}$ is C^1, and if $\{\gamma^t, 0 \le t < T\}$ is a weak solution, then there is a C^1 family of diffeomorphisms $s^t : \bar\Omega_0 \to \bar\Omega_0$ such that $\gamma^t = (s^t \times \mathrm{id})_\# \gamma^0$; i.e., $\{\gamma^t\}$ is a classical solution.*

Proof. The velocity field v^t defined by the first line in (30), i.e.,

$$v^t(s) = -\int_{\Omega_0\times\Omega_1} \frac{K(s,\xi)}{\mu_0(s)\mu_0(\xi)} \cdot \Phi_x(\xi,\eta)\,\mathrm{d}\gamma^t(\xi,\eta),$$

is C^1 in $s \in \bar\Omega_0$. Therefore the ODE $\dot s = v^t(s)$ defines a unique family of diffeomorphisms s^t, $0 \le t < T$, with $s^0(x) \equiv x$. We now verify that $\gamma^t = (s^t \times \mathrm{id})_\# \gamma^0$.

Consider the measure $\lambda^t = \big((s^t)^{-1} \times \mathrm{id}\big)_\# \gamma^0$. We have $\gamma^t = (s^t \times \mathrm{id})_\# \lambda^t$, and $\lambda^0 = \gamma^0$. We will show that λ^t is constant.

For any test function $\varphi \in C^1(\bar\Omega_0 \times \bar\Omega_1)$ we differentiate

$$\frac{d}{dt}\left\langle \lambda^t, \varphi \circ (s^t \times \mathrm{id})\right\rangle = \frac{d}{dt}\left\langle \gamma^t, \varphi\right\rangle$$

(using the fact that γ^t is a weak solution)

$$\begin{aligned}
&= \left\langle \gamma^t, \varphi_x(x,y)\cdot v^t(x)\right\rangle \\
&= \left\langle \gamma^t, \frac{\partial \varphi \circ (s^t \times \mathrm{id})}{\partial t} \circ ((s^t)^{-1} \times \mathrm{id})\right\rangle \\
&= \left\langle \lambda^t, \frac{\partial}{\partial t}\big(\varphi \circ (s^t \times \mathrm{id})\big)\right\rangle.
\end{aligned}$$

On the other hand we also have

$$\frac{d}{dt}\left\langle \lambda^t, \varphi \circ (s^t \times \mathrm{id})\right\rangle = \left\langle \frac{\partial \lambda^t}{\partial t}, \varphi \circ (s^t \times \mathrm{id})\right\rangle + \left\langle \lambda^t, \frac{\partial}{\partial t}\big(\varphi \circ (s^t \times \mathrm{id})\big)\right\rangle.$$

We see that $\langle \partial_t \lambda^t, \varphi \circ (s^t \times \mathrm{id})\rangle$ vanishes for arbitrary test functions φ. Since s^t is C^1, this implies that $\tilde\varphi = \varphi \circ (s^t \times \mathrm{id})$ can also be any C^1 test function, and we conclude that $\partial_t \lambda^t = 0$.

Since $\lambda^0 = \gamma^0$, we get $\lambda^t = \gamma^0$ for all t, and finally, $\gamma^t = (s^t \times \mathrm{id})_\# \lambda^t = (s^t \times \mathrm{id})_\# \gamma^0$, as claimed. ∎

5. General energy bounds. By setting $\varphi = \Phi$ in (35) we get the following.

LEMMA 5.1 (energy identity). *For any weak solution $\{\gamma^t, t \in [0,T)\}$ and any $0 \le t_0 < t_1 < T$ one has*

$$M(\gamma^{t_1}) + \int_{t_0}^{t_1} \|\mathcal{A}\mathcal{P}\mathbb{E}_{\gamma^t}(\Phi_x \mid x)\|_{\mathfrak{H}}^2 \,\mathrm{d}t = M(\gamma^{t_0}).$$

This immediately leads to the following lemma.

LEMMA 5.2. *For any weak solution $\{\gamma^t, t \in [0,T)\}$ the Monge–Kantorovich cost functional is nonincreasing. It remains constant if and only if $\mathcal{P}\mathbb{E}_{\gamma^t}(\Phi_x \mid x) = 0$ for almost all $t \in [0,T)$, i.e., if and only if*

$$\mathbb{E}_{\gamma^t}(\Phi_x \mid x) = \nabla p^t$$

for some function $p^t : \Omega_0 \to \mathbb{R}$ and almost all $t \in [0,T)$.

Proof. It is clear that $M(\gamma^t)$ cannot increase. If $M(\gamma^{t_1}) = M(\gamma^{t_0})$ for certain $t_0 < t_1$, then (35) implies that $\mathcal{AP}\left(\mathbb{E}_{\gamma^t}(\Phi_x \mid x)\right) = 0$ for almost all $t_0 < t < t_1$. Since we assume the smoothing operator $\mathcal{A}$ is injective, this forces $\mathcal{P}\left(\mathbb{E}_{\gamma^t}(\Phi_x \mid x)\right) = 0$. ∎

LEMMA 5.3 (uniform Lipschitz bound). *If $\{\gamma^t, 0 \le t < T\}$ is a weak solution to (24), then for any test function $\varphi \in C^1(\bar{\Omega}_0 \times \bar{\Omega}_1)$ the function $t \mapsto \langle \gamma^t, \varphi \rangle$ is Lipschitz continuous, with*

$$\left| \frac{d\,\langle \mu_n^t, \varphi \rangle}{dt} \right| \le \|\mathcal{A}\|^2_{L(\mathfrak{H})} \left\| \frac{\partial \varphi}{\partial x} \right\|_{L^\infty} \left\| \frac{\partial \Phi}{\partial x} \right\|_{L^\infty}. \tag{36}$$

One could formulate this lemma as follows: weak solutions γ^t are uniformly Lipschitz continuous functions of t with values in $\left(C^1(\bar{\Omega}_0 \times \bar{\Omega}_1)\right)^*$ (the dual of C^1 functions on $\bar{\Omega}_0 \times \bar{\Omega}_1$), with Lipschitz constant depending only on the smoothing operator $\|\mathcal{A}\|$ and the cost function Φ.

Proof. This follows directly from (34) and the fact that for almost all $x \in \Omega_0$

$$|\mathbb{E}_\gamma(f(x,y) \mid x)| \le \operatorname*{ess\,sup}_{y \in \Omega_1} |f(x,y)|$$

for any $f \in L^\infty(\Omega_0 \times \Omega_1)$. ∎

LEMMA 5.4 (equicontinuity). *For any $\varphi \in C^0(\bar{\Omega}_0 \times \bar{\Omega}_1)$ there is a modulus of continuity $\sigma : \mathbb{R}_+ \to \mathbb{R}_+$ which depends only on φ, $\|\mathcal{A}\|_{L(\mathfrak{H})}$, $\|\Phi_x\|_{L^\infty}$, and the total mass $\mu_0(\Omega_0)$ such that*

$$\left| \langle \gamma^{t_1}, \varphi \rangle - \langle \gamma^{t_0}, \varphi \rangle \right| \le \sigma(|t_1 - t_0|).$$

Proof. For test functions $\tilde{\varphi} \in C^1(\bar{\Omega}_0 \times \bar{\Omega}_1)$ the previous lemma gives us a uniform Lipschitz bound. We now approximate our given $\varphi \in C^0$ by a $\tilde{\varphi} \in C^1$ and compute

$$\begin{aligned} \left| \langle \gamma^{t_1}, \varphi \rangle - \langle \gamma^{t_0}, \varphi \rangle \right| &= \left| \langle \gamma^{t_1} - \gamma^{t_0}, \varphi \rangle \right| \\ &\le \left| \langle \gamma^{t_1} - \gamma^{t_0}, \tilde{\varphi} \rangle \right| + \left| \langle \gamma^{t_1} - \gamma^{t_0}, \tilde{\varphi} - \varphi \rangle \right| \\ &\le C \|\tilde{\varphi}_x\|_\infty \delta + 2\mu_0(\Omega_0) \|\varphi - \tilde{\varphi}\|_\infty, \end{aligned}$$

where $C = \|\mathcal{A}\|^2 \|\Phi_x\|_\infty$, and where we have used the fact that all measures γ^t have the same total mass $\gamma^t(\Omega_0 \times \Omega_1) = \mu_0(\Omega_0)$ (see (12)) to estimate the second term.

Thus we see that the modulus of continuity σ is given by

$$\sigma(\delta) = \inf_{\tilde{\varphi} \in C^1} \left\{ \|\mathcal{A}\|^2 \|\Phi_x\|_\infty \|\tilde{\varphi}_x\|_\infty \delta + 2\mu_0(\Omega_0) \|\varphi - \tilde{\varphi}\|_\infty \right\}.$$

Clearly $\sigma(\delta)$ is monotone in $\delta > 0$, and $\lim_{\delta \to 0} \sigma(\delta) = 0$, which makes σ a modulus of continuity. ∎

We note that the modulus of continuity $\sigma(\delta)$ is actually bounded by

$$\sigma(\delta) \le C \sup_{|x - x'| + |y - y'| < \delta} |\varphi(x,y) - \varphi(x',y')|,$$

where C depends only on $\|A\|_{L(\mathfrak{H})}$ and $\|\Phi_x\|_\infty$.

6. Weak compactness. In this section, we study limits of sequences of weak solutions.

LEMMA 6.1. *Let $\mathcal{A}_n$ be a sequence of operators satisfying (21) and let $\Phi_n \in C^1(\bar{\Omega}_0 \times \bar{\Omega}_1)$ be a sequence of cost functions. Assume the $\mathcal{A}_n$ are uniformly bounded in $L(\mathfrak{H})$ and the Φ_n are uniformly bounded in $C^1(\bar{\Omega}_0 \times \bar{\Omega}_1)$.*

Given a family of weak solutions $\{\gamma_n^t, t \in [0,T)\}$*, to* (24) *with* $\mathcal{A} = \mathcal{A}_n$*,* $\Phi = \Phi_n$*, there is a subsequence such that* $\gamma_{n_k}^t \rightharpoonup \gamma_\infty^t$ *for some weak* continuous family of measures* $\{\gamma_\infty^t, t \in [0,T)\}$*.*

Proof. Given any $t \in [0,T)$ weak* compactness of $\mathfrak{X}$ enables us to find a subsequence of $\{\gamma_n^t\}$ which weak* converges. Given a finite subset $\{t_1, \dots, t_m\} \subset [0,T)$ we can repeat this argument m times and obtain a subsequence $n_k \in \mathbb{N}$ such that $\mu_{n_k}^t$ weak* converges for $t = t_1, t_2, \dots, t_m$. A diagonalization trick then gives us a further subsequence $n_k \in \mathbb{N}$ such that $\mu_{n_k}^t$ weak* converges for all rational $t \in [0,T)$.

We now argue that this subsequence $\gamma_{n_k}^t$ actually weak* converges for *all* $t \in [0,T)$ rather than just for all rational t.

Let $\varphi \in C^0(\bar{\Omega}_0 \times \bar{\Omega}_1)$ and some $t \in [0,T)$ be given. By Lemma 5.4 the functions $t \mapsto \langle \gamma_{n_k}^t, \varphi \rangle$ are equicontinuous. By Ascoli–Arzelà they form a precompact subset of $C^0([0,T))$. Since they converge pointwise on a dense subset of the interval $[0,T)$ they must converge uniformly for $0 \le t < T$.

It follows that $\langle \gamma_{n_k}^t, \varphi \rangle$ converges for all $t \in [0,T)$, as claimed.

To complete the proof we check weak* continuity in time of the limit measures γ^t. But this is immediate since the μ_n^t all share the same modulus of continuity from Lemma 5.4. Passing to the limit we find that γ^t also has this modulus of continuity. ∎

PROPOSITION 6.2. *Let* γ_n^t*,* $\mathcal{A}_n$*, and* Φ_n *be as in the previous lemma.*

Assume that the operators $\mathcal{A}_n$ *converge strongly to some operator* $\mathcal{A}_\infty$*. Assume also that the cost functions* Φ_n *converge in* C^1 *to some* $\Phi_\infty \in C^1$*.*

Then $W_n^t \stackrel{\text{def}}{=} \mathbb{E}_{\gamma_n^t}(\Phi_{n,x} \mid x)$ *weak* converges in* $L^\infty(\Omega_0; \mathbb{R}^d) = L^1(\Omega_0; \mathbb{R}^d)^*$ *to* $W_\infty^t \stackrel{\text{def}}{=} \mathbb{E}_{\gamma_\infty^t}(\Phi_{\infty,x} \mid x)$*.*

The limiting family γ_∞^t *satisfies the energy inequality*

$$M(\gamma^{t_1}) + \int_{t_0}^{t_1} \left\| \mathcal{A}_\infty \mathcal{P} W_\infty^t \right\|_{L^2}^2 \, \mathrm{d}t \le M(\gamma^{t_0}) \tag{37}$$

for all $0 \le t_0 < t_1 < T$*.*

Proof. For any $\zeta \in L^1(\Omega_0)$ we have

$$\begin{aligned} \int W_n^t(x) \cdot \zeta(x) \mathrm{d}\mu_0(x) &= \iint_{\Omega_0 \times \Omega_1} \zeta(x) \cdot \frac{\partial \Phi_n(x,y)}{\partial x} \mathrm{d}\mu_n^t(x,y) \\ &\to \iint_{\Omega_0 \times \Omega_1} \zeta(x) \cdot \frac{\partial \Phi_\infty(x,y)}{\partial x} \mathrm{d}\mu_\infty^t(x,y) \qquad \text{as } n \to \infty \\ &= \int W_\infty^t(x) \cdot \zeta(x) \mathrm{d}\mu_0(x), \end{aligned}$$

which establishes the weak* convergence of the W_n^t.

Since $M(\gamma_n^t) = \langle \gamma_n^t, \Phi_n \rangle$ weak* convergence of the measures γ_n^t directly implies convergence of the corresponding costs,

$$\lim_{n \to \infty} M(\gamma_n^t) = M(\gamma_\infty^t),$$

for all $t \in [0,T)$.

To prove the energy inequality (37) we need the following.

LEMMA 6.3. $\mathcal{A}_n \mathcal{P} W_n^t$ *converges weakly to* $\mathcal{A}_\infty \mathcal{P} W_\infty^t$ *in* $\mathfrak{H}$*, and hence*

$$\| \mathcal{A}_\infty \mathcal{P} W_\infty^t \| \le \liminf_{n \to \infty} \| \mathcal{A}_n \mathcal{P} W_n^t \|.$$

Given this lemma, we use Fatou's lemma to get

$$\int_{t_0}^{t_1} \left\|\mathcal{A}\mathcal{P}W_\infty^t\right\|_{L^2}^2 \,\mathrm{d}t \le \liminf_{k\to\infty} \int_{t_0}^{t_1} \left\|\mathcal{A}\mathcal{P}W_k^t\right\|_{L^2}^2 \,\mathrm{d}t.$$

The energy identity in Lemma 5.1 then directly leads to the energy inequality (37).

It therefore remains only to verify Lemma 6.3. To this end we recall that the W_n^t converge in the weak* topology on $L^\infty(\Omega_0)\otimes\mathbb{R}^d$ and hence converge weakly in $L^2(\Omega_0)\otimes\mathbb{R}^d = \mathfrak{H}$.

The Helmholtz projection $\mathcal{P}$ is bounded on $\mathfrak{H}$, so $\mathcal{P}W_n^t$ converges weakly to $\mathcal{P}W_\infty^t$.

The operators $\mathcal{A}_n$ converge strongly to $\mathcal{A}_\infty$, so for an arbitrary $f\in\mathfrak{H}$ we have $\|\mathcal{A}_n f - \mathcal{A}_\infty f\|_{\mathfrak{H}} \to 0$.

Altogether this gives us

$$(f, \mathcal{A}_n\mathcal{P}W_n^t)_{\mathfrak{H}} = (\mathcal{A}_n f, \mathcal{P}W_n^t)_{\mathfrak{H}} \to (\mathcal{A}_\infty f, \mathcal{P}W_\infty^t)_{\mathfrak{H}} = (f, \mathcal{A}_\infty\mathcal{P}W_\infty^t)_{\mathfrak{H}}$$

as $n\to\infty$ and for arbitrary $f\in\mathfrak{H}$; i.e., we find that $\mathcal{A}_n\mathcal{P}W_n^t$ converges weakly to $\mathcal{A}_\infty\mathcal{P}W_\infty^t$, and we are done. ∎

At this point it is not clear whether the limiting family $\{\gamma_\infty^t, t\in[0,T)\}$ is a weak solution.

LEMMA 6.4. *Assume that the integral kernel $K(x,\xi)$ of the operator $\mathcal{P}\mathcal{A}^2\mathcal{P}$ is a continuous function on $\bar\Omega_0\times\bar\Omega_1$.*

Then any weak* limit $\{\gamma_\infty^t\}$ of weak solutions $\{\gamma_k^t\}$ is again a weak solution.

Proof. Since the γ_k^t weak* converge to γ_∞^t, the product measures $\gamma_k^t\times\gamma_k^t$ also weak* converge[2] to $\gamma_\infty^t\times\gamma_\infty^t$.

The measures γ_n^t are uniformly bounded (their total mass is fixed by (12)), so, using the dominated convergence theorem, one easily shows that the Borel measures $\mathrm{d}\gamma_n^t\times\mathrm{d}\gamma_n^t\times\mathrm{d}t$ on $\bar\Omega_0\times\bar\Omega_0\times[0,T)$ converge weakly to $\mathrm{d}\gamma_\infty^t\times\mathrm{d}\gamma_\infty^t\times\mathrm{d}t$.

To prove that γ_∞^t is a weak solution to (24) we must show that γ_∞^t satisfies (35) for all test functions φ. Using the integral kernel $K(x,\xi)$ we rewrite (35) as

$$\int_{t_0}^{t_1}\int_{\Omega_0} \mathbb{E}_{\gamma^t}(\varphi_x \mid x)\cdot\mathcal{P}\mathcal{A}^2\mathcal{P}\{\mathbb{E}_{\gamma^t}(\Phi_x\mid x)\}\,\mathrm{d}x\,\mathrm{d}t = \left\langle\gamma^{t_0},\varphi\right\rangle - \left\langle\gamma^{t_1},\varphi\right\rangle,$$

i.e.,

$$\int_{t_0}^{t_1}\iint_{\Omega_0\times\Omega_0} \mathbb{E}_{\gamma^t}(\varphi_x\mid x)\cdot K(x,\xi)\cdot\mathbb{E}_{\gamma^t}(\Phi_x\mid\xi)\,\mathrm{d}x\mathrm{d}\xi\,\mathrm{d}t = \left\langle\gamma^{t_0},\varphi\right\rangle - \left\langle\gamma^{t_1},\varphi\right\rangle.$$

Using the definition of $\mathbb{E}_{\gamma^t}(\cdot\mid\cdot)$ we see that (35) is equivalent to

$$\int_{t_0}^{t_1}\iint_{(\Omega_0\times\Omega_1)^2} \frac{\varphi_x(x,y)\cdot K(x,\xi)\cdot\Phi_x(\xi,\eta)}{\mu_0(x)\mu_0(\xi)}\,\mathrm{d}\gamma^t(x,y)\mathrm{d}\gamma^t(\xi,\eta)\,\mathrm{d}t = \left\langle\gamma^{t_0},\varphi\right\rangle - \left\langle\gamma^{t_1},\varphi\right\rangle. \tag{38}$$

[2]If Borel measures μ_n and ν_n on compact Hausdorff spaces X and Y, respectively, converge weakly to measures γ and ν, then the product measures $\mu_n\times\nu_n$ converge weakly to $\gamma\times\nu$. Indeed, the $\mu_n\times\nu_n$ are uniformly bounded so one only has to check $\langle\mu_n\times\nu_n, f\rangle\to\langle\gamma\times\nu, f\rangle$ for a dense set of $f\in C^0(X\times Y)$. By the Stone–Weierstraß theorem we may therefore assume that $f(x,y) = g_1(x)h_1(y)+\cdots+g_k(x)h_k(y)$ for continuous functions g_i and h_i. By linearity we may assume that $k=1$. But if $f(x,y)=g_1(x)h_1(y)$, then $\langle\mu_n\times\nu_n,f\rangle = \langle\mu_n,g_1\rangle\langle\nu_n,h_1\rangle$, which converges to $\langle\gamma,g_1\rangle\langle\nu,h_1\rangle = \langle\gamma\times\nu,f\rangle$.

All measures γ_n^t satisfy (38) since they are weak solutions. Our hypotheses are such that the integrand in the triple integral in (38) is a continuous function for any choice of the test function φ. Weak* convergence of the measures $\mathrm{d}\gamma_n^t \times \mathrm{d}\gamma_n^t \times \mathrm{d}t$ then allows us to complete the proof by passing to the limit in (35). □

7. The Helmholtz decomposition. Any vector field $w : \Omega_0 \to \mathbb{R}^d$ can be decomposed into a divergence-free part and a gradient; i.e., one can find a vector field $\mathcal{P}w$ and a function $p = p_w$ such that

$$w = \mathcal{P}w + \nabla p, \qquad \operatorname{div} \mathcal{P}w = 0 \tag{39}$$

holds. We call $\mathcal{P}w$ the *Helmholtz projection* of w, and by analogy with fluid dynamics we will call p_w the corresponding pressure. The pressure p_w is determined by (39) up to an additive constant, at best. We can remove this freedom by imposing some normalization on p_w, such as

$$\int_{\Omega_0} p_w(x)\, \mathrm{d}x = 0.$$

To uniquely specify $\mathcal{P}w$ and p_w we must impose boundary conditions: we will always require $\mathcal{P}w$ to be tangential to the boundary. Thus if ν denotes the outward unit normal to $\partial\Omega_0$, then we require

$$\nu \cdot \nabla(\mathcal{P}w) = 0, \quad \text{or, equivalently,} \quad \nu \cdot \nabla p_w = \nu \cdot w \quad \text{on } \partial\Omega_0. \tag{40}$$

A brief construction of $\mathcal{P}w$ uses Hilbert space theory. Indeed, let $\mathfrak{H}_{\mathrm{div}}$ be the closed subspace of $\mathfrak{H} = L^2(\Omega; \mathbb{R}^d)$ determined by

$$\mathfrak{H}_{\mathrm{div}} = \{w \in \mathfrak{H} \mid (w, \nabla\varphi)_{\mathfrak{H}} = 0 \text{ for all } \varphi \in C^1(\bar{\Omega})\}.$$

Then the Helmholtz projection $\mathcal{P}$ is simply the orthogonal projection of $\mathfrak{H}$ onto $\mathfrak{H}_{\mathrm{div}}$. This implies the following.

LEMMA 7.1. *The operator $\mathcal{P}$ is bounded on $\mathfrak{H}$, with $\|\mathcal{P}\|_{L(\mathfrak{H})} = 1$.*

This construction does not show how $\mathcal{P}$ preserves smoothness of the vector field w. Therefore we now recall a different description of the Helmholtz decomposition.

7.1. Smooth domains. The defining equations (39) and (40) imply

$$\begin{cases} \Delta p_w = \operatorname{div} w \ \text{ on } \Omega_0, \\ \nu \cdot \nabla p_w = \nu \cdot w \text{ on } \partial\Omega_0. \end{cases} \tag{41}$$

Standard elliptic theory tells us that under minimal smoothness assumptions on $\partial\Omega_0$ and w a solution in the weak sense exists for this boundary value problem. Moreover, the vector field $\mathcal{P}w$ defined by

$$\mathcal{P}w := w - \nabla p_w$$

is divergence free and tangential to the boundary.

LEMMA 7.2. *Assume the boundary $\partial\Omega_0$ is $C^{1,\alpha}$ smooth. Then the Helmholtz projection is a bounded operator on $C^{1,\alpha}(\Omega_0; \mathbb{R}^d)$; i.e., for any vector field $w \in C^{1,\alpha}(\Omega_0; \mathbb{R}^d)$ one has $\mathcal{P}w \in C^{1,\alpha}(\Omega_0; \mathbb{R}^d)$ and $\|\mathcal{P}w\|_{C^{1,\alpha}} \le C\|w\|_{C^{1,\alpha}}$.*

Proof. If $w \in C^{1,\alpha}$, then $w \in C^{0,\alpha}$, while $\nu \cdot w \in C^{1,\alpha}$. Furthermore the data satisfy

$$\int_{\partial\Omega_0} \nu \cdot w = \int_{\Omega_0} \operatorname{div} w,$$

so the boundary value problem (41) has a unique solution $p_w \in C^{2,\alpha}(\Omega_0)$ with $\int_{\Omega_0} p_w \, \mathrm{d}x = 0$ (see [10]). One has

$$\begin{aligned}\|p_w\|_{C^{2,\alpha}} &\leq C_{\alpha,\Omega}\left\{\|\nu \cdot \nabla p_w\|_{C^{1,\alpha}} + \|\Delta p_w\|_{C^\alpha}\right\} \\ &= C_{\alpha,\Omega}\left\{\|\nu \cdot w\|_{C^{1,\alpha}(\partial\Omega_0)} + \|\operatorname{div} w\|_{C^\alpha(\Omega_0)}\right\}\end{aligned}$$

for some constant C_{α,Ω_0}. The representation $\mathcal{P}w = w - \nabla p_w$ then implies the lemma. ∎

7.2. Helmholtz decomposition on rectangles. In the case that Ω_0 is a rectangle, i.e., $\Omega_0 = [0, L_1] \times \cdots \times [0, L_d]$, we can give a more explicit representation of the Helmholtz projection by using Fourier series.

Assume for simplicity of notation that all sides of Ω_0 have length π, i.e., $L_j = \pi$. Then one can write any $w \in \mathfrak{H}$ as a series,

$$w(x) = \sum_{j=1}^{d} \sum_{\ell_1,\ldots,\ell_d \geq 0} \hat{w}_{j,\ell_1,\ldots,\ell_d} \cos(\ell_1 x_1) \cdots \sin(\ell_j x_j) \cdots \cos(\ell_d x_d) \mathbf{e}_j. \tag{42}$$

Observe that due to the presence of the factor $\sin \ell_j x_j$ the term with $\ell = 0$, i.e., with $\ell_1 = \cdots = \ell_d = 0$, is absent from the sum. We will not try to incorporate this fact into our notation, but it will allow us to divide by $|\ell|$ in what follows.

The $L^2(\Omega_0; \mathbb{R}^d) = \mathfrak{H}$-norm of such a vector field is

$$\|w\|^2_{\mathfrak{H}} = \left(\frac{\pi}{2}\right)^d \sum_{j=1}^{d} \sum_{\ell_1,\ldots,\ell_d \geq 0} 2^{\mathsf{C}_\ell} \left|\hat{w}_{j,\ell_1,\ldots,\ell_d}\right|^2, \tag{43}$$

where C_ℓ denotes the number of components of $\ell = (\ell_1, \ldots, \ell_d)$ which vanish.

Any L^2 vector field given by (42) extends to a vector field on all of $\mathbb{R}^d$ which is 2π periodic in each of the variables $x_1, \ldots, x_d$. Moreover, any w given by (42) has the symmetry

$$w(R_j x) = R_j\big(w(x)\big) \quad \text{for } j = 1, \ldots, d, \tag{44}$$

in which R_j is the reflection $R_j(x_1, \ldots, x_d) = (x_1, \ldots, x_{j-1}, -x_j, x_{j+1}, \ldots, x_d)$. See Figure 2. Conversely, any vector field $w \in L^2\big([-\pi, \pi]^d; \mathbb{R}^d\big)$ which has the symmetries (44) can be written as a Fourier series of the form (42).

The Helmholtz projection of w is then given by

$$\mathcal{P}w(x) = \sum_{j=1}^{d} \sum_{\ell_1,\ldots,\ell_d \geq 0} \widehat{(\mathcal{P}w)}_{j,\ell_1,\ldots,\ell_d} \cos(\ell_1 x_1) \cdots \sin(\ell_j x_j) \cdots \cos(\ell_d x_d) \mathbf{e}_j, \tag{45}$$

with

$$\widehat{(\mathcal{P}w)}_{j,\ell} = \sum_{k=1}^{d} \left(\delta_{jk} - \frac{\ell_j \ell_k}{|\ell|^2}\right) \hat{w}_{j,\ell}$$

and where $|\ell|^2 = \ell_1^2 + \cdots + \ell_d^2$. The corresponding "pressure" is given by

$$p_w = \sum_{|\ell|>0} \widehat{(p_w)}_\ell \cos(\ell_1 x_1) \cdots \cos(\ell_d x_d),$$

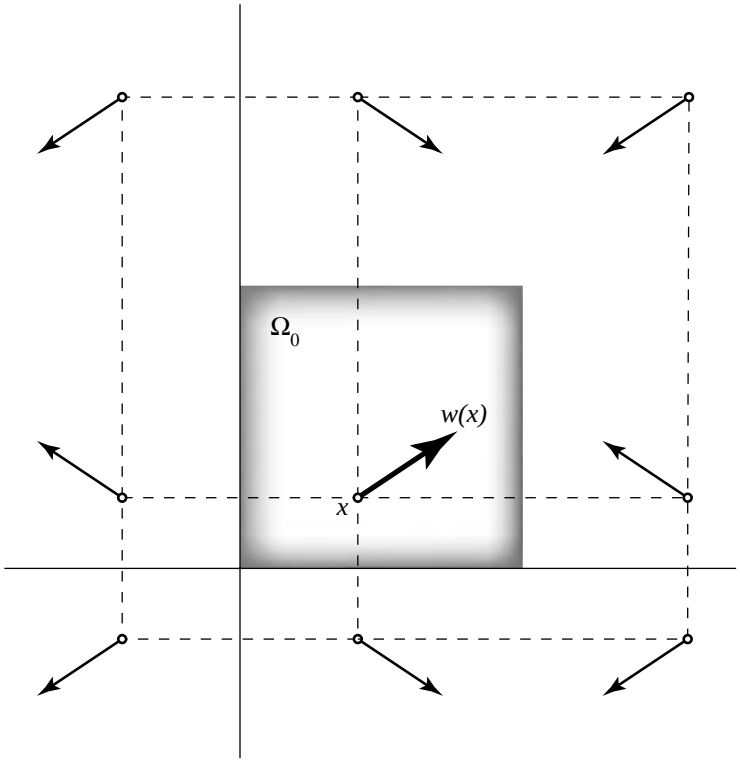


FIG. 2. *The symmetry* (44).

with

$$\widehat{(p_w)}_\ell = \frac{\ell_1 \hat{w}_{1,\ell} + \cdots + \ell_d \hat{w}_{d,\ell}}{|\ell|^2}.$$

Let $C^{1,\alpha}(\mathbb{T}^d;\mathbb{R}^d)$ be the space of all $C^{1,\alpha}$ vector fields on $\mathbb{R}^d$ which are 2π periodic in all variables $x_1, \ldots, x_d$, and let $\mathfrak{C}^{1,\alpha}$ be the closed subspace of $C^{1,\alpha}(\mathbb{T}^d;\mathbb{R}^d)$ consisting of all vector fields which have the symmetry (44).

LEMMA 7.3. *The Helmholtz projection is a bounded operator on* $C^{1,\alpha}(\mathbb{T}^d;\mathbb{R}^d)$ *which leaves the subspace* $\mathfrak{C}^{1,\alpha}$ *invariant.*

Proof. We can write $\mathcal{P}$ on $C^{1,\alpha}(\mathbb{T}^d;\mathbb{R}^d)$ as $\mathcal{P} = \mathrm{I} - \mathrm{grad} \circ (\Delta)^{-1} \circ \mathrm{div}$, where for any f with $\int_{[-\pi,\pi]^d} f \mathrm{d}x = 0$ we define $u = (\Delta)^{-1} f$ to be the unique solution of $\Delta u = f$ with $\int_{[-\pi,\pi]^d} u \mathrm{d}x = 0$. Classical Schauder estimates imply that $(\Delta)^{-1}$ is bounded from $C^{0,\alpha}$ to $C^{2,\alpha}$. This implies that $\mathcal{P} = \mathrm{I} - \mathrm{grad} \circ (\Delta)^{-1} \circ \mathrm{div}$ is bounded on $C^{1,\alpha}$ vector fields.

The Helmholtz decomposition is easily seen to commute with the symmetries (44), so that $\mathfrak{C}^{1,\alpha}$ is an invariant subspace. ∎

8. The smoothing operator $\mathcal{A}$. In this section, we exhibit smoothing operators $\mathcal{A}$ which satisfy all assumptions made so far.

8.1. Smoothing vector fields on $C^{1,\alpha}$ domains. Many different smoothing operators can be constructed. The following is one choice.

LEMMA 8.1. *Let Ω_0 be a domain with $C^{1,\alpha}$ boundary, and let $\mathcal{A}_\varepsilon$ be the operator*

$$\mathcal{A}_\varepsilon = e^{\varepsilon\Delta_N},\ \textit{i.e.,}\ \mathcal{A}_\varepsilon w = \big(e^{\varepsilon\Delta_N} w_1, \dots, e^{\varepsilon\Delta_N} w_d\big),$$

where Δ_N is the Neumann–Laplacian on Ω_0.

Then $\mathcal{A}_\varepsilon$ is a bounded, injective, self-adjoint operator on $\mathfrak{H}$, with $\|A_\varepsilon\|_{L(\mathfrak{H})} \le 1$ for any $\varepsilon > 0$.

The operator $\mathcal{A}_\varepsilon$ is also bounded from $C^{1,\alpha}$ to $C^{1,\alpha}$, with $\|\mathcal{A}_\varepsilon\|_{L(C^{1,\alpha})} \le C$ for some C that does not depend on $\varepsilon > 0$.

Proof. The operator $\mathcal{A}_\varepsilon$ acts on each individual component w_i of a vector field w in the same way. The fact that $e^{\varepsilon\Delta_N}$ is a contraction of L^2 and uniformly bounded on $C^{1,\alpha}$ respectively follows from linear parabolic theory.

The Neumann Laplacian is well known to be a self-adjoint operator on $L^2(\Omega_0)$, so that $\mathcal{A}_\varepsilon$ is self-adjoint. Self-adjointness of Δ_N implies via the spectral theorem for self-adjoint operators that $e^{\varepsilon\Delta_N}$ is injective for all $\varepsilon > 0$. □

LEMMA 8.2. *Let $\mathcal{A}$ be as above.*

The operator $\mathcal{P}\mathcal{A}^2\mathcal{P}$ is bounded from $\mathfrak{H}$ to $C^{1,\alpha}(\Omega;\mathbb{R}^d)$ for any $0 < \alpha < 1$.

The operator $\mathcal{P}\mathcal{A}^2\mathcal{P}$ has an integral kernel $K \in C^{1,\alpha}(\bar\Omega_0 \times \bar\Omega_0)$.

Proof. Boundedness of $\mathcal{A} : \mathfrak{H} \to C^{1,\alpha}$ follows from the smoothing property of the heat equation. (But the operator norm $\|\mathcal{A}\|_{L(\mathfrak{H},C^{1,\alpha})}$ blows up as $\varepsilon \searrow 0$.) Since $\mathcal{P}$ is bounded on both $\mathfrak{H}$ and $C^{1,\alpha}$ it follows that $\mathcal{P}\mathcal{A}^2\mathcal{P}$ is bounded from $\mathfrak{H}$ to $C^{1,\alpha}$.

To study the kernel of the operator $\mathcal{P}\mathcal{A}^2\mathcal{P}$ we write $\mathcal{P}\mathcal{A}^2\mathcal{P}$ as $T \circ T^*$, where $T = \mathcal{P}\mathcal{A}$, and show that T has an integral kernel.

Let $\Gamma_\varepsilon^y(x)$ be the heat kernel for Δ_N; i.e., for any function $\phi \in L^1(\Omega_0)$ one has

$$e^{\varepsilon\Delta_N} f(x) = \int_{\Omega_0} \phi(y)\Gamma_\varepsilon^y(x)\,\mathrm{d}y. \tag{46}$$

Then $(x,y) \mapsto \Gamma^y(x)$ is a $C^{1,\alpha}$ function.

We expand the vector-valued function f into its components, $f(x) = f_1(x)\mathbf{e}_1 + \cdots + f_d(x)\mathbf{e}_d$, $f_1, \dots, f_d$ being scalar L^2 functions. From (46) we then get the following representation for $\mathcal{P}\mathcal{A}f$:

$$\mathcal{P}\mathcal{A}f = \sum_{j=1}^{d} \int_{\Omega_0} f_j(y)\mathcal{P}(\Gamma_\varepsilon^y \otimes \mathbf{e}_j)\,\mathrm{d}y.$$

Let $N_j^y(x)$ be the function $N_j^y = \mathcal{P}(\Gamma_\varepsilon^y \otimes \mathbf{e}_j)$. We get

$$\mathcal{P}\mathcal{A}f(x) = \int_{\Omega_0} \sum_{j=1}^{d} N_j^y(x) f_j(y)\,\mathrm{d}y,$$

which means that $\mathcal{P}\mathcal{A}$ is an integral operator with matrix-valued kernel

$$\mathsf{N}(x,y) = \big[N_1^y(x), \dots, N_d^y(x)\big];$$

i.e., the jth column of the matrix $\mathsf{N}(x,y)$ is $N_j^y(x)$.

For each $y \in \Omega_0$ we have $\Gamma_\varepsilon^y \in C^{1,\alpha}(\bar\Omega_0)$, so by Lemma 7.2 we get $\mathcal{P}\Gamma_\varepsilon^y \in C^{1,\alpha}(\bar\Omega_0;\mathbb{R}^d)$. Moreover, $\Gamma_\varepsilon^y \in C^{1,\alpha}(\bar\Omega_0;\mathbb{R}^d)$ depends $C^{1,\alpha}$ smoothly on y, so we see that the kernel N is a $C^{1,\alpha}$ function on $\bar\Omega_0 \times \bar\Omega_0$.

The operator $\mathcal{P}\mathcal{A}^2\mathcal{P} = (\mathcal{P}\mathcal{A})(\mathcal{P}\mathcal{A})^*$ must now also be an integral operator, and its kernel must be

$$K(x,\xi) = \int_{\Omega_0} \mathsf{N}(x,y)\mathsf{N}(\xi,y)^T \, \mathrm{d}y.$$

This is clearly again a $C^{1,\alpha}$ function on $\bar{\Omega}_0 \times \bar{\Omega}_0$. ∎

The operators $\mathcal{A}_\varepsilon$ do not preserve the boundary condition $n \cdot w = 0$, so they will not commute with the Helmholtz projection $\mathcal{P}$.

8.2. Smoothing operator on a rectangle. In the case that Ω_0 is a rectangle, i.e., $\Omega_0 = [0,\pi]^d$, we can construct a different smoothing operator by using the Fourier series (42).

We define the smoothing operator $\mathcal{A}_\varepsilon$ by

$$\mathcal{A}_\varepsilon w = \big(e^{\varepsilon\Delta_1} w_1, \ldots, e^{\varepsilon\Delta_d} w_d\big),$$

in which Δ_j is the Laplacian with Neumann boundary conditions on the sides $x_j = 0$ and $x_j = \pi$, and Dirichlet boundary conditions on all other sides of the rectangle Ω_0.

An equivalent description of $\mathcal{A}_\varepsilon$ goes like this: to compute $\mathcal{A}_\varepsilon w$ for some vector field w on $\Omega_0 = [0,\pi]^d$ extend w to a vector field $\tilde{w}$ on all of $\mathbb{R}^d$ by imposing the symmetries (44) and by requiring the extension to be 2π periodic in all variables. We then set $\mathcal{A}_\varepsilon = e^{\varepsilon\Delta}\tilde{w}$, in which $e^{\varepsilon\Delta}$ is the standard heat semigroup on $\mathbb{R}^d$; i.e., we have

$$\mathcal{A}_\varepsilon w(x) = \big(4\pi\varepsilon\big)^{-d/2} \int_{\mathbb{R}^d} e^{-|x-\xi|^2/4\varepsilon} \tilde{w}(\xi) \, \mathrm{d}\xi. \tag{47}$$

If w is given by the Fourier series (42), then $\mathcal{A}_\varepsilon w$ is given by

$$\mathcal{A}_\varepsilon w(x) = \sum_{j=1}^{d} \sum_{\ell_1,\ldots,\ell_d \geq 0} \widehat{(\mathcal{A}_\varepsilon w)}_{j,\ell_1,\ldots,\ell_d} \cos(\ell_1 x_1) \cdots \sin(\ell_j x_j) \cdots \cos(\ell_d x_d) \mathbf{e}_j,$$

with

$$\widehat{(\mathcal{A}_\varepsilon w)}_{j,\ell} = e^{-\varepsilon|\ell|^2} \hat{w}_{j,\ell}. \tag{48}$$

LEMMA 8.3. *The smoothing operators $\mathcal{A}_\varepsilon$ are uniformly bounded on $\mathfrak{H}$ and $\mathfrak{C}^{1,\alpha}$. They are self-adjoint and injective, and they commute with the Helmholtz projection.*

Proof. The statements concerning the behavior of the operators on the Hilbert space $\mathfrak{H}$ follow directly from the series expansion (42), the Fourier multiplier descriptions (45) and (47) of $\mathcal{P}$ and $\mathcal{A}_\varepsilon$, respectively, and the Plancherel identity (43).

The $C^{1,\alpha}$ bounds follow from the representation (47). ∎

LEMMA 8.4. *Let $\mathcal{A} = \mathcal{A}_\varepsilon$ be as above.*

The operator $\mathcal{P}\mathcal{A}^2\mathcal{P}$ is bounded from $\mathfrak{H}$ to $\mathfrak{C}^{1,\alpha}(\Omega;\mathbb{R}^d)$ for any $0 < \alpha < 1$.

The operator $\mathcal{P}\mathcal{A}^2\mathcal{P}$ has an integral kernel $K \in C^{1,\alpha}(\bar{\Omega}_0 \times \bar{\Omega}_0)$.

The kernel $K(x,\xi)$ satisfies

$$K(R_j x, \xi) = R_j K(x,\xi), \qquad j = 1,\ldots,d. \tag{49}$$

Proof. This lemma is analogous to Lemma 8.2, and its proof proceeds along the same lines.

Boundedness from $\mathfrak{H}$ to $\mathfrak{C}^{1,\alpha}$ again follows from the smoothing property of the heat equation, i.e., of $e^{\varepsilon\Delta}$. The integral kernel is constructed in the same way, starting from the explicit representation (47) of $e^{\varepsilon\Delta}w$.

For any vector field $f \in \mathfrak{H}$ the smoothed-out projection $w = \mathcal{P}\mathcal{A}^2\mathcal{P}f$ belongs to $\mathfrak{C}^{1,\alpha}$. We therefore may conclude from $w(R_jx) \equiv R_jw(x)$ that

$$\int_{\Omega_0} K(R_jx,\xi)f(\xi)\,\mathrm{d}\xi = R_j\int_{\Omega_0} K(x,\xi)f(\xi)\,\mathrm{d}\xi$$

for all $x \in \mathbb{R}^d$, $j = 1, \ldots, d$, and all $f \in \mathfrak{H}$. This implies (49). □

9. Existence and well-posedness for the regularized flow. In this section, we construct classical solutions γ^t of the initial value problem (24) by writing them as $\gamma^t = (s^t \times \mathrm{id})_\#(\gamma^0)$ and solving the initial value problem (30)

$$\left.\begin{aligned} &\frac{\partial s^t}{\partial t} = -\int_{\Omega_0\times\Omega_1} \frac{K(s^t(x),s^t(\xi))}{\mu_0(s^t(x))\mu_0(s^t(\xi))}\cdot\Phi_x(s^t(\xi),\eta)\,\mathrm{d}\gamma^0(\xi,\eta), && (50)\\ &s^0(x) = x \quad (x\in\Omega) && (51)\end{aligned}\right\}$$

for s^t.

THEOREM 9.1. *Let the cost function Φ be C^1. Assume also that the smoothing operator $\mathcal{A}$ is such that the kernel K is $C^{1,\alpha}$. Then for any initial measure $\gamma^0 \in \mathfrak{X}$ the initial value problem* (50) *has a solution $\{s^t \in C^{1,\alpha}(\bar\Omega_0;\bar\Omega_0) : 0 \le t < \infty\}$.*

If the cost function Φ is C^2, then the solution $\{s^t : t \ge 0\}$ is unique.

We begin the proof by observing that (50) is of the form

$$\frac{\partial s^t}{\partial t} = \int_{\Omega_0\times\Omega_1} F(s^t(x),s^t(\xi);\xi,\eta)\,\mathrm{d}\gamma^0(\xi,\eta), \tag{52}$$

where the map $F : \Omega_0\times\Omega_0\times\Omega_0\times\Omega_1 \to \mathbb{R}^d$ is given by

$$F(s,\sigma;\xi,\eta) = \frac{K(s,\sigma)\cdot\Phi_x(\sigma,\eta)}{\mu_0(s)\mu_0(\sigma)}. \tag{53}$$

If $\Omega_0 = [0,\pi]^d$ is a rectangle, then $K(s,\sigma)$ is defined for all $s \in \mathbb{R}^d$. We agree to extend $\mu_0(x)$ to be 2π periodic and even in each variable so that $F(s,\sigma;\xi,\eta)$ is also defined for $s \in \mathbb{R}^d$.

LEMMA 9.2.

(A) *The map $(s,\sigma;\xi,\eta) \mapsto F(s,\sigma;\xi,\eta)$ is continuous.*

(B) *$F(s,\sigma;\xi,\eta)$ is also C^1 in $s \in \Omega_0$, and the partial derivative $\frac{\partial F}{\partial s}$ is uniformly bounded:*

$$\left|\frac{\partial F}{\partial s}(s,\sigma;\xi,\eta)\right| \le C, \tag{54}$$

with $C < \infty$ independent of s, σ, ξ, and η.

(C1) *If $\partial\Omega_0$ is $C^{1,\alpha}$ smooth, and if $s \in \partial\Omega_0$, then $s\cdot F(s,\sigma,\xi,\eta) = 0$ for all $\sigma,\xi \in \Omega_0$ and $\eta \in \Omega_1$.*

(C2) *If $\Omega_0 = [0,\pi]^d$, then $F(s,\sigma;\xi,\eta)$ is 2π periodic in each component of $s = (s_1,\ldots,s_d)$, and $s \mapsto F(s,\sigma;\xi,\eta)$ satisfies the symmetries* (44), *i.e.,*

$$F(R_js,\sigma;\xi,\eta) = R_jF(s,\sigma;\xi,\eta).$$

Proof. (A) and (B) are immediate from the representation (53), the known continuity and smoothness properties of the kernel K, and the cost function Φ.

(C1). The kernel $K(x,\xi)$ satisfies $x \cdot K(x,\xi) = 0$, since for any vector field $w \in L^2(\Omega;\mathbb{R}^d)$ the vector field $\mathcal{P}\mathcal{A}^2\mathcal{P}w(x) = \int_{\Omega_0} K(x,\xi)w(\xi)d\xi$ is everywhere tangent to $\partial\Omega$. This implies

$$x \cdot \mathcal{P}\mathcal{A}^2\mathcal{P}w(x) = \int_{\Omega_0} x \cdot K(x,\xi) \cdot w(\xi)d\xi = 0$$

for arbitrary w, which can happen only if $x \cdot K(x,\xi) \equiv 0$. Equation (53) then implies (C1).

The kernel $K(s,\sigma)$ and the density $\mu_0(s)$ are periodic and have the appropriate symmetries, so (C2) follows immediately from (53). □

9.1. Construction of a solution to (50). We regard the initial value problem (50) as a fixed point problem for the map $\mathcal{F} : \sigma \mapsto s$, where $s = \mathcal{F}(\sigma)$ is the solution of the ODE

$$\frac{\partial s^t(x)}{\partial t} = \int_{\Omega_0} F(s^t(x), \sigma^t(\xi), \xi, \eta)\ \mathrm{d}\gamma^0(\xi,\eta),$$

with initial data $s^0 = \mathrm{id}$, i.e., $s^0(x) \equiv x$.

To set up a fixed point argument (and, in particular, to use the Brouwer–Leray–Schauder fixed point theorem) we must overcome a technical difficulty, namely, the space of maps $\{s^t : \Omega_0 \to \Omega_0, 0 \le t \le T\}$ is not a linear space, since the target Ω_0 is not a vector space. To deal with this we extend the domain of the definition of the nonlinear map $F(s,\sigma;\xi,\eta)$ to include all $(s,\sigma,\xi,\eta) \in \mathbb{R}^d \times \mathbb{R}^d \times \Omega_0 \times \Omega_1$; in other words, we lift the restriction $s,\sigma \in \Omega_0$. Then we can regard maps $s^t : \Omega_0 \to \Omega_0$ as maps $s^t : \Omega_0 \to \mathbb{R}^d$, and the space of such maps (defined below as $\mathcal{C}_T$) is a vector space.

We only have to go through this extension process in the case where Ω_0 is a smoothly bounded domain. When Ω_0 is a rectangle the function $F(s,\sigma;\xi,\eta)$ is already defined for all $s,\sigma \in \mathbb{R}^d$.

9.1.1. Extending *F*. We choose a defining function $\varrho \in C^{1,\alpha}(\mathbb{R}^d)$ for Ω_0. This means that $\Omega_0 = \{x \in \mathbb{R}^d : \varrho(x) > 0\}$ and $\nabla\varrho \neq 0$ on $\partial\Omega_0$. We can choose ϱ so that $\nabla\varrho(x) \neq 0$ if $-1 \le \varrho(x) \le 1$, while $\varrho(x) = -2$ outside some compact set $K \supset \bar{\Omega}_0$.

Let $U = \{x \in \mathbb{R}^d : \varrho(x) \ge -1\}$, and choose a retraction $\pi : U \to \bar{\Omega}_0$. One possible choice is $\pi(x_0) = x_0$ for $x_0 \in \bar{\Omega}_0$, and

$$\pi(x_0) = \left\{ \begin{matrix} \text{the first point in } \bar{\Omega}_0 \text{ on the orbit } x(t) \text{ of} \\ \text{the gradient flow } \dot{x} = \nabla\varrho(x) \text{ which starts} \\ \text{at } x(0) = x_0 \in U \setminus \bar{\Omega}_0 \end{matrix} \right\}$$

for all $x_0 \in U \setminus \bar{\Omega}_0$. The retraction π is Lipschitz continuous on U and even C^1 on $U \setminus \partial\Omega_0$.

The extension F_* of F will now be defined in several stages. First we introduce a map $F_1 : U \times U \times \Omega_0 \times \Omega_1 \to \mathbb{R}^d$ given by

$$F_1(s,\sigma,\xi,\eta) = F(\pi(s),\pi(\sigma),\xi,\eta).$$

Next, let $\chi : \mathbb{R} \to \mathbb{R}$ be a Lipschitz cutoff function, e.g., $\chi(t) = 1$ for $t \ge 0$, $1 + t$ for $-1 \le t \le 0$, and 0 for $t \ge -1$. Put

$$F_2(s,\sigma,\xi,\eta) = \chi(\varrho(s))\chi(\varrho(\sigma))F_1(s,\sigma,\xi,\eta)$$

when $(s,\sigma) \in U \times U$, and $F_2(s,\sigma,\xi,\eta) = 0$ otherwise. Then $F_2 : \mathbb{R}^d \times \mathbb{R}^d \times \Omega_0 \times \Omega_1 \to \mathbb{R}^d$ is an extension of F which is uniformly Lipschitz continuous in $(s,\sigma) \in \mathbb{R}^d \times \mathbb{R}^d$ and continuous in $\xi \in \Omega_0$, $\eta \in \Omega_1$.

Finally, we define $F_* : \mathbb{R}^d \times \mathbb{R}^d \times \Omega_0 \times \Omega_1 \to \mathbb{R}^d$ by setting

$$F_*(s,\sigma,\xi,\eta) = \begin{cases} F_2 & \text{for } s \in \bar{\Omega}_0, \\ F_2 - \dfrac{\nabla\varrho(s)\cdot F_2}{|\nabla\varrho(s)|^2}\nabla\varrho(s) & \text{for } s \in U \setminus \bar{\Omega}_0, \\ 0 & \text{for } s \in \mathbb{R}^d \setminus U, \end{cases} \tag{55}$$

where $F_2 = F_2(s,\sigma,\xi,\eta)$.

LEMMA 9.3. *The extension $F_* : \mathbb{R}^d \times \mathbb{R}^d \times \Omega_0 \times \Omega_1 \to \mathbb{R}^d$ of F is continuous in (s,σ) and uniformly Lipschitz in $s \in \mathbb{R}^d$; i.e., for all $\sigma \in \mathbb{R}^d$ and $\xi \in \Omega_0$, $\eta \in \Omega_1$ one has*

$$|F_*(s,\sigma,\xi,\eta) - F_*(s',\sigma,\xi,\eta)| \le M|s-s'| \tag{56}$$

for some $M < \infty$. Furthermore, F_* is uniformly bounded,

$$|F_*(s,\sigma,\xi,\eta)| \le M' \tag{57}$$

for some constant $M' < \infty$ and for all $s,\sigma \in \mathbb{R}^d$ and $\xi \in \Omega_0$, $\eta \in \Omega_1$.

If the cost function Φ is C^2, then $F_*(s,\sigma,\xi,\eta)$ is uniformly Lipschitz in $(s,\sigma) \in \mathbb{R}^d$; i.e., for some finite M'' one has

$$|F_*(s,\sigma,\xi,\eta) - F_*(s',\sigma',\xi,\eta)| \le M''\{|s-s'| + |\sigma-\sigma'|\} \tag{58}$$

for all $s,s',\sigma,\sigma' \in \mathbb{R}^d$ and $\xi \in \Omega_0$, $\eta \in \Omega_1$.

Finally, F_* satisfies

$$\nabla\varrho(s)\cdot F_*(s,\sigma,\xi,\eta) = 0 \quad \textit{when} \quad \varrho(s) \le 0. \tag{59}$$

9.1.2. The fixed point argument. We prove existence and uniqueness of solutions for the case where $\partial\Omega_0$ is $C^{1,\alpha}$ smooth. The same arguments with minor modifications apply to the case $\Omega_0 = [0,\pi]^d$.

With the extended F in hand we can set up the fixed point problem. Let $\mathcal{C}_T$ be the Banach space

$$\mathcal{C}_T = C^0([0,T]\times\Omega_0;\mathbb{R}^d).$$

LEMMA 9.4 (definition of $\mathcal{F}$). *Let $\sigma \in \mathcal{C}_T$ be given. Define $s = \mathcal{F}(\sigma)$ to be the solution of*

$$\frac{\partial s^t}{\partial t} = \int_{\Omega_0} F_*(s^t(x),\sigma^t(\xi);\xi,\eta)\,\mathrm{d}\gamma^0(\xi,\eta), \qquad s^0(x) = x. \tag{60}$$

Then

$$\left|\frac{\partial s^t}{\partial t}\right| \le M'|\Omega_0| \tag{61}$$

and

$$|s^t(x) - s^t(x')| \le e^{M|\Omega_0|t}|x-x'|. \tag{62}$$

Proof. Equation (60) is an ODE for $s^t(x)$ of the form $\partial_t s^t = v^t(s^t)$, where

$$v^t(s) = \int_{\Omega_0} F_*(s, \sigma^t(\xi); \xi, \eta)\, \mathrm{d}\gamma^0(\xi, \eta).$$

The estimates (56) and (57) for F_* imply that $|v^t(s)| \le M'|\Omega_0|$ and $|v^t(s) - v^t(s')| \le M|\Omega_0||s - s'|$. Standard theorems for ODEs then imply (61), (62). ∎

LEMMA 9.5. *Let s^t, $0 \le t \le T$, be the solution of (60) for some $\sigma \in \mathcal{C}_T$. Then the s^t are C^1 diffeomorphisms of $\bar\Omega_0$.*

Proof. The s^t are the flow of a vector field v^t, so we only have to show that $s^t(\bar\Omega_0) = \bar\Omega_0$. But our construction of F_* is such that $v^t(s) \cdot \nabla\varrho(s) = 0$ whenever $\varrho(s) \le 0$. Indeed, one has

$$v^t(s) \cdot \nabla\varrho(s) = \int_{\Omega_0} \nabla\varrho(s) \cdot F_*(s, \sigma^t(\xi); \xi, \eta)\, \mathrm{d}\gamma^0(\xi, \eta) = 0$$

by (59). Therefore ϱ is a conserved quantity outside of Ω_0, and in particular $\partial\Omega_0$ is invariant under the flow of v^t. So $s^t(\bar\Omega_0) = \bar\Omega_0$.

Since the vector field v^t is C^1 on $\bar\Omega$ the flow s^t is also C^1. ∎

Existence. The estimates (61) and (62) imply that $\mathcal{F}$ maps all of $\mathcal{C}_T$ into a compact subset of $\mathcal{C}_T$. Hence the Brouwer–Leray–Schauder fixed point theorem applies, and we can conclude the existence of a fixed point $s_T \in \mathcal{C}_T$ for $\mathcal{F}$. The initial value problem for the rearrangement map s^t therefore has a solution on any finite time interval $0 \le t \le T$. Since we have not established uniqueness of the solution, the solutions s_T might actually depend on T. However, they all satisfy the a priori estimates (61), (62) so, as $T \nearrow \infty$, one can extract a subsequence which converges uniformly on any finite time interval. The limit of this subsequence is then a global solution $\{s^t\}_{t\ge0}$.

Uniqueness. If the cost function Φ is C^2, then there is only one solution. To see this let $s, \bar s \in \mathcal{C}_T$ be any two solutions and consider their difference $w^t(x) = s^t(x) - \bar s^t(x)$.

Both s and $\bar s$ are solutions to (52), so subtracting the two equations we get

$$|\partial_t w^t(x)| \le M''|\Omega_0| \sup_{\xi\in\Omega_0} |w^t(\xi)|,$$

where we have used that $F_*(s, \sigma, \xi, \eta)$ is uniformly Lipschitz in $(s, \sigma) \in \mathbb{R}^d \times \mathbb{R}^d$ (by (58)).

This implies that $\sup|w^t| \le e^{M''|\Omega_0|t} \sup|w^0|$. Since $w^0 = s^0 - \bar s^0 = 0$ we find that $w^t \equiv 0$.

9.2. The regularized flow on $\mathfrak{X}$. If the cost function Φ is C^2, then there is another way of proving existence and uniqueness of solutions to (50). Namely, we observe that (50) is an ODE on a Banach space. One can write (50) as

$$\frac{\partial s^t}{\partial t} = \mathrm{V}(s),$$

where V is given by

$$\mathrm{V}(s)(x) = \iint_{\Omega_0\times\Omega_1} F_*(s^t(x), s^t(\xi); \xi, \eta)\, \mathrm{d}\gamma^0(\xi, \eta).$$

Here F_* is the extension of F constructed in section 9.1.1.

The properties of F_* derived in Lemma 9.3 imply that V is a globally Lipschitz vector field on the Banach space $\mathfrak{Z} = C^0(\bar{\Omega}_0; \mathbb{R}^d)$. It follows immediately that $\partial_t s^t = \mathrm{V}(s^t)$ generates a global flow on $\mathfrak{Z}$, since $t \mapsto s^t(x)$ is a solution of the ODE

$$\frac{ds}{dt} = v^t(s),$$

where

$$v^t(\sigma) = \iint_{\Omega_0 \times \Omega_1} F_*(\sigma, s^t(\xi); \xi, \eta)\, \mathrm{d}\gamma^0(\xi, \eta).$$

Thus $s^t = S^t \circ s^0$, where S^t is the flow of the vector field v^t.

9.3. ω-limit sets of the regularized flow. Let $\gamma^0 \in \mathfrak{X}$ be any initial measure, and let $\{\gamma^t, t \geq 0\}$ be a solution of (24) starting at γ^0. In dynamical systems one defines the ω-limit set of the solution $\{\gamma^t\}$ to be

$$\omega(\{\gamma^t\}) = \big\{\lambda \in \mathfrak{X} \mid \exists t_k \nearrow \infty : \gamma^{t_k} \rightharpoonup \lambda\big\} = \bigcap_{s \geq 0} \overline{\{\gamma^t \mid t \geq s\}}.$$

The second description shows that $\omega(\{\gamma^t\})$ is a closed (hence compact) and connected subset of $\mathfrak{X}$.

PROPOSITION 9.6. *$\omega(\{\gamma^t\})$ consists of critical points for* (24).

Proof. For given $\lambda \in \omega(\{\gamma^t\})$ we choose a sequence $t_k \nearrow \infty$ with $\gamma^{t_k} \rightharpoonup \lambda$ and consider the weak solutions $\lambda_k^t = \gamma^{t_k + t}$. By Lemma 6.4 we can find a subsequence t_{k_j} for which the λ_k^t weak* converge to a new weak solution $\lambda_\dagger^t$. The $\lambda_\dagger^t$, being weak solutions, satisfy the energy identity from Lemma 5.1. Furthermore,

$$M(\lambda_\dagger^t) = \lim_{j \to \infty} M(\lambda_{k_j}^t) = \lim_{j \to \infty} M(\gamma^{t_{k_j} + t}) = \lim_{t \to \infty} M(\gamma^t),$$

where the latter limit must exist since $M(\gamma^t)$ is a nonincreasing and bounded quantity.

The energy identity for $\lambda_\dagger^t$ together with constancy of $M(\lambda_\dagger^t)$ imply that the $\lambda_\dagger^t$ are critical points. In particular, $\lambda_\dagger^0 = \lim_{j \to \infty} \gamma^{t_{k_j}} = \lambda$ must be a critical point, as claimed. ∎

10. The unregularized flow. In the unregularized case, where one takes $\mathcal{A} = I_{\mathfrak{H}}$, one can try to construct weak solutions of (35) by solving the equation for a sequence of smoothing operators $\mathcal{A}$ which approximate the identity, and extract a weak limit of the solutions of the regularized equations. In this section, we study the limits of weak solutions which arise in this way. Although we do not show they are weak solutions, these limits still have many of the properties of weak solutions.

10.1. Choice of $\mathcal{A}_\varepsilon$. We let $\mathcal{A}_\varepsilon$ be given by the heat equation with Neumann boundary conditions, $\mathcal{A}_\varepsilon = e^{\varepsilon \Delta_N}$. It is classical that the heat equation defines a strongly continuous semigroup on $L^2(\Omega_0; \mathbb{R}^d)$, so that the $\mathcal{A}_\varepsilon$ converge strongly to the identity operator on $\mathfrak{H}$ as $\varepsilon \searrow 0$.

If Ω_0 is a rectangle, then we choose $\mathcal{A}_\varepsilon$ as in (47), (48).

10.2. Construction of a generalized solution. Let $\gamma^0 \in \mathfrak{X}$ be a given initial measure, and denote by $\{\gamma_\varepsilon^t, t \geq 0\}$ the global solutions to (35) with $\mathcal{A} = \mathcal{A}_\varepsilon$ which exist by Theorem 9.1.

Lemma 6.1 provides us with a convergent sequence $\gamma_{\varepsilon_k}^t$: write $\gamma_\dagger^t$ for the weak limit. We declare this family of measures to be a generalized solution of (24).

By Proposition 6.2 any generalized solution we construct in this way satisfies the energy inequality (37). Thus a generalized solution decreases the cost functional at least as fast as a smooth solution would, i.e.,

$$\frac{d}{dt}M(\gamma_\dagger^t) \le -\left\|\mathcal{P}\mathbb{E}_{\gamma_\dagger^t}(\Phi_x \mid x)\right\|_{\mathfrak{H}}^2.$$

10.3. ω-limit sets of generalized solutions. We define

$$\omega(\{\gamma_\dagger^t, t \ge 0\}) = \left\{\lambda \in \mathfrak{X} \mid \exists t_j \nearrow \infty : \gamma_\dagger^{t_j} \rightharpoonup \lambda\right\} = \bigcap_{s\ge 0} \overline{\{\gamma_\dagger^t \mid t \ge s\}}.$$

PROPOSITION 10.1. *The ω-limit set of a generalized solution is a closed and connected subset of $\mathfrak{X}$ which consists of critical points for the Monge–Kantorovich functional.*

Proof. Connectedness and closedness follow through entirely conventional arguments from the second description of $\omega(\{\gamma_\dagger^t\})$ given above.

Let $\nu \in \omega(\{\gamma_\dagger^t\})$ be given. Choose a sequence of times $t_k \nearrow \infty$ from which $\gamma_\dagger^{t_k} \rightharpoonup \nu$, and consider the families of measures $\nu_k^t = \gamma_\dagger^{t_k+t}$, $t \in \mathbb{R}$. The arguments in the proof of Proposition 6.2 imply that we can select a subsequence $\nu_{k_j}^t$ which weak* converges for all t. The limit $\nu_\dagger^t$ of this subsequence again satisfies the energy inequality. Moreover, the cost functional is constant on $\nu_\dagger^t$, since

$$M(\nu_\dagger^t) = \lim_{k\to\infty} M(\gamma_\dagger^{t_k+t}) = \lim_{t\to\infty} M(\gamma_\dagger^t).$$

The last limit must exist because $M(\gamma_\dagger^t)$ is a nonincreasing bounded quantity.

The energy inequality for $\nu_\dagger^t$ states that

$$\int_{t_0}^{t_1} \|\mathcal{P}\mathbb{E}_{\nu_\dagger^t}(\Phi_x \mid x)\|_{\mathfrak{H}}^2 \,\mathrm{d}t \le M(\nu_\dagger^{t_0}) - M(\nu_\dagger^{t_1}) = 0$$

so that $\mathcal{P}\mathbb{E}_{\nu_\dagger^t}(\Phi_x \mid x) = 0$ for almost all t. Weak* continuity of ν^t with respect to t strengthens this to $\mathcal{P}\mathbb{E}_{\nu_\dagger^t}(\Phi_x \mid x) = 0$ for all t.

Recalling that $\nu = \nu_\dagger^0$, we conclude that $\mathcal{P}\mathbb{E}_\nu(\Phi_x \mid x) = 0$; i.e., $\nu \in \omega(\{\gamma_\dagger^t\})$ is a critical point. ∎

11. The unregularized flow—smooth solutions. If we omit the smoothing operator, i.e., if we set $\mathcal{A} = I_{\mathfrak{H}}$, then (30) for the rearrangement map s^t,

$$\frac{\partial s^t(x)}{\partial t} = -\int_{\Omega_0\times\Omega_1} \frac{K(s^t(x), s^t(\xi))}{\mu_0(s^t(x))\mu_0(s^t(\xi))} \cdot \Phi_x(s^t(\xi), \eta)\,\mathrm{d}\gamma^0(\xi,\eta),$$

is highly singular, since the kernel K now is the kernel of the Helmholtz projection. The fixed point arguments of section 9 no longer work. Nonetheless, it turns out that a short time existence theorem for solutions of this equation *does* hold if one assumes the initial data are sufficiently regular. In this section we prove such a theorem.

We will assume in this section that the measures γ^t are all defined by measure preserving maps $u^t : \Omega_0 \to \Omega_1$, i.e., $\gamma^t = (\mathrm{id} \times u^t)_\#(\mu_0)$.

Our strategy will be to consider the regularized equation in which $\mathcal{A} = \mathcal{A}_\varepsilon$ is given by a heat operator, as in section 10.1. For each positive ε we have already shown that a global solution exists. The heart of this section is an estimate for how fast the $C^{1,\alpha}$

norm of the map u_ε^t grows with time. The estimate is independent of the mollifying parameter ε if the initial data is smooth. Letting $\varepsilon \searrow 0$ then gives an estimate and short time existence result in $C^{1,\alpha}$ for the unregularized equation.

LEMMA 11.1. *Let $s^t : \Omega_0 \to \Omega_0$ be a solution of the regularized equation* (30) *with $s^0 = \mathrm{id}$. If the initial map $u^0 : \Omega_0 \to \Omega_1$ is $C^{1,\alpha}$, then s^t remains $C^{1,\alpha}$ for a short time $T_* > 0$, and one has $\|\mathrm{d}s^t\|_{0,\alpha} \le C_*$ for $0 \le t \le T_*$, where C_* and T_* depend on the initial data but not on $\varepsilon > 0$.*

11.1. Notation for Hölder norms. For any map $f : \Omega_0 \to \mathbb{R}^N$, we write

$$\begin{aligned}
[f]_\alpha &= \sup_{x,x' \in \Omega_0} \frac{|f(x) - f(x')|}{|x - x'|^\alpha}, \\
\|f\|_{0,\alpha} &= \|f\|_\infty + [f]_\alpha, \\
\|f\|_{1,\alpha} &= \|f\|_\infty + \|\mathrm{d}f\|_\infty + [f]_\alpha.
\end{aligned}$$

The Hölder seminorm $[\cdot]_\alpha$ satisfies the "product-rule estimate,"

$$[f \cdot g]_\alpha \le \|f\|_\infty [g]_\alpha + \|g\|_\infty [f]_\alpha,$$

which one easily derives from $f(x)g(x) - f(y)g(y) = f(x)g(x) - f(x)g(y) + f(x)g(y) - f(y)g(y)$. One then also finds

$$\|f \cdot g\|_{0,\alpha} \le \|f\|_{0,\alpha}\, \|g\|_{0,\alpha}.$$

11.2. Estimates of inverses and compositions. The following proposition shows that we will never have to bother with the case of small $\|s\|_\infty$.

PROPOSITION 11.2. *If $s : \bar\Omega_0 \to \bar\Omega_0$ is a C^1 diffeomorphism, then $\|\mathrm{d}s\|_\infty \ge 1$.*

Consequently we also have $\|s\|_{1,\alpha} \ge \|\mathrm{d}s\|_\infty \ge 1$.

Proof. Since $s(\partial\Omega_0) = \partial\Omega_0$, s cannot be a contraction on all of $\partial\Omega_0$, so somewhere on $\partial\Omega_0$ one has $|\mathrm{d}s| \ge 1$. □

Let $s : \Omega_0 \to \Omega_0$ be a C^1 diffeomorphism which preserves μ_0, i.e., for which

$$\mu_0(s(x)) \det \mathrm{d}s(x) = \mu_0(x) \tag{63}$$

holds.

LEMMA 11.3. *Assume the diffeomorphism $s : \Omega_0 \to \Omega_0$ satisfies* (63). *If*

$$K = \max_{x,x' \in \Omega_0} \frac{\mu_0(x)}{\mu_0(x')},$$

then

$$\sup_{\Omega_0} |(\mathrm{d}s(x))^{-1}| \le C_d \left(K \sup_{\Omega_0} |\mathrm{d}s(x)| \right)^{d-1}$$

for some constant C_d which only depends on the dimension d. In particular, if s is Lipschitz continuous with Lipschitz constant L, then s^{-1} is Lipschitz continuous with constant at most $C_d(KL)^{d-1}$.

Proof. We represent $\mathrm{d}s(x)$ as a $d \times d$ matrix. Then

$$(\mathrm{d}s(x))^{-1} = \frac{1}{\det \mathrm{d}s(x)} (\mathrm{d}s(x))^\# = \frac{\mu_0(x)}{\mu_1(s(x))} (\mathrm{d}s(x))^\#, \tag{64}$$

where $\mathrm{d}s^t(x)^\#$ is the cofactor matrix. This matrix is a polynomial of degree $d-1$ in the entries of the matrix $\mathrm{d}s^t$, hence the lemma. ∎

LEMMA 11.4. *Assume the diffeomorphism $s:\Omega_0\to\Omega_0$ satisfies* (63). *If $s\in C^{1,\alpha}$, then $s^{-1}\in C^{1,\alpha}$, and*

$$\|s^{-1}\|_{1,\alpha}\le C\|s\|_{1,\alpha}^{(d-1)(2+\alpha)+1}, \tag{65}$$

where the constant C only depends on μ_0 and the dimension d.

Proof. We will henceforth write $\mathrm{d}s(x)=\mathrm{d}s_x$ if it seems to improve the notation.

We get an estimate for the supremum norm of $\mathrm{d}(s^{-1})$ from the inverse function theorem, which says $\mathrm{d}(s^{-1})=(\mathrm{d}s)^{-1}\circ s^{-1}$, so that $\|\mathrm{d}(s^{-1})\|_\infty=\|(\mathrm{d}s)^{-1}\|_\infty\le C\|\mathrm{d}s\|_\infty^{d-1}$.

To estimate the Hölder seminorm $[\mathrm{d}s^{-1}]_\alpha$ we compute for $x,y\in\Omega_0$

$$\begin{aligned}
|\mathrm{d}(s^{-1})(x)-\mathrm{d}(s^{-1})(y)| &= \left|(\mathrm{d}s)^{-1}_{s^{-1}(x)}-(\mathrm{d}s)^{-1}_{s^{-1}(y)}\right|\\
&= \left|(\mathrm{d}s)^{-1}_{s^{-1}(y)}\left\{\mathrm{d}s_{s^{-1}(x)}-\mathrm{d}s_{s^{-1}(y)}\right\}\mathrm{d}(s^{-1})_{s^{-1}(x)}\right|\\
&\le \|(\mathrm{d}s)^{-1}\|_\infty^2\left\|\mathrm{d}s_{s^{-1}(x)}-\mathrm{d}s_{s^{-1}(y)}\right\|\\
&\le \|(\mathrm{d}s)^{-1}\|_\infty^2[\mathrm{d}s]_\alpha\left|s^{-1}(x)-s^{-1}(y)\right|^\alpha\\
&\le \|(\mathrm{d}s)^{-1}\|_\infty^{2+\alpha}[\mathrm{d}s]_\alpha|x-y|^\alpha\\
&\le C\|\mathrm{d}s\|_\infty^{(d-1)(2+\alpha)}[\mathrm{d}s]_\alpha|x-y|^\alpha.
\end{aligned}$$

Hence we get

$$[d(s^{-1})]_\alpha\le C\|\mathrm{d}s\|_\infty^{(d-1)(2+\alpha)}[\mathrm{d}s]_\alpha\le C\|s\|_{1,\alpha}^{(d-1)(2+\alpha)+1}.$$

To estimate the full $C^{1,\alpha}$ norm of s^{-1} we add the lower order terms,

$$\begin{aligned}
\|s^{-1}\|_{1,\alpha} &= \|s^{-1}\|_\infty+\|\mathrm{d}s^{-1}\|_\infty+[d(s^{-1})]_\alpha\\
&\le C+C\|\mathrm{d}s\|_\infty^{d-1}+C\|s\|_{1,\alpha}^{(d-1)(2+\alpha)+1}\\
&\le C+C\|s\|_{1,\alpha}^{(d-1)(2+\alpha)+1}.
\end{aligned}$$

Finally we use $\|s\|_{1,\alpha}\ge 1$ to get (65). ∎

We will occasionally use the following crude estimate for the $C^{1,\alpha}$ norm of the composition of two maps.

LEMMA 11.5. *For two $C^{1,\alpha}$ maps f,g one has*

$$\begin{aligned}
[f\circ g]_\alpha &\le \|\mathrm{d}g\|_\infty\cdot[f]_\alpha,\\
\|f\circ g\|_{0,\alpha} &\le \left(1+\|\mathrm{d}g\|_\infty^\alpha\right)\|f\|_{0,\alpha},\\
\|f\circ g\|_{1,\alpha} &\le 3\big(1+\|g\|_{1,\alpha}^{1+\alpha}\big)\|f\|_{1,\alpha}.
\end{aligned}$$

Proof. The first inequality follows directly from

$$|f(g(x))-f(g(y))|\le[f]_\alpha|g(x)-g(y)|^\alpha\le[f]_\alpha\|\mathrm{d}g\|_\infty^\alpha\,|x-y|^\alpha.$$

The second inequality follows from

$$\begin{aligned}
\|f\circ g\|_{0,\alpha} &= \|f\circ g\|_\infty+[f\circ g]_\alpha\\
&\le \|f\|_\infty+\|\mathrm{d}g\|_\infty^\alpha\cdot[f]_\alpha\\
&\le \|f\|_{0,\alpha}+\|\mathrm{d}g\|_\infty^\alpha\|f\|_{0,\alpha}.
\end{aligned}$$

To prove the third inequality, we compute

$$\begin{aligned}\|f\circ g\|_{1,\alpha} &= \|f\circ g\|_\infty + \|(\mathrm{d}f\circ g)\cdot \mathrm{d}g\|_{0,\alpha}\\ &\le \|f\|_\infty + \|\mathrm{d}f\circ g\|_{0,\alpha}\|\mathrm{d}g\|_{0,\alpha}\\ &\le \|f\|_\infty + \bigl(1+\|\mathrm{d}g\|_\infty^\alpha\bigr)\|\mathrm{d}f\|_{0,\alpha}\|\mathrm{d}g\|_{0,\alpha}\\ &\le \bigl(1+\|\mathrm{d}g\|_{0,\alpha}+\|\mathrm{d}g\|_{0,\alpha}^{1+\alpha}\bigr)\|f\|_{1,\alpha}\\ &\le 3\bigl(1+\|\mathrm{d}g\|_{0,\alpha}^{1+\alpha}\bigr)\|f\|_{1,\alpha}\end{aligned}$$

since $1+x+x^{1+\alpha}\le 3(1+x^{1+\alpha})$ for all $x\ge 0$. □

11.3. Proof of Lemma 11.1. We summarize the relations that define the maps u^t.

First, u^t and the initial map u^0 are related by

$$u^t = u^0\circ (s^t)^{-1}. \tag{66}$$

The rearrangement maps s^t move with velocity field v^t. This gives two equations, one for s^t and one for its space derivative:

$$\frac{\partial s^t}{\partial t} = v^t\circ s^t, \qquad \frac{\partial \mathrm{d}s^t}{\partial t} = (\mathrm{d}v^t\circ s^t)\cdot \mathrm{d}s^t. \tag{67}$$

The velocity field v^t is determined by the map u^t via

$$v^t(x) = \frac{-1}{\mu_0(x)}\mathcal{P}\mathcal{A}_\varepsilon^2\mathcal{P}W^t, \tag{68}$$

while W^t is given by $W^t = \mathbb{E}_{\gamma^t}(\Phi_x \mid x)$, i.e., by

$$W^t(x) = \Phi_x(x,u^t(x)). \tag{69}$$

We begin our estimate of $\|\partial_t s^t\|_{1,\alpha}$ as follows:

$$\begin{aligned}\left\|\frac{\partial}{\partial t}s^t\right\|_{1,\alpha} &= \|\partial_t s^t\|_\infty + \left\|\frac{\partial}{\partial t}\mathrm{d}s^t\right\|_{0,\alpha}\\ &= \|v^t\|_\infty + \|(\mathrm{d}v^t\circ s^t)\cdot \mathrm{d}s^t\|_{0,\alpha}\\ &\le \|v^t\|_\infty + \|\mathrm{d}v^t\circ s^t\|_{0,\alpha}\cdot\|\mathrm{d}s^t\|_{0,\alpha}\\ &\le \|v^t\|_\infty + \|\mathrm{d}v^t\|_{0,\alpha}\bigl(1+\|\mathrm{d}s^t\|_\infty^\alpha\bigr)\|\mathrm{d}s^t\|_{0,\alpha}\\ &\le \|v^t\|_\infty + 2\|\mathrm{d}v^t\|_{0,\alpha}\|\mathrm{d}s^t\|_{0,\alpha}^{1+\alpha}\\ &\le \|v^t\|_\infty + 2\|v^t\|_{1,\alpha}\|s^t\|_{1,\alpha}^{1+\alpha}\\ &\le 3\|v^t\|_{1,\alpha}\|s^t\|_{1,\alpha}^{1+\alpha},\end{aligned} \tag{70}$$

where we have used $\|\mathrm{d}s^t\|_\infty\ge 1$ again.

Next, we estimate the $C^{1,\alpha}$ norm of the velocity field:

$$\begin{aligned}\|v^t\|_{1,\alpha} &= \left\|\frac{-1}{\mu_0(x)}\mathcal{P}\mathcal{A}^2\mathcal{P}\bigl(\Phi_x(x,u^t(x))\bigr)\right\|_{1,\alpha}\\ &\le C\|\Phi_x(x,u^t(x))\|_{1,\alpha}\\ &\le 3C\|\Phi\|_{2,\alpha}\bigl(1+\|u^t\|_{1,\alpha}^{1+\alpha}\bigr)\\ &\le C\bigl(1+\|u^t\|_{1,\alpha}^{1+\alpha}\bigr).\end{aligned} \tag{71}$$

Here we have used the facts that $\mu_0 \in C^{1,\alpha}$, that the smoothing operators $\mathcal{A}$ are uniformly bounded from $C^{1,\alpha}$ to $C^{1,\alpha}$, and that the Helmholtz decomposition is also bounded on $C^{1,\alpha}$.

Finally we estimate $\|u^t\|_{1,\alpha}$ in terms of $\|s^t\|_{1,\alpha}$. We have

$$\begin{aligned} \|u^t\|_{1,\alpha} &= \|u^0 \circ \big(s^t\big)^{-1}\|_{1,\alpha} \\ &\le 3\big(1 + \|(s^t)^{-1}\|_{1,\alpha}^{1+\alpha}\big)\ \|u^0\|_{1,\alpha} \\ &\le C\ \|(s^t)^{-1}\|_{1,\alpha}^{1+\alpha}\ \|u^0\|_{1,\alpha} \\ &\le C\ \|s^t\|_{1,\alpha}^{(d-1)(2+\alpha)(1+\alpha)+1+\alpha}. \end{aligned} \tag{72}$$

Combining (70), (71), and (72), we arrive at

$$\frac{d}{dt}\|s^t\|_{1,\alpha} \le \|\partial_t s^t\|_{1,\alpha} \le C\left(\|s^t\|_{1,\alpha}\right)^{\kappa}, \tag{73}$$

where

$$\begin{aligned} \kappa &= \{(d-1)(2+\alpha)(1+\alpha) + 1 + \alpha\}(1+\alpha) + 1 + \alpha \\ &= \big((d-1)\alpha^2 + 3(d-1)\alpha + 2d\big)(1+\alpha). \end{aligned}$$

Integrate this ODE, using the initial data $\|s^0\|_{1,\alpha} = 1$ which derives from $s^0 = \mathrm{id}$, and you get

$$\|s^t\|_{1,\alpha} \le \big(1 - Ct\big)^{-\frac{1}{\kappa-1}}.$$

The constant C depends on the initial map u^0 but not on the smoothing parameter $\varepsilon > 0$, as claimed in Lemma 11.1.

12. Numerical methods and examples. In this section, we describe some of the techniques we use to numerically solve (9), as well as how we compute the initial mapping. Briefly, we have employed an upwinding scheme when computing ∇u^t and the FFT when inverting the Laplacian on a rectangular grid. Standard centered differences were used for the other spatial derivatives. In practice, we iterate until the mean absolute curl is sufficiently small. More details of the numerical implementation for solving (9) are given below. See also [11, 12].

12.1. Finding an initial mapping. In this section, we describe our procedure for finding the initial mass preserving mapping u for (9). We work here on the unit square. An initial mapping for general domains can also be obtained using a method of Moser [6].

So we work in $\mathbb{R}^2$ and assume $\Omega_0 = \Omega_1 = [0,1]^2$, the generalization to higher dimensions being straightforward. The idea of this construction is that we solve a family of one-dimensional mass transport problems. In one dimension, the optimal transport map can be found by simple quadrature. We first transport mass along lines parallel to the x-axis, and then afterward transport mass along lines parallel to the y-axis. Accordingly, we define a function $a = a(x)$ by the equation

$$\int_0^{a(x)} \int_0^1 \mu_1(\eta, y)\ dy\ d\eta = \int_0^x \int_0^1 \mu_0(\eta, y)\ dy\ d\eta, \tag{74}$$

which gives by differentiation with respect to x

$$a'(x)\int_0^1 \mu_1(a(x),y)\,dy = \int_0^1 \mu_0(x,y)\,dy. \tag{75}$$

We may now define a function $b = b(x,y)$ by the equation

$$a'(x)\int_0^{b(x,y)} \mu_1(a(x),\rho)\,d\rho = \int_0^y \mu_0(x,\rho)\,d\rho \tag{76}$$

and set $u(x,y) = (a(x), b(x,y))$. Since $a_y = 0$, $|Du| = a_x b_y$, and differentiating (76) with respect to y we find

$$\begin{aligned} a'(x)\; b_y(x,y)\; \mu_1(a(x),b(x,y)) &= \mu_0(x,y), \\ |Du|\; \mu_1 \circ u &= \mu_0, \end{aligned}$$

which is the mass preserving property we need. In practice, a and b can be found with simple numerical integration techniques.

12.2. Defining the warping map. Typically in elastic registration, one wants to see an explicit warping which smoothly deforms one image into the other [11]. This can easily be done using the solution of the Monge–Kantorovich problem. Thus, we assume now that we have applied our gradient descent process as described above and that it has converged to the optimal L^2 Monge–Kantorovich mapping u_{MK}.

Following the work of Benamou and Brenier [3] (see also [9]), we consider the related problem

$$\inf \int\int_0^1 \mu(t,x)|v(t,x)|^2\,dt\,dx \tag{77}$$

over all time varying densities μ and velocity fields v satisfying

$$\frac{\partial \mu}{\partial t} + \operatorname{div}(\mu v) = 0, \tag{78}$$

$$\mu(0,\cdot) = \mu_0, \quad \mu(1,\cdot) = \mu_1. \tag{79}$$

It is shown in [3] that this infimum is attained for some μ_{min} and v_{min}, and that it is equal to the L^2 Kantorovich–Wasserstein distance between μ_0 and μ_1. Recall that this distance is defined by

$$d_2(\mu_0,\mu_1)^2 := \inf_u \int_{\Omega_0} |u(x) - x|^2\,dx,$$

the infimum taken over all diffeomorphisms which satisfy the Jacobian condition (1). Further, the flow $X = X(x,t)$ corresponding to the minimizing velocity field v_{min} via

$$X(x,0) = x, \quad X_t = v_{min} \circ X \tag{80}$$

is given simply as

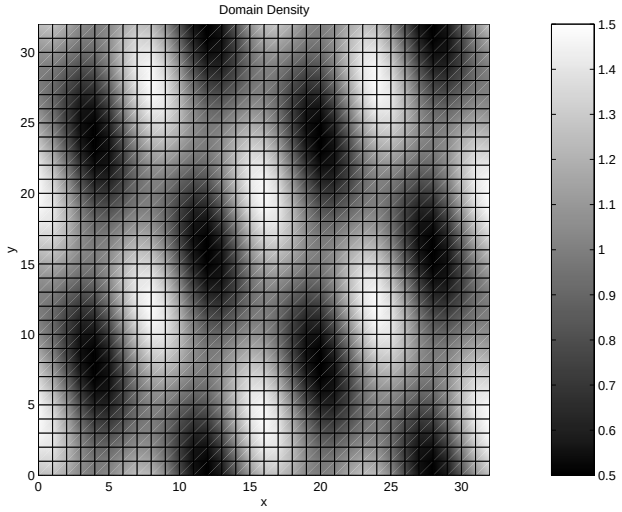


FIG. 3. *Density* μ_1 *on* Ω_0.

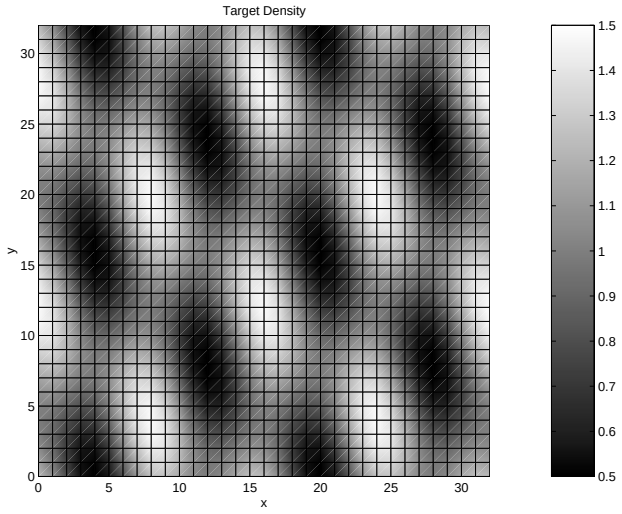


FIG. 4. *Density* μ_1 *on* Ω_1.

$$X(x,t) = x + t\ (u_{MK}(x) - x). \tag{81}$$

Note that when $t = 0$, X is the identity map, and when $t = 1$, it is the solution u_{MK} to the Monge–Kantorovich problem. This analysis provides appropriate justification for using (81) to *define* our continuous warping map X between the densities μ_0 and μ_1.

13. Implementation and examples. We illustrate our methods with the following examples. The first is the mapping of one synthetic density onto another. Figure 3 shows a mass distribution μ_0 on Ω_0, with dark regions representing little mass, lighter regions representing more. Similarly, Figure 4 indicates the density μ_1 on Ω_1. Figure 5 represents the initial mapping u, which was obtained by the method described above. The shading in this figure represents the Jacobian of u. Figure 6

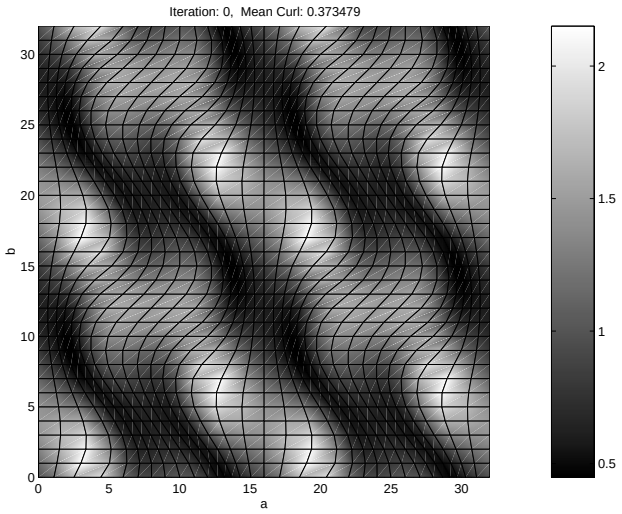


FIG. 5. *Initial mapping from* Ω_0 *to* Ω_1.

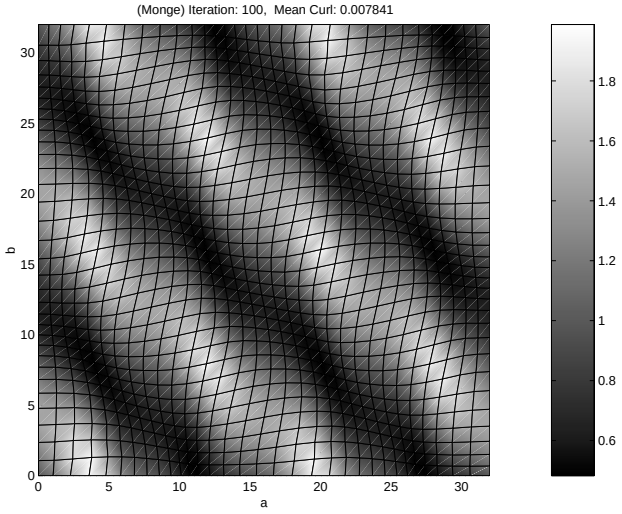


FIG. 6. *Final Monge–Kantorovich mapping from* Ω_0 *to* Ω_1.

shows the nearly optimal Monge–Kantorovich mapping obtained using the nonlocal first order equation (9). One can see that the effect of removing the curl is to straighten out the grid lines somewhat. On a Sun Ultra10, this process took just a few seconds.

In Figures 7 through 10 we show a brain deformation sequence obtained with MRI. The first and last images were given, and the intermediate two were found using our process. This type of elastic brain deformation occurs during surgery, after the skull is opened. These two-dimensional slices were extracted from an original three-dimensional data set ($256 \times 256 \times 124$) to which the registration algorithm was applied. We should note that in contrast to other elastic approaches based on fluid and continuum mechanics ideas (see [17], especially Chapters 1 and 18 for a general discussion) in which the computations may take hours, in our case the three-dimensional set was processed in about half an hour with very reasonable results.

FIG. 7. *Brain warping:* $t = 0.00$.

FIG. 8. *Brain warping:* $t = 0.33$.

In general, the target domain Ω_1 need not be rectangular when using the nonlocal method. However, we note that if the periodic boundary conditions are used on the displacement, as in section 7.1, then the Laplacian in (9) can be inverted using the FFT alone, without the need to solve a subsequent matrix system. For the brain warp, this reduced the processing time by about 1/3.

Acknowledgment. We would like to thank Robert McCann of the University of Toronto for some very useful discussions about the Monge–Kantorovich problem.

FIG. 9. *Brain warping:* $t = 0.66$.

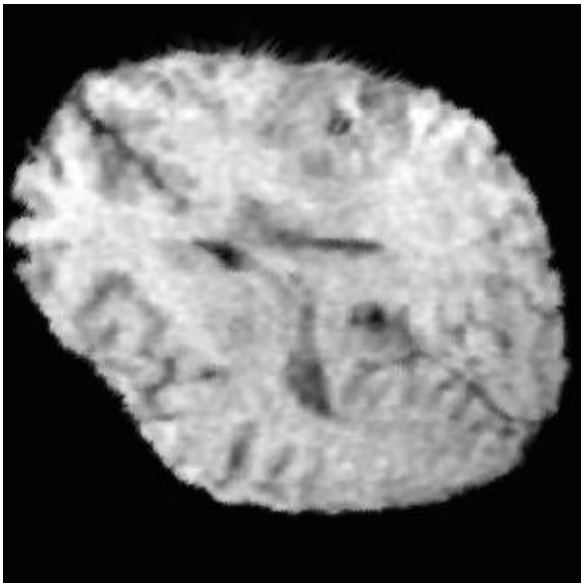


FIG. 10. *Brain warping:* $t = 1.00$.

REFERENCES

[1] L. AMBROSIO, *Lecture Notes on Optimal Transport Theory*, CIME Series of Springer Lecture Notes, Euro Summer School Mathematical Aspects of Evolving Interfaces, Madeira, Portugal, Springer-Verlag, New York, 2000.

[2] S. ANGENENT, S. HAKER, A. TANNENBAUM, AND R. KIKINIS, *On area preserving maps of minimal distortion,* in System Theory: Modeling, Analysis, and Control, T. Djaferis and I. Schick, eds., Kluwer Academic, The Netherlands, 1999, pp. 275–287.

[3] J.-D. BENAMOU AND Y. BRENIER, *A computational fluid mechanics solution to the Monge–*

Kantorovich mass transfer problem, Numer. Math., 84 (2000), pp. 375–393.
[4] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions,* Comm. Pure Appl. Math., 64 (1991), pp. 375–417.
[5] M. Cullen and R. Purser, *An extended Lagrangian theory of semigeostrophic frontogenesis,* J. Atmospheric Sci., 41 (1984), pp. 1477–1497.
[6] B. Dacorogna and J. Moser, *On a partial differential equation involving the Jacobian determinant,* Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 1–26.
[7] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1998.
[8] L. C. Evans, *Partial differential equations and Monge–Kantorovich mass transfer,* in Current Developments in Mathematics, International Press, Boston, MA, 1999, pp. 65–126.
[9] W. Gangbo and R. McCann, *The geometry of optimal transportation,* Acta Math., 177 (1996), pp. 113–161.
[10] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order,* Springer-Verlag, Berlin, 2001.
[11] S. Haker, A. Tannenbaum, and S. Angenent, *Optimal transport and image registration,* Internat. J. Comput. Vision, submitted.
[12] S. Haker, A. Tannenbaum, and R. Kikinis, *Mass-preserving mappings and surface registration,* in Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI'01), 4th International Conference, Utrecht, The Netherlands, 2001.
[13] L. V. Kantorovich, *On a problem of Monge,* Uspekhi Mat. Nauk., 3 (1948), pp. 225–226.
[14] M. Knott and C. Smith, *On the optimal mapping of distributions,* J. Optim. Theory, 43 (1984), pp. 39–49.
[15] S. Rachev and L. Rüschendorf, *Mass Transportation Problems*, Vol. I, Probab. Appl., Springer-Verlag, New York, 1998.
[16] S. Rachev and L. Rüschendorf, *Mass Transportation Problems*, Vol. II, Probab. Appl., Springer-Verlag, New York, 1998.
[17] A. Toga, *Brain Warping*, Academic Press, San Diego, 1999.

SIAM J. MATH. ANAL.
Vol. 35, No. 1, pp. 98–122

FRACTIONAL RATE OF CONVERGENCE FOR VISCOUS APPROXIMATION TO NONCONVEX CONSERVATION LAWS*

TAO TANG†, ZHEN-HUAN TENG‡, AND ZHOUPING XIN§

Abstract. This paper considers the viscous approximations to conservation laws with nonconvex flux function. It is shown that if the entropy solutions are piecewise smooth, then the rate of L^1-convergence is a *fractional* number in $(0.5, 1]$. This is in contrast to the corresponding result for the convex conservation laws. Numerical experiments indicate that the theoretical prediction for the convergence rate is optimal.

Key words. rate of convergence, error estimate, viscous approximation, conservation law, nonconvex flux

AMS subject classifications. 35L65, 49M25, 34A65

DOI. 10.1137/S0036141001388993

1. Introduction.

In this paper, we consider the initial value problem for *nonconvex* conservation laws

$$\partial_t u + \partial_x f(u) = 0, \qquad t > 0\,, \quad x \in \mathbf{R}, \tag{1.1}$$

which is subject to the initial condition prescribed at $t = 0$,

$$u(x, 0) = u_0(x), \tag{1.2}$$

where $f \in C^2$. We shall investigate viscous approximations to the entropy solution of (1.1):

$$\partial_t u^\epsilon + \partial_x f(u^\epsilon) = \epsilon \partial_{xx} u^\epsilon \tag{1.3}$$

subject to the initial data

$$u^\epsilon(x, 0) = u_0(x). \tag{1.4}$$

In this work, we assume that $f''(u)$ vanishes at a *finite number* of points. It is also assumed that the entropy solution to (1.1) and (1.2) is *piecewise smooth* with finitely many shock discontinuities. The existence and uniqueness of the solutions to (1.1) in the class of piecewise smooth weak solutions were studied by Ballou [1].

When the flux f is convex, the solution structure for (1.1) and (1.2) has been obtained; see, e.g., Lax [10] and Dafermos [2]. If f has inflection points, then the

*Received by the editors May 8, 2001; accepted for publication (in revised form) November 30, 2002; published electronically June 10, 2003.
http://www.siam.org/journals/sima/35-1/38899.html

†Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong (ttang@math.hkbu.edu.hk). The research of this author was supported by the Hong Kong Research Grants Council.

‡School of Mathematical Sciences, Peking University, Beijing 100871, People's Republic of China (tengzh@math.pku.edu.cn). The research of this author was supported by the China State Major Key Project for Basic Research.

§Courant Institute of Mathematical Sciences, New York University, New York, NY, and Department of Mathematics and Institute of Mathematical Sciences, The Chinese University of Hong Kong, Hong Kong (zpxin@ims.cuhk.edu.hk). The research of this author was supported by the Hong Kong Research Grants Council, NSF grants, and a DOE grant.

situation is more complicated. In this case, some analysis for the solution structure and asymptotic behavior has been done; see, e.g., Dafermos [3], Liu [11], and Zumbrun [32]. However, we are still far from having a complete understanding of this general case, since the geometric structure of the solution, when f changes convexity, is much more complicated due to the presence of contact discontinuities, and there is a large variety of asymptotic states.

The asymptotic convergence of solutions to the viscous problem (1.3) and (1.4) to the corresponding discontinuous solutions of the inviscid problem (1.1) and (1.2) has been the main driving force for the mathematical theory of shock waves from both theoretical and numerical points of view. Substantial progress has been made in the past in this regard (see [29, 20] and the references therein), pioneered by Hopf, Lax [10], Oleinik [18], and Krushkov [7], to name a few. For BV entropy solutions, Kuznetsov [8] was the first to establish the half-order rate of L^1-convergence for viscosity approximation and monotone schemes. It was proved by Tang and Teng [25] that this half-order rate of convergence is optimal in the BV solution class; see also Sabac [19]. However, for convex conservation laws with piecewise smooth solutions the L^1-convergence rate can be improved to first-order; see, e.g., Teng and Zhang [27] for the monotone scheme, Tang and Teng [24] for viscosity approximation, and Teng [26] for the relaxation method. The basic method in obtaining the first-order rate of convergence is the matching asymptotic method developed by Goodman and Xin [5] and Liu and Xin [14]. One of the key ingredients in this method is the nonlinear large asymptotic stability of viscous shock profiles. For systems of viscous conservation laws, this stability theory has been extensively studied in the past decade. Important progress has been made by Goodman [4], Matsumura and Nishihara [16], Liu [12], and Szepessy and Xin [21]; see also some recent new approaches by Howard and Zumbrun [6], Liu [13], and Kreiss and Kreiss [9]. In particular, convergence with a rate to viscous shock profiles was obtained by Liu [13] by using a pointwise estimate for the approximate Green's function. Even in the case of nonconvex fluxes, the nonlinear large time asymptotic stability has been established for some special systems; see, e.g., [15] and [17]. The convergence of viscous solutions to piecewise smooth solutions for general systems was established by Goodman and Xin [5]; see also [31] for a recent improvement. For the convergence of viscous solutions in the presence of physical boundaries, we refer to [30] and the references therein. We also point out that there are some first-order *pointwise* convergence results for viscous approximations to convex conservation laws; see, e.g., Tadmor and Tang [22, 23], who used the energy method with some bootstrap extrapolation technique. It is proved in [24] that, for convex conservation laws whose entropy solution consists of finitely many discontinuities, the L^1-error between the viscosity solution u^ϵ and its inviscid limit u is bounded by $\mathcal{O}(\epsilon|\ln\epsilon|)$. If neither central rarefaction waves nor spontaneous shocks occur, the error bound is improved to $\mathcal{O}(\epsilon)$; see also [28]. In this work, we will show that for *nonconvex* conservation laws, the L^1-error between the viscosity solution and its inviscid limit is bounded by $\mathcal{O}(\epsilon^\alpha|\ln\epsilon|)$, where $\frac{1}{2} < \alpha \le 1$, even in the piecewise smooth solution class. The constant α is determined by the index numbers of shock curves to be defined in the next section. Based on the form of the flux function, the rate α can be any number between $\frac{1}{2}$ and 1. This result suggests that for the viscous approximations the L^1-convergence rate of the nonconvex conservation laws is substantially different from that of the convex ones.

We close the introduction by outlining the rest of the paper. In the next section, we give some preliminaries, define an index number for a shock discontinuity, and list

some properties of the index number. In section 3, we state our main convergence theorem, whose proof occupies section 4 to section 8. Finally, in section 9 numerical experiments are performed to verify the theoretical estimates.

2. Piecewise smooth solution. Throughout this paper, we assume that the entropy solution of (1.1) and (1.2) is *piecewise smooth*, with finitely many shock discontinuities. More precisely, we can divide the given time interval $[0, T]$ into finite intervals $\{[t_{m-1}, t_m]\}_{m=1}^{M}$ such that in each interval $[t_{m-1}, t_m]$ the entropy solution is a finite combination of the cases plotted in Figures 1 and 2 (demonstrated in the case with three inflection points for $f(u)$). Thus, if we denote by $S(t)$ the set of the discontinuous curve of $u(\cdot, t)$ in the time interval $[t_{m-1}, t_m]$, then it consists of finitely many shocks:

$$S(t) := \{(x,t) \,|\, x = X_k(t),\ 1 \le k \le K;\ t_{m-1} \le t \le t_m\},$$

where $X_k(t) < X_{k+1}(t)$ for $t \in (t_{m-1}, t_m)$. It is understood that u is smooth with bounded limits $u(X_k(t) \pm 0, t)$ (denote by $u_k^{\pm}(t)$) and $u_x(X_k(t) \pm 0, t)$. For simplicity, we will not consider the newly formed shock wave here, although this case was investigated extensively in [24]. As a consequence, we always have $u_k^+(t) - u_k^-(t) \neq 0$. For ease of notation we omit the dependence of $S(t)$, $X_k(t)$, and K on m. Each of the *noncontact* shocks $X_k(t)$, plotted in Figure 1, satisfies the Rankine–Hugoniot and the Lax conditions

$$X_k'(t) = \sigma(u_k^+(t), u_k^-(t)) := \frac{f(u_k^+(t)) - f(u_k^-(t))}{u_k^+(t) - u_k^-(t)}, \tag{2.1}$$

$$a(u_k^-(t)) > X_k'(t) > a(u_k^+(t)), \quad \text{where} \quad a(v) := f'(v). \tag{2.2}$$

Each of the *contact shocks* $X_k(t)$, plotted in Figure 2, satisfies the Rankine–Hugoniot and the contact conditions

$$X_k'(t) = \sigma(u_k^+(t), u_k^-(t)), \tag{2.3}$$

$$a(u_k^-(t)) > X_k'(t) = a(u_k^+(t)), \quad \text{and/or} \tag{2.4}$$

$$a(u_k^-(t)) = X_k'(t) > a(u_k^+(t)). \tag{2.5}$$

We now define *index numbers* $\beta_k^{\pm}$ for a shock curve $x = X_k(t)$:

1. If $X_k'(t) > a(u_k^+)$, then the index number $\beta_k^+ = 0$.
2. If $X_k'(t) = a(u_k^+)$ and there exists a positive number $\beta > 0$ such that

$$|a(u_k^+) - a(u)| \sim |u_k^+ - u|^{\beta} \quad \text{as} \quad u \to u_k^+, \tag{2.6}$$

   then the index number $\beta_k^+ = \beta$. In (2.6), the notation " $\sim$ " means equivalence. More precisely, $g(u) \sim h(u)$ as $u \to c$ means that there exists a constant $D > 0$ such that $D^{-1}h \le g \le Dh$ as $u \to c$.
3. Similarly, we can define β_k^- for the shock curve $x = X_k(t)$.

The following result gives a rule for calculating the index number.

THEOREM 2.1. *If $f(u) \in C^r(\mathbf{R})$, $\sigma(u_k^+, u_k^-) = a(u_k^+)$, the derivative of $a(u)$ is zero at $u = u_k^+$ up to $(r-1)$th order but $a^{(r)}(u_k^+) \neq 0$, then $\beta_k^+ = r$.*

Proof. Applying Taylor's theorem to $a(u)$ gives

$$a(u) - a(u_k^+) = \frac{1}{r!} a^{(r)}(u_k^+)(u - u_k^+)^r + o(|u - u_k^+|^r).$$

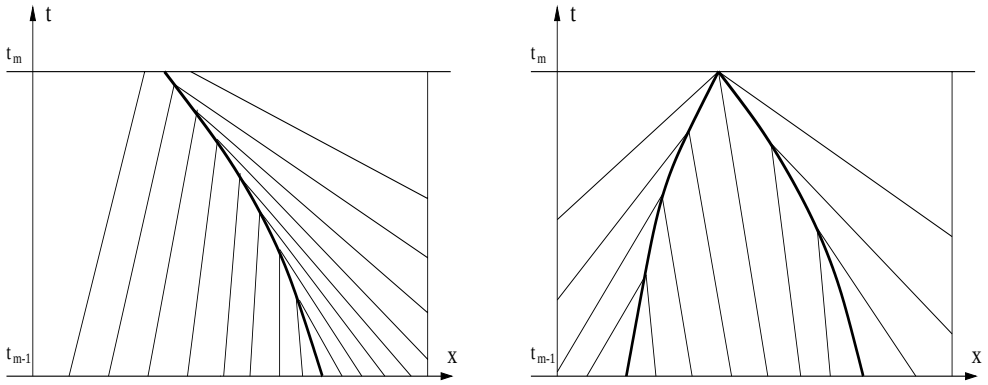


FIG. 1. *Illustration of noncontact shocks: Thin lines are characteristics, and thick ones are* noncontact *shock curves. Here characteristics come into shocks from both sides.*

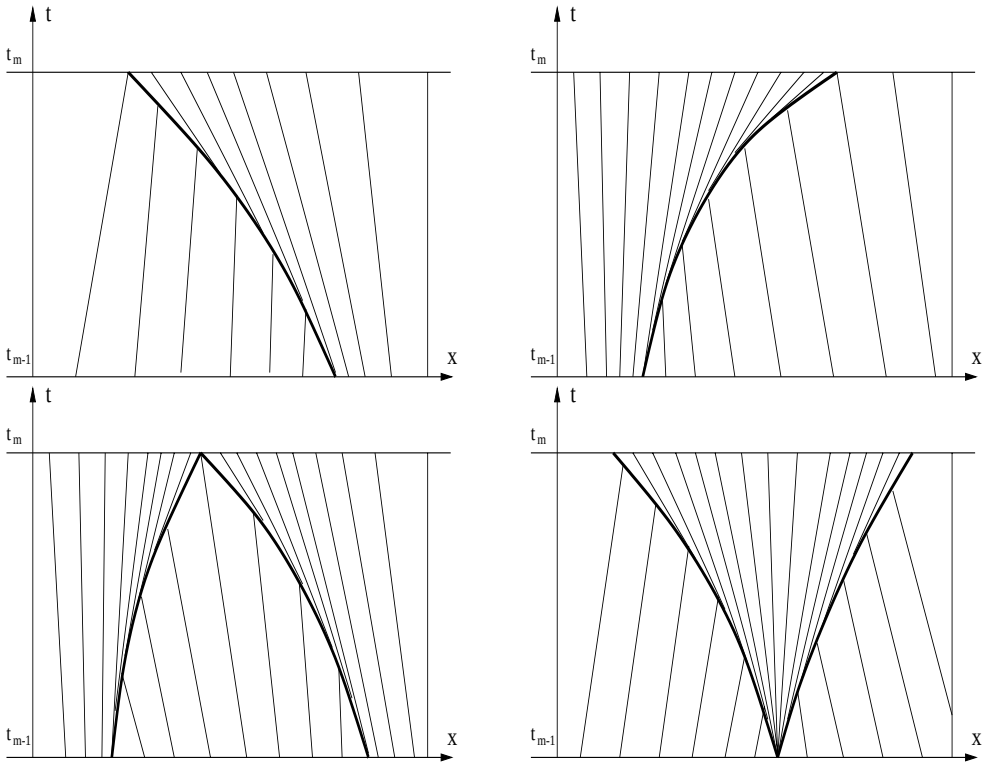


FIG. 2. *Illustration of contact shocks: Thin lines are characteristics, and thick ones are* contact *shock curves. Here characteristics are tangent to shocks at least in one side.*

This means that $|a(u) - a(u_k^+)| \sim |u - u_k^+|^r$ as $u \to u_k^+$. Therefore, it follows from the definition (2.6) that $\beta_k^+ = r$. □

The following result is an immediate consequence of the above theorem.

COROLLARY 2.1. *If $f''(u)$ has only a finite number of zero points, then the index number can take only a finite number of values.*

Example 2.1. Let $f(u) = u^{2m+1}$. Assume the entropy solution is of the form

$$u(x,t) = \begin{cases} u_- & \text{for } x < \sigma(u_+, u_-)t, \\ u_+ & \text{for } x \geq \sigma(u_+, u_-)t, \end{cases} \tag{2.7}$$

where $u_- > 0$ is a given number, and $u_+ < 0$ is the solution of the equation $\sigma(u_+, u_-) = a(u_+)$. In other words, u_+ is determined by

$$\sum_{s=0}^{2m} (u_+)^{2m-s}(u_-)^s = (2m+1)(u_+)^{2m}.$$

It is easy to show that $a(u_-) > \sigma(u_+, u_-) = a(u_+)$, and hence $\beta^- = 0$. Since $a'(u_+) > 0$, it follows from Theorem 2.1 that $\beta^+ = 1$.

Example 2.2. If $f(u) = (1-u)^p(1+u)^q$ with $p \geq 1$, $q \geq 1$, and $p+q > 2$ and an entropy solution is given by

$$u(x,t) = \begin{cases} -1, & x < 0, \\ +1, & x \geq 0, \end{cases} \tag{2.8}$$

then the curve $x = X(t) = 0$ is a contact shock with $\sigma(1,-1) = a(1) = a(-1)$, and the index numbers are $\beta^+ = p - 1$ and $\beta^- = q - 1$.

3. Main theorem. In this section, the main result of this paper presented; its proof will be given in the next few sections.

THEOREM 3.1. *Let $f \in C^2$ and assume that f'' may change its sign at most at a finite number of points. Let u be the piecewise smooth entropy solution of* (1.1)–(1.2) *with finitely many shock discontinuities, and let u^ϵ be the viscosity solution of* (1.3)–(1.4). *Then the following error estimates hold for any $0 < t \leq T$:*

$$\|u^\epsilon(\cdot,t) - u(\cdot,t)\|_{L^1(\mathbf{R})} \leq \begin{cases} C(T)\epsilon|\ln\epsilon| & \text{for } \bar\beta < 1, \\ C(T)\epsilon^{(1+1/\bar\beta)/2}|\ln\epsilon| & \text{for } \bar\beta \geq 1, \end{cases} \tag{3.1}$$

where $\bar\beta = \max_{0\leq t\leq T}\beta(t)$, $\beta = \max\{\beta^+, \beta^-\}$, $\beta^\pm = \max_k\{\beta_k^\pm\}$, and $\beta_k^\pm$ are index numbers defined by (2.6).

The above theorem will be established by using a matched asymptotic analysis, a stability lemma, and some detailed analysis for the traveling wave solution. The stability lemma to be used is valid only for the scalar conservation laws, which makes the present analysis much simpler than the system case. For the hyperbolic system, Goodman and Xin [5] constructed high-order approximations in obtaining a local first-order rate of convergence for the viscous approximations.

In the analysis of this work, we have to deal with the L^1 estimate of piecewisely continuous functions, some of which involve derivatives of some other piecewisely continuous functions; see, e.g., (7.10). In order to avoid confusion, we define the L^1-norm for a piecewisely smooth function q by

$$\|q(\cdot)\|_{\mathrm{pis}(\mathbf{R})} = \sum_{i=1}^{I+1} \|q(\cdot)\|_{L^1(Y_{i-1}, Y_i)},$$

where Y_i, with $Y_0 := -\infty$ and $Y_{I+1} := \infty$, are all the possible discontinuous points of $q(x)$. The proof of the following stability lemma can be found in [24].

LEMMA 3.1. *Let u^ϵ be the viscous solution of* (1.3)–(1.4). *Let $v^\epsilon \in C(\mathbf{R}\times[0,T])$ be a piecewisely smooth function with jumps in the derivative in the set $\mathcal{A} = \{(x,t) \mid x = Y_i(t),\ 1 \le i \le I\}$. If v^ϵ satisfies*

$$\partial_t v^\epsilon + \partial_x f(v^\epsilon) = \epsilon \partial_{xx} v^\epsilon + g(x,t) \tag{3.2}$$

everywhere except on the set $\mathcal{A}$, then for any $0 \le \tau < t \le T$

$$\begin{aligned}\|u^\epsilon(\cdot,t) - v^\epsilon(\cdot,t)\|_{L^1(\mathbf{R})} &\le \|u^\epsilon(\cdot,\tau) - v^\epsilon(\cdot,\tau)\|_{L^1(\mathbf{R})} \\ &+ \epsilon \sum_{i=1}^{I} \int_\tau^t \Big| \left[\partial_x v^\epsilon(x,t)\right] |_{x=Y_i(t)} \Big| dt + \int_\tau^t \|g(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})} dt,\end{aligned}$$

where the jumps are defined by

$$[w(x,t)]|_{x=Y(t)} := w(Y(t)+0,t) - w(Y(t)-0,t).$$

Remark 3.1. It will be seen in sections 7 and 8 that $g(x,t)$ in (3.2) may involve some derivatives of a discontinuous function, so we use the norm $\| \bullet \|_{\mathrm{pis}(\mathbf{R})}$ to define its L^1-norm.

It follows from Theorem 3.1 that both first-order and fractional-order rates of convergence may occur for nonconvex conservation laws, which is in contrast with that for the convex conservation laws. We will demonstrate this fact with the following examples.

Example 3.1. Let $f(u) = u^{2m+1}$. If the entropy solution $u(x,t)$ is defined by (2.7), then it follows from Example 2.1 and Theorem 3.1 that

$$\|u^\epsilon(\cdot,t) - u(\cdot,t)\|_{L^1(\mathbf{R})} \le C(T)\epsilon|\ln\epsilon|. \tag{3.3}$$

Example 3.2. Let $f(u) = (1-u)^p(1+u)^q$, with $s := \max(p,q) \ge 1$. If the entropy solution $u(x,t)$ is defined by (2.8), then it follows from Example 2.2 and Theorem 3.1 that

$$\|u^\epsilon(\cdot,t) - u(\cdot,t)\|_{L^1(\mathbf{R})} \le \begin{cases} C(T)\epsilon|\ln\epsilon| & \text{for} \quad 1 \le s \le 2, \\ C(T)\epsilon^{\frac{s}{2(s-1)}}|\ln\epsilon| & \text{for} \quad s > 2. \end{cases} \tag{3.4}$$

4. Traveling wave solution of viscous equation. Our construction of an approximation solution is based on some detailed properties of viscous shock profiles for (1.3), whose nonlinear asymptotic stability was studied by Matsumura and Nishihara [15]. We will summarize some of their results in this section, which will be used in our error analysis. Some results not obtained in [15] can be derived by using the techniques developed in [24]. Let

$$u^\epsilon(x,t) = V^\epsilon(x - \sigma t; u_+, u_-), \tag{4.1}$$

which is subject to the boundary conditions

$$V^\epsilon(\xi; u_+, u_-) \to u_\pm \quad \text{as} \quad \xi \to \pm\infty.$$

If $V^\epsilon(x-\sigma t; u_+, u_-)$ satisfies (1.3), then it is called a *traveling wave solution* of (1.3). Applying the solution from (4.1) to (1.3) gives

$$\epsilon V^\epsilon_{\xi\xi} = -\sigma V^\epsilon_\xi + f(V^\epsilon)_\xi. \tag{4.2}$$

Integrating the above equation over $(-\infty, \xi)$ gives

$$\epsilon V^\epsilon_\xi = -\sigma(V^\epsilon - u_-) + f(V^\epsilon) - f(u_-). \tag{4.3}$$

It is easy to show by rescaling $\eta = \xi/\epsilon$ that $V^\epsilon(\xi; u_+, u_-) = V^1(\xi/\epsilon; u_+, u_-)$. In the following we will use the notation $V(\eta; u_+, u_-)$ to denote $V^1(\eta; u_+, u_-)$. We also denote by $V(\eta; a, b)_a$ and $V(\eta; a, b)_b$ the partial derivatives of V with respect to a and b, respectively. Note that $V^\epsilon(\xi; u_+, u_-) = V(\xi/\epsilon; u_+, u_-)$, which satisfies

$$V' = -\sigma(V - u_-) + f(V) - f(u_-). \tag{4.4}$$

It is well known that a necessary and sufficient condition for the existence of a traveling wave solution is that the constants $u_\pm$ and σ satisfy the Rankine–Hugoniot condition

$$-\sigma(u_+ - u_-) + f(u_+) - f(u_-) = 0 \tag{4.5}$$

and the entropy condition

$$\Phi(u; u_+, u_-) =: -\sigma(u - u_\pm) + f(u) - f(u_\pm) \begin{cases} < 0 & \text{if } u_+ < u < u_-, \\ > 0 & \text{if } u_- < u < u_+. \end{cases} \tag{4.6}$$

LEMMA 4.1. *Let $u_\pm$ and σ satisfy* (4.5)–(4.6) *and*

$$|\Phi(u; u_+, u_-)| \sim |u - u_\pm|^{1+\beta_\pm} \quad \text{as} \quad u \to u_\pm \tag{4.7}$$

with $\beta_\pm \geq 0$. Then there exists $V(\eta; u_+, u_-)$, unique up to a shift, which is determined by the ordinary differential equation (4.4). *Moreover, for $k = 1, 2$*

1. *if $\beta_\pm = 0$, then $f'(u_+) < \sigma < f'(u_-)$ and for $\eta \to \pm\infty$*

$$\begin{aligned} &|V(\eta; u_+, u_-) - H(\eta; u_+, u_-)| \sim \exp(-c|\eta|), \\ &|V^{(k)}(\eta; u_+, u_-)| \sim \exp(-c|\eta|), \\ &|V(\eta; u_+, u_-)_{u_\pm} - 1| \sim \exp(-c|\eta|), \end{aligned} \tag{4.8}$$

   where $H(\eta; u_+, u_-)$ is the Heaviside function satisfying $H = u_+$ for $\eta > 0$ and $H = u_-$ for $\eta < 0$;

2. *if $\beta_+ > 0$, then $\sigma = f'(u_+)$ and for $\eta \to +\infty$*

$$\begin{aligned} &|V(\eta; u_+, u_-) - u_+| \sim |\eta|^{-1/\beta_+}, \\ &|V^{(k)}(\eta; u_+, u_-)| \sim |\eta|^{-1/\beta_+ - k}, \\ &|V(\eta; u_+, u_-)_{u_+} - 1| \sim |\eta|^{-1/\beta_+}; \end{aligned} \tag{4.9}$$

3. *if $\beta_- > 0$, then $\sigma = f'(u_-)$ and for $\eta \to -\infty$*

$$\begin{aligned} &|V(\eta; u_+, u_-) - u_-| \sim |\eta|^{-1/\beta_-}, \\ &|V^{(k)}(\eta; u_+, u_-)| \sim |\eta|^{-1/\beta_- - k}, \\ &|V(\eta; u_+, u_-)_{u_-} - 1| \sim |\eta|^{-1/\beta_-}. \end{aligned} \tag{4.10}$$

Proof. For completeness, we briefly outline the proof for this lemma. It follows from (4.4) that $V(\eta; u_+, u_-)$ can be defined implicitly by

$$\eta = \int_{(u_+ + u_-)/2}^{V} \Phi(v; u_+, u_-)^{-1} dv. \tag{4.11}$$

The proof of this lemma is mainly based on the above definition and the assumption (4.7). Here we show only some of the estimates in (4.9); other estimates can be obtained similarly. The assumption (4.7) implies

$$D^{-1}|v-u_-|^{1+\beta_-}|v-u_+|^{1+\beta_+} \le |\Phi(v;u_+,u_-)| \le D|v-u_-|^{1+\beta_-}|v-u_+|^{1+\beta_+},$$

where $D>0$ is a constant. It follows from the above inequalities and (4.11) that, for $\eta>0$,

$$\begin{aligned} D^{-1}2^{-(1+\beta_-)}\left|\frac{2}{u_- - u_+}\right|^{(1+\beta_-)}\left|\int_{(u_++u_-)/2}^{V}(v-u_+)^{-1-\beta_+}dv\right| \\ \le \eta \le D\left|\frac{2}{u_- - u_+}\right|^{(1+\beta_-)}\left|\int_{(u_++u_-)/2}^{V}(v-u_+)^{-1-\beta_+}dv\right|. \end{aligned}$$

Solving the above inequalities for V gives

$$\begin{aligned} &\left|\frac{u_+-u_-}{2}\right|\left(1+D2^{(1+\beta_-)}\left|\frac{u_+-u_-}{2}\right|^{1+\beta_++\beta_-}\beta_+\eta\right)^{-1/\beta_+} \\ &\quad \le |V-u_+| \le \left|\frac{u_+-u_-}{2}\right|\left(1+D^{-1}\left|\frac{u_+-u_-}{2}\right|^{1+\beta_++\beta_-}\beta_+\eta\right)^{-1/\beta_+}. \end{aligned}$$

This proves the first estimate in (4.9). It is easy to show that as $\eta\to+\infty$

$$\begin{aligned} |V'(\eta;u_+,u_-)| = |\Phi(V;u_+,u_-)| &\sim |V-u_+|^{1+\beta_+} \\ &\sim |\eta|^{-1/\beta_+(1+\beta_+)} = |\eta|^{-1/\beta_+-1}. \end{aligned}$$

Thus the second estimate in (4.9) follows. □

COROLLARY 4.1. *If* (4.7) *holds and* $\beta_\pm>0$, *then*

$$|a(u)-a(u_\pm)| \sim |u-u_\pm|^{\beta_\pm} \quad as \quad u\to u_\pm. \tag{4.12}$$

COROLLARY 4.2. *Under the same assumptions as in Lemma* 4.1, *the following results hold:*

1. *If* $\beta_\pm=0$, *then* $\forall\eta\in\mathbf{R}$ *and* $k=1,2$

$$\begin{aligned} &|V(\eta;u_+,u_-)-H(\eta;u_+,u_-)| \le C\exp(-c|\eta|), \\ &|V^{(k)}(\eta;u_+,u_-)| \le C\exp(-c|\eta|). \end{aligned} \tag{4.13}$$

2. *If* $\beta_+>0$, *then* $\forall\eta\in\mathbf{R}^+$ *and* $k=1,2$

$$\begin{aligned} &|V(\eta;u_+,u_-)-u_+| \le C(1+|\eta|)^{-1/\beta_+}, \\ &|V^{(k)}(\eta;u_+,u_-)| \le C(1+|\eta|)^{-1/\beta_+-k}. \end{aligned} \tag{4.14}$$

3. *If* $\beta_->0$, *then* $\forall\eta\in\mathbf{R}^-$ *and* $k=1,2$

$$\begin{aligned} &|V(\eta;u_+,u_-)-u_-| \le C(1+|\eta|)^{-1/\beta_-}, \\ &|V^{(k)}(\eta;u_+,u_-)| \le C(1+|\eta|)^{-1/\beta_--k}. \end{aligned} \tag{4.15}$$

4. *If* $\beta_\pm > 0$, *then* $\forall \eta \in \mathbf{R}$

$$\begin{aligned} &\left|V(\eta; u_+, u_-)_{u_+}\dot{u}_+ + V(\eta; u_+, u_-)_{u_-}\dot{u}_- - H(\eta; \dot{u}_+, \dot{u}_-)\right| \\ &\le C(1+|\eta|)^{-1/\beta}, \end{aligned} \tag{4.16}$$

where $\beta = \max\{\beta_-, \beta_+\}$ *and* $\dot{w} = w'(t)$.

Remark 4.1. It is noted that the constant C in the above inequalities depends on $|u_+ - u_-|^{-1}$. Since it is assumed that $u_k^+(t) - u_k^-(t) \neq 0$ on all $[t_{m-1}, t_m]$, C can be regarded as a constant uniform with respect to both $t \in [0, T]$ and $k = 1, \ldots, K$.

5. Construction of an approximate solution. In this section, we construct an approximate solution $\widehat{u}^\epsilon$ to u and u^ϵ by using the method of matching asymptotic expansions. As in [5] and [24], the main idea of constructing $\widehat{u}^\epsilon$ is that $\widehat{u}^\epsilon$ is a small perturbation of u in the smooth region that posseses a viscous shock profile in places of discontinuities. We begin with the simpler case of one single shock.

5.1. An approximation to u and u^ϵ with one shock. Assume that there is only one shock curve $x = X_1(t)$ in the entropy solution $u(x,t)$ in the time interval $[t_{m-1}, t_m]$. We construct a *continuous* approximate solution $\widehat{u}^\epsilon$ to u and u^ϵ in $[t_{m-1}, t_m]$,

$$\widehat{u}^\epsilon(x,t) = m\left(\frac{x - X_1(t)}{\epsilon^\gamma}\right) I(x,t) + \left(1 - m\left(\frac{x - X_1(t)}{\epsilon^\gamma}\right)\right) O(x,t), \tag{5.1}$$

where

$$\begin{aligned} I(x,t) = u(x,t) &+ V\left(\frac{x - X_1(t)}{\epsilon}; u_+(t), u_-(t)\right) \\ &- H\left(\frac{x - X_1(t)}{\epsilon}; u_+(t), u_-(t)\right), \end{aligned} \tag{5.2}$$

$$O(x,t) = u(x,t) \tag{5.3}$$

are called first-order *inner* and *outer* solutions, respectively, $u_\pm(t) = u(X_1(t) \pm 0, t)$, $H(\xi; u_+, u_-)$ is the Heaviside function, $0 < \gamma < 1$ is a constant to be determined later, $m(\xi) \in C^\infty(\mathbf{R})$ satisfying $0 \le m(\xi) \le 1$, and

$$m(\xi) = \begin{cases} 1, & |\xi| \le 1, \\ 0, & |\xi| \ge 2. \end{cases} \tag{5.4}$$

The approximate solution can also be written in an equivalent form:

$$\widehat{u}^\epsilon(x,t) = u(x,t) + m\left(\frac{x - X_1(t)}{\epsilon^\gamma}\right)(V - H)\left(\frac{x - X_1(t)}{\epsilon}; u_+(t), u_-(t)\right). \tag{5.5}$$

The following lemma shows that $\widehat{u}^\epsilon$ is a good approximation to u in the L^1 space.

LEMMA 5.1. *Assume that there is only one shock curve* $x = X_1(t)$ *for the entropy solution* $u(x,t)$ *in the time interval* $[t_{m-1}, t_m]$. *Then, for any* $t \in [t_{m-1}, t_m]$,

$$\|\widehat{u}^\epsilon(\cdot,t) - u(\cdot,t)\|_{L^1(\mathbf{R})} \le \begin{cases} C\epsilon, & \bar{\beta} < 1, \\ C\epsilon^{\gamma + (1-\gamma)/\bar{\beta}}\,|\ln \epsilon|, & \bar{\beta} \ge 1, \end{cases} \tag{5.6}$$

where $\bar{\beta} = \max_t \beta(t)$ *and* $\beta = \max\{\beta_1^+, \beta_1^-\}$.

Proof. It follows from (5.5) and (5.4) that

$$\begin{aligned}
&\|\widehat{u}^\epsilon(\cdot,t)-u(\cdot,t)\|_{L^1(\mathbf{R})}\\
&=\int_{-2\epsilon^\gamma}^{0}+\int_0^{2\epsilon^\gamma} m\left(\frac{x}{\epsilon^\gamma};u_+(t),u_-(t)\right)\left|(V-H)\left(\frac{x}{\epsilon};u_+(t),u_-(t)\right)\right|dx\\
&=: I_-+I_+.
\end{aligned}$$

Using the change of variables $\xi = x/\epsilon$ and the estimate (4.14), we have for $0<\beta_1^+\le 1$ that

$$\begin{aligned}
I_+ &= \epsilon\int_0^{2\epsilon^{-1+\gamma}} m(\epsilon^{1-\gamma}\xi)|V(\xi;u_+,u_-)-H(\xi;u_+,u_-)|d\xi \qquad (5.7)\\
&= C\epsilon\int_0^{2\epsilon^{-1+\gamma}}(1+|\xi|)^{-1/\beta_1^+}d\xi \le \begin{cases} C\epsilon, & 0<\beta_1^+<1,\\ C\epsilon|\ln\epsilon|, & \beta_1^+=1.\end{cases}
\end{aligned}$$

On the other hand, for $\beta_1^+>1$ using the change of variables $\xi^{\beta_1^+}=x/\epsilon$ gives

$$\begin{aligned}
I_+ &= \beta_1^+\epsilon\int_0^{2^{1/\beta_1^+}\epsilon^{-(1-\gamma)/\beta_1^+}} m\Big(\epsilon^{1-\gamma}\xi^{\beta_1^+}\Big)\Big|(V-H)\Big(\xi^{\beta_1^+};u_+,u_-\Big)\Big|\xi^{\beta_1^+-1}d\xi\\
&\le C\epsilon^{1-(1-\gamma)(\beta_1^+-1)/\beta_1^+}\int_0^{2^{1/\beta_1^+}\epsilon^{-(1-\gamma)/\beta_1^+}}(1+|\xi|^{\beta_1^+})^{-1/\beta_1^+}d\xi\\
&\le C\epsilon^{1-(1-\gamma)(\beta_1^+-1)/\beta_1^+}|\ln\epsilon|. \qquad (5.8)
\end{aligned}$$

It follows from the above results that

$$I_+\le\begin{cases} C\epsilon, & 0<\beta_1^+<1,\\ C\epsilon|\ln\epsilon|, & \beta_1^+=1,\\ C\epsilon^{1-(1-\gamma)(\beta_1^+-1)/\beta_1^+}|\ln\epsilon|, & 1<\beta_1^+.\end{cases} \qquad (5.9)$$

Similarly, we can obtain the estimates for I_-:

$$I_-\le\begin{cases} C\epsilon, & 0<\beta_1^-<1,\\ C\epsilon|\ln\epsilon|, & \beta_1^-=1,\\ C\epsilon^{1-(1-\gamma)(\beta_1^--1)/\beta_1^-}|\ln\epsilon|, & 1<\beta_1^-.\end{cases} \qquad (5.10)$$

Combining the estimates for I_+ and I_- gives the desired result (5.6). ∎

We can also estimate the difference between $\widehat{u}^\epsilon$ and u^ϵ. The result will be given below, but its proof will be deferred to section 7.

LEMMA 5.2. *Assume that there is only one shock curve $x=X_1(t)$ for the entropy solution $u(x,t)$ in the time interval $[t_{m-1},t_m]$. Then, for any $t\in[t_{m-1},t_m]$,*

$$\begin{aligned}
&\|\widehat{u}^\epsilon(\cdot,t)-u^\epsilon(\cdot,t)\|_{L^1(\mathbf{R})} \qquad (5.11)\\
&\le \|\widehat{u}^\epsilon(\cdot,t_{m-1})-u^\epsilon(\cdot,t_{m-1})\|_{L^1(\mathbf{R})}+C\epsilon^{(1-\gamma)(\bar\beta+1)/\bar\beta}+C\epsilon^{\gamma+(1-\gamma)/\bar\beta}.
\end{aligned}$$

5.2. An approximation to u and u^ϵ with two shocks. Assume that in the time interval $[t_{m-1},t_m]$ there exist two shock curves $x=X_1(t)$ and $x=X_2(t)$ for the entropy solution $u(x,t)$ which either collide at $t=t_m$, i.e., $X_1(t_m)=X_2(t_m)$, or at $t=t_{m-1}$, i.e., $X_1(t_{m-1})=X_2(t_{m-1})$. We construct a *continuous* approximate

solution $\widehat{u}^\epsilon$ to u and u^ϵ in $[t_{m-1}, t_m]$ by using the method of matching asymptotic expansions:

$$\widehat{u}^\epsilon(x,t) = m\left(\frac{x-X_1(t)}{\epsilon^\gamma}\right) I_1(x,t) + \left(1 - m\left(\frac{x-X_1(t)}{\epsilon^\gamma}\right)\right) O_1(x,t) + m\left(\frac{x-X_2(t)}{\epsilon^\gamma}\right) I_2(x,t) + \left(1 - m\left(\frac{x-X_2(t)}{\epsilon^\gamma}\right)\right) O_2(x,t), \tag{5.12}$$

where

$$I_i(x,t) = u(x,t) + (V-H)\left(\frac{x-X_i(t)}{\epsilon}; u_{i+}(t), u_{i-}(t)\right), \qquad i = 1,2, \tag{5.13}$$

$$O_i(x,t) = u(x,t), \qquad i = 1,2, \tag{5.14}$$

are the first-order inner and outer solutions, respectively. Here, $u_{i\pm}(t) = u(X_i(t) \pm 0, t)$, $H(\xi; u_+, u_-)$ is the Heaviside function, γ is a constant to be determined later. This approximation can be also written in an equivalent form:

$$\begin{aligned}\widehat{u}^\epsilon(x,t) = u(x,t) &+ m\left(\frac{x-X_1(t)}{\epsilon^\gamma}\right)(V-H)\left(\frac{x-X_1(t)}{\epsilon}; u_{1+}(t), u_{1-}(t)\right) \\ &+ m\left(\frac{x-X_2(t)}{\epsilon^\gamma}\right)(V-H)\left(\frac{x-X_2(t)}{\epsilon}; u_{2+}(t), u_{2-}(t)\right).\end{aligned} \tag{5.15}$$

LEMMA 5.3. *Assume that in the time interval $[t_{m-1}, t_m]$ there exist two shock curves $x = X_1(t)$ and $x = X_2(t)$ for the entropy solution $u(x,t)$ which either collide at $t = t_m$ or at $t = t_{m-1}$. Then, for any $t \in [t_{m-1}, t_m]$,*

$$\|\widehat{u}^\epsilon(\cdot,t) - u(\cdot,t)\|_{L^1(\mathbf{R})} \le \begin{cases} C\epsilon, & \bar\beta < 1, \\ C\epsilon^{\gamma+(1-\gamma)/\bar\beta}\,|\ln\epsilon|, & \bar\beta \ge 1, \end{cases} \tag{5.16}$$

where $\bar\beta = \max_t \beta(t)$ and $\beta = \max\{\beta_1^+, \beta_1^-, \beta_2^+, \beta_2^-\}$.

LEMMA 5.4. *Assume that in the time interval $[t_{m-1}, t_m]$ there exist two shock curves $x = X_1(t)$ and $x = X_2(t)$ for the entropy solution $u(x,t)$ which either collide at $t = t_m$ or at $t = t_{m-1}$. Then, for any $t \in [t_{m-1}, t_m]$,*

$$\begin{aligned}&\|\widehat{u}^\epsilon(\cdot,t) - u^\epsilon(\cdot,t)\|_{L^1(\mathbf{R})} \\ &\le \|\widehat{u}^\epsilon(\cdot,t_{m-1}) - u^\epsilon(\cdot,t_{m-1})\|_{L^1(\mathbf{R})} + C\epsilon^{(1-\gamma)(\bar\beta+1)/\bar\beta} + C\epsilon^{\gamma+(1-\gamma)/\bar\beta} + C\epsilon^{2\gamma}.\end{aligned} \tag{5.17}$$

The proof of Lemma 5.3 is similar to that of Lemma 5.1 and will be omitted here. We defer the proof of Lemma 5.4 to section 8.

6. Proof of main theorem. We will prove Theorem 3.1 by considering only the case $\bar\beta \ge 1$, i.e., the nonconvex case; the convex result was obtained in [24]. Recall that it is assumed in each time interval $[t_{m-1}, t_m]$ the entropy solution u is a finite combination of some noncontact shocks, contact shocks, etc. Theorem 3.1 will be established by induction on m. Namely, we will prove

$$\|u(\cdot,t_m) - u^\epsilon(\cdot,t_m)\|_{L^1(\mathbf{R})} \le \begin{cases} C(T)\epsilon|\ln\epsilon| & \text{for } \bar\beta < 1, \\ C(T)\epsilon^{(1+1/\bar\beta)/2}|\ln\epsilon| & \text{for } \bar\beta \ge 1 \end{cases} \tag{6.1}$$

under the induction assumption

$$\|u(\cdot,t_{m-1}) - u^\epsilon(\cdot,t_{m-1})\|_{L^1(\mathbf{R})} \le \begin{cases} C(T)\epsilon|\ln\epsilon| & \text{for } \bar\beta < 1, \\ C(T)\epsilon^{(1+1/\bar\beta)/2}|\ln\epsilon| & \text{for } \bar\beta \ge 1. \end{cases} \tag{6.2}$$

The induction assumption holds for $m = 1$ due to the fact $u(x,0) = u^\epsilon(x,t)$. Observe that

$$\begin{aligned}(6.3)\qquad &\|u(\cdot,t_m) - u^\epsilon(\cdot,t_m)\|_{L^1(\mathbf{R})}\\ &\le \|u(\cdot,t_m) - \hat{v}^\epsilon(\cdot,t_m)\|_{L^1(\mathbf{R})} + \|\hat{v}^\epsilon(\cdot,t_m) - u^\epsilon(\cdot,t_m)\|_{L^1(\mathbf{R})}.\end{aligned}$$

It follows from Lemmas 5.1 and 5.3 that

$$(6.4)\qquad \|u(\cdot,t_m) - \hat{v}^\epsilon(\cdot,t_m)\|_{L^1(\mathbf{R})} \le C\epsilon^{\gamma+(1-\gamma)/\bar{\beta}}|\ln\epsilon|.$$

On the other hand, Lemmas 5.2 and 5.4 imply that

$$\begin{aligned}(6.5)\qquad &\|\hat{v}^\epsilon(\cdot,t_m) - u^\epsilon(\cdot,t_m)\|_{L^1(\mathbf{R})}\\ &\le \|\hat{v}^\epsilon(\cdot,t_{m-1}) - u^\epsilon(\cdot,t_{m-1})\|_{L^1(\mathbf{R})} + C\epsilon^{(1-\gamma)(\bar{\beta}+1)/\bar{\beta}} + C\epsilon^{\gamma+(1-\gamma)/\bar{\beta}} + C\epsilon^{2\gamma}.\end{aligned}$$

Using the induction assumption (6.2) and (6.3)–(6.5) gives

$$\begin{aligned}&\|u(\cdot,t_m) - u^\epsilon(\cdot,t_m)\|_{L^1(\mathbf{R})}\\ &\le C\epsilon^{\gamma+(1-\gamma)/\bar{\beta}}|\ln\epsilon| + C\epsilon^{(1+1/\bar{\beta})/2}|\ln\epsilon| + C\epsilon^{(1-\gamma)(\bar{\beta}+1)/\bar{\beta}} + C\epsilon^{\gamma+(1-\gamma)/\bar{\beta}} + C\epsilon^{2\gamma}.\end{aligned}$$

Setting $\gamma = 1/2$ in the above estimate leads to (6.1), which completes the induction proof.

7. Proof of Lemma 5.2. The main tool for establishing Lemma 5.2 is the stability lemma, Lemma 3.1. Let $v^\epsilon = \widehat{u}^\epsilon$ as defined by (5.1). Then v^ϵ satisfies (3.2) on its smooth region $\{(x,t) : x \ne X_1(t)\}$, with

$$\begin{aligned}g(x,t) &= u_t + (m(V-H))_t + f(u + m(V-H))_x - \epsilon(u_{xx} + (m(V-H))_{xx})\\ &= -a(u)u_x + m_t(V-H) - mV'\frac{\dot{X}_1(t)}{\epsilon} + m(V-H)_{u_+}\dot{u}_+ + m(V-H)_{u_-}\dot{u}_-\\ &\quad + (a(u+m(V-H))\left(u_x + m_x(V-H) + mV'\frac{1}{\epsilon}\right)\\ &\quad - \epsilon\left(u_{xx} + m_{xx}(V-H) + 2m_xV'\frac{1}{\epsilon} + mV''\frac{1}{\epsilon^2}\right).\end{aligned}$$

It follows from Lemma 3.1 that, for any $t \in [t_{m-1}, t_m]$,

$$\begin{aligned}(7.1)\qquad &\|\widehat{u}^\epsilon(\cdot,t) - u^\epsilon(\cdot,t)\|_{L^1(\mathbf{R})}\\ &\le \|\widehat{u}^\epsilon(\cdot,t_{m-1}) - u^\epsilon(\cdot,t_{m-1})\|_{L^1(\mathbf{R})}\\ &\quad + \epsilon\int_{t_{m-1}}^t \Big|\left[\partial_x\widehat{u}^\epsilon(x,t)\right]|_{x=X_1(t)}\Big|dt + \int_{t_{m-1}}^t \|g(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})}dt\\ &\le \|\widehat{u}^\epsilon(\cdot,t_{m-1}) - u^\epsilon(\cdot,t_{m-1})\|_{L^1(\mathbf{R})} + C\epsilon + \int_{t_{m-1}}^t \|g(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})}dt,\end{aligned}$$

where we have used the facts that $[\partial_x\widehat{u}^\epsilon(x,t)]\,|_{x=X_1(t)} = [u_x(x,t)]\,|_{x=X_1(t)}$ and the limits of $u_x(X(t)\pm 0,t)$ are uniformly bounded on $[t_{m-1},t_m]$. We now claim that g is

sufficiently small such that $\widehat{u}^\epsilon$ satisfies (1.3) approximately. To this end, we rewrite g as follows:

$$\begin{aligned} g(x,t) &= m\left(\big(a(u+m(V-H))-\dot{X}_1(t)\big)V'-V''\right)\epsilon^{-1} \\ &\quad + \big(a(u+m(V-H))-a(u)\big)u_x \\ &\quad + \left(\big(a(u+m(V-H))-\dot{X}_1(t)\big)m'-m''\epsilon^{1-\gamma}\right)\epsilon^{-\gamma}(V-H) \\ &\quad - 2m'\epsilon^{-\gamma}V'-\epsilon u_{xx}+m(V-H)_{u_+}\dot{u}_+ + m(V-H)_{u_-}\dot{u}_- \\ &= \big(a(u+m(V-H))-\dot{X}_1(t)\big)\left(mV'\epsilon^{-1}+u_x+m'(V-H)\epsilon^{-\gamma}\right) \\ &\quad - mV''\epsilon^{-1}+(\dot{X}_1(t)-a(u))u_x - m''(V-H)\epsilon^{1-2\gamma}-2m'\epsilon^{-\gamma}V'-\epsilon u_{xx} \\ &\quad + m\left(V_{u_+}\dot{u}_+ + V_{u_-}\dot{u}_- - H\big(x-X_1(t);\dot{u}_+(t),\dot{u}_-(t)\big)\right). \end{aligned}$$

Using the traveling wave equation $V''=(a(V)-\dot{X}_1(t))V'$ gives

$$\begin{aligned} g(x,t) &= \big(a(u+m(V-H))-a(V)\big)mV'\epsilon^{-1} \\ &\quad + \big(a(u+m(V-H))-\dot{X}_1(t)\big)\left(u_x+m'(V-H)\epsilon^{-\gamma}\right) \\ &\quad + (\dot{X}_1(t)-a(u))u_x - m''(V-H)\epsilon^{1-2\gamma}-2m'\epsilon^{-\gamma}V'-\epsilon u_{xx} \\ &\quad + m\left(V_{u_+}\dot{u}_+ + V_{u_-}\dot{u}_- - H\big(x-X_1(t);\dot{u}_+(t),\dot{u}_-(t)\big)\right). \end{aligned} \tag{7.2}$$

Without loss of generality, we will consider only contact shock curves, i.e., $\beta_1^\pm \geq 1$. The assumption $\beta_1^\pm \geq 1$ implies that $\dot{X}_1(t)=a(u_\pm)$, which leads to

$$\begin{aligned} g(x,t) &= \big(a(u+m(V-H))-a(V)\big)mV'\epsilon^{-1} \\ &\quad + \big(a(u+m(V-H))-a(u_\pm)\big)\left(u_x+m'(V-H)\epsilon^{-\gamma}\right) \\ &\quad + (a(u_\pm)-a(u))u_x - m''(V-H)\epsilon^{1-2\gamma}-2m'\epsilon^{-\gamma}V'-\epsilon u_{xx} \\ &\quad + m\left(V_{u_+}\dot{u}_+ + V_{u_-}\dot{u}_- - H\big(x-X_1(t);\dot{u}_+(t),\dot{u}_-(t)\big)\right). \end{aligned} \tag{7.3}$$

We now split $\|g(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})}$ into the following three parts:

$$\begin{aligned} \|g(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})} &= \int_{0<|x-X_1(t)|\leq\epsilon^\gamma} + \int_{\epsilon^\gamma\leq|x-X_1(t)|\leq 2\epsilon^\gamma} + \int_{|x-X_1(t)|\geq 2\epsilon^\gamma} |g(x,t)|dt \\ &= I + II + III. \end{aligned} \tag{7.4}$$

7.1. Piecewise constant solution. In order to estimate I, II, and III above we first consider a simple but important case: the piecewise constant solution, i.e.,

$$u(x,t)=\begin{cases} u_+, & x>X_1(t), \\ u_-, & x\leq X_1(t), \end{cases} \tag{7.5}$$

where u_+ and u_- are constants and $\dot{X}_1(t)=(f(u_-)-f(u_+))/(u_--u_+)$. It is easy to show that $g(x,t)=0$ for $0<|x-X_1(t)|\leq\epsilon^\gamma$ and $|x-X_1(t)|>2\epsilon^\gamma$. Therefore $I=III=0$, and what we need to estimate is the term II. Let

$$II = II_+ + II_-, \tag{7.6}$$

where

$$
\begin{aligned}
II_\pm =: \pm \int_{X_1(t)\pm\epsilon^\gamma}^{X_1(t)\pm 2\epsilon^\gamma} \Big| & \big(a(u_\pm + m(V-u_\pm)) - a(V)\big) mV'\epsilon^{-1} \\
& + \big(a(u_\pm + m(V-u_\pm)) - a(u_\pm)\big) m'(V-u_\pm)\epsilon^{-\gamma} \\
& - m''(V-u_\pm)\epsilon^{1-2\gamma} - 2m'\epsilon^{-\gamma}V' \Big| dx \qquad (7.7)\\
= \pm \int_{X_1(t)\pm\epsilon^\gamma}^{X_1(t)\pm 2\epsilon^\gamma} \Big| & \big(a(u_\pm + m(V-u_\pm)) - a(u_\pm)\big)\big(mV'\epsilon^{-1} + m'(V-u_\pm)\epsilon^{-\gamma}\big) \\
& - (a(V) - a(u_\pm))mV'\epsilon^{-1} - m''(V-u_\pm)\epsilon^{1-2\gamma} - 2m'\epsilon^{-\gamma}V' \Big| dx.
\end{aligned}
$$

Using the change of variables $\xi = (x - X_1(t))/\epsilon^\gamma$ gives

$$
\begin{aligned}
II_\pm = \pm \int_{\pm 1}^{\pm 2} \Big| & \big(a(u_\pm + m(V-u_\pm)) - a(u_\pm)\big)\big(mV'\epsilon^{-1+\gamma} + m'(V-u_\pm)\big) \\
& - (a(V) - a(u_\pm))mV'\epsilon^{-1+\gamma} - m''(V-u_\pm)\epsilon^{1-\gamma} - 2m'V' \Big| d\xi, \qquad (7.8)
\end{aligned}
$$

where $m = m(\xi)$ and $V = V(\xi/\epsilon^{1-\gamma})$. The following estimates can be obtained from Lemma 4.1 and Corollary 4.1:

$$
\begin{aligned}
& |V - u_\pm| \le C\epsilon^{(1-\gamma)/\beta_1^\pm}, \qquad |V'| \le C\epsilon^{(1-\gamma)(1+1/\beta_1^\pm)}, \\
& |a(u_\pm + m(V-u_\pm)) - a(u_\pm)| \le C|V-u_\pm|^{\beta_1^\pm} \le C\epsilon^{1-\gamma}, \\
& |a(V) - a(u_\pm)| \le C|V-u_\pm|^{\beta_1^\pm} \le C\epsilon^{1-\gamma},
\end{aligned}
$$

provided ϵ is sufficiently small. It follows from (7.8) and the above estimates that

$$
II_\pm \le C\epsilon^{(1-\gamma)(1+1/\beta_1^\pm)}. \qquad (7.9)
$$

The above results, together with the facts $I = III = 0$, give the desired upper bound for $\|g(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})}$. Therefore, Lemma 5.2 is established in the case of the piecewise constant solution.

7.2. Piecewise smooth solution. We now consider a more general case, i.e., u is piecewise smooth. It follows from (7.3) that $g(x,t) = -\epsilon u_{xx}$ for $|x - X_1(t)| > 2\epsilon^\gamma$, and

$$
\begin{aligned}
g(x,t) = & \big(a(u + (V-H)) - a(V)\big)V'\epsilon^{-1} \\
& + \big(a(u + (V-H)) - a(u)\big)u_x - \epsilon u_{xx} \\
& + V_{u_+}\dot{u}_+ + V_{u_-}\dot{u}_- - H(x - X_1(t); \dot{u}_+, \dot{u}_-)
\end{aligned} \qquad (7.10)
$$

for $0 < |x - X_1(t)| \le \epsilon^\gamma$. It is easy to see from (7.4) that

$$
III = \epsilon \int_{|x-X_1(t)|\ge 2\epsilon^\gamma} |u_{xx}| dx \le C\epsilon, \qquad (7.11)
$$

where $u_{xx}(\cdot,t)$ is assumed piecewisely in L^1. It follows from (7.10) that

$$
I \le I_1 + I_2 + I_3 + I_4, \qquad (7.12)
$$

where

$$
\begin{aligned}
I_1 &= \int_{0<|x-X_1(t)|\le\epsilon^\gamma} -|a(u+(V-H))-a(V)|V'\epsilon^{-1}dx,\\
I_2 &= \int_{0<|x-X_1(t)|\le\epsilon^\gamma} \big|\big(a(u+(V-H))-a(u)\big)u_x\big|dx,\\
I_3 &= \int_{0<|x-X_1(t)|\le\epsilon^\gamma} \epsilon|u_{xx}|dx,\\
I_4 &= \int_{0<|x-X_1(t)|\le\epsilon^\gamma} \big|V_{u_+}\dot u_+ + V_{u_-}\dot u_- - H(x-X_1(t);\dot u_+,\dot u_-)\big|dx.
\end{aligned}
$$

We now estimate $I_i, 1\le i\le 4$. Since u is piecewise smooth and $u-H\to 0$ as $x\to X_1(t)\pm 0$, we have

$$
\begin{aligned}
|a(u+(V-H))-a(V)| &= |a'(V+\theta(u-H))(u-H)|\\
&\le C|x-X_1(t)|.
\end{aligned}
$$

Therefore, we can find a positive function $A(\xi)$ such that $\forall x\in(-\infty,\infty)$

- $|a(u+(V-H))-a(V)|\le A\left(x-X_1(t)\right)$;
- $A(\xi)\le C|\xi|$;
- $|A'(\xi)|\le M$,

where M is a constant. It follows from the above auxiliary function A and the estimates (4.9) and (4.10) that

$$
\begin{aligned}
I_1 &\le \int_{|x-X_1(t)|\le\epsilon^\gamma} -A(x-X_1(t))V'((x-X_1(t))/\epsilon)\epsilon^{-1}dx\\
&= \int_{-\epsilon^{\gamma-1}}^{\epsilon^{\gamma-1}} -A(\epsilon\xi)V'(\xi)d\xi\\
&= \int_{-\epsilon^{\gamma-1}}^{\epsilon^{\gamma-1}} -A(\epsilon\xi)\big(V(\xi)-H(\xi)\big)'d\xi \qquad \text{(using the fact } A(0)=0)\\
&\le A(\epsilon^\gamma)|V(\epsilon^{\gamma-1})-H(\epsilon^{\gamma-1})| + A(-\epsilon^\gamma)|V(-\epsilon^{\gamma-1})-H(-\epsilon^{\gamma-1})|\\
&\qquad +M\epsilon\int_{-\epsilon^{\gamma-1}}^{\epsilon^{\gamma-1}}|V(\xi)-H(\xi)|d\xi \qquad \text{(using integration by parts)}\\
&\le C\epsilon^{\gamma+(1-\gamma)/\beta} + C\epsilon^{1-(1-\gamma)(1-1/\beta)} \le \epsilon^{\gamma+(1-\gamma)/\beta}.
\end{aligned}
\tag{7.13}
$$

Observe that $|a(u+(V-H))-a(u)|\le C|V-H|$, which, if applied to I_2, gives

$$
\begin{aligned}
I_2 &\le C\int_{|x-X_1(t)|\le\epsilon^\gamma}|V-H|dx\\
&\le C\epsilon\int_{-\epsilon^{\gamma-1}}^{\epsilon^{\gamma-1}}|V(\xi)-H(\xi)|d\xi\\
&\le C\epsilon\int_{-\epsilon^{\gamma-1}}^{\epsilon^{\gamma-1}}(1+|\xi|)^{-1/\beta}d\xi\\
&\le C\epsilon^{1-(1-\gamma)(1-1/\beta)} = C\epsilon^{\gamma+(1-\gamma)/\beta}.
\end{aligned}
\tag{7.14}
$$

Moreover, by the definition of I_3 we can easily obtain $I_3 \le C\epsilon^{1+\gamma}$. Using the change of variables $\xi = (x - X_1(t))/\epsilon$ and the estimate (4.16) gives

$$I_4 \le C\epsilon \int_{-\epsilon^{\gamma-1}}^{\epsilon^{\gamma-1}} (1+|\xi|)^{-1/\beta} d\xi \le C\epsilon^{1-(1-\gamma)(1-1/\beta)} = C\epsilon^{\gamma+(1-\gamma)/\beta}. \tag{7.15}$$

Combining the above estimates yields

$$I \le C\left(\epsilon^{\gamma+(1-\gamma)/\beta} + \epsilon^{\gamma(1+\beta)} + \epsilon|\ln\epsilon| + \epsilon^{(1+\gamma)}\right). \tag{7.16}$$

It remains to estimate II. It follows from (7.3) that

$$II \le II_+ + II_- + \sum_{i=1}^{5} II_+^{(i)} + \sum_{i=1}^{5} II_-^{(i)},$$

where $II_\pm$ is defined by (7.7) and

$$\begin{aligned}
II_\pm^{(1)} &= \pm \int_{X_1(t)\pm\epsilon^\gamma}^{X_1(t)\pm 2\epsilon^\gamma} |(a(u + m(V-H)) - a(u_\pm + m(V-H)))mV'|\epsilon^{-1} dx,\\
II_\pm^{(2)} &= \pm \int_{X_1(t)\pm\epsilon^\gamma}^{X_1(t)\pm 2\epsilon^\gamma} |(a(u + m(V-H)) - a(u))u_x| dx,\\
II_\pm^{(3)} &= \pm \int_{X_1(t)\pm\epsilon^\gamma}^{X_1(t)\pm 2\epsilon^\gamma} |(a(u + m(V-H)) - a(u_\pm + m(V-H)))m'(V-H)|\epsilon^{-\gamma} dx,\\
II_\pm^{(4)} &= \pm \int_{X_1(t)\pm\epsilon^\gamma}^{X_1(t)\pm 2\epsilon^\gamma} \epsilon|u_{xx}| dx,\\
II_\pm^{(5)} &= \pm \int_{X_1(t)\pm\epsilon^\gamma}^{X_1(t)\pm 2\epsilon^\gamma} m\big|V_{u_+}\dot{u}_+ + V_{u_-}\dot{u}_- - H(x - X_1(t); \dot{u}_+, \dot{u}_-)\big| dx.
\end{aligned}$$

It follows from (7.9) that

$$II_\pm \le C\epsilon^{(1-\gamma)(1+1/\beta_1^\pm)}. \tag{7.17}$$

The estimate for $II_\pm^{(1)}$ is similar to that for I_1, with the same error bound as (7.13), namely,

$$II_\pm^{(1)} \le C\epsilon^{\gamma+(1-\gamma)/\beta}. \tag{7.18}$$

The estimate for $II_\pm^{(2)}$ is similar to that for I_2, with the same error bound as (7.14), namely,

$$II_\pm^{(2)} \le C\left(\epsilon^{\gamma(1+\beta)} + \epsilon|\ln\epsilon|\right). \tag{7.19}$$

The estimate for $II_\pm^{(5)}$ is similar to that for I_4, with the same error bound as (7.15), namely,

$$II_\pm^{(5)} \le C\epsilon^{1-(1-\gamma)(1-1/\beta)} = C\epsilon^{\gamma+(1-\gamma)/\beta}. \tag{7.20}$$

Using the facts that $a(u+m(V-H))-a(u_\pm+m(V-H))=a'(\bullet)(u-u_\pm)$, and $|u-u_\pm|\le C|x-X_1(t)|$ for $\pm(x-X_1(t))>0$, we obtain

$$I\!I_\pm^{(3)} \le \pm C\epsilon^\gamma \int_{X_1(t)\pm\epsilon^\gamma}^{X_1(t)\pm 2\epsilon^\gamma} |(V-H)|\epsilon^{-\gamma}dx.$$

Applying the change of variables $(x-X_1(t))/\epsilon=\xi$ to the above integration gives

$$I\!I_\pm^{(3)} \le \pm C\epsilon \int_{\pm\epsilon^{\gamma-1}}^{\pm 2\epsilon^{\gamma-1}} |V(\xi)-H(\xi)|d\xi.$$

It then follows from (4.14) and (4.15) that

$$\begin{aligned} I\!I_\pm^{(3)} &\le \pm C\epsilon \int_{\pm\epsilon^{\gamma-1}}^{\pm 2\epsilon^{\gamma-1}} (1+|\xi|)^{-1/\beta_1^\pm} d\xi \le C\epsilon^{1-(1-\gamma)(1-1/\beta_1^\pm)} \\ &= C\epsilon^{\gamma+(1-\gamma)/\beta_1^\pm}. \end{aligned} \tag{7.21}$$

Moreover, using the definition of $I\!I_\pm^{(4)}$ gives

$$I\!I_\pm^{(4)} \le \pm C \int_{X_1(t)\pm\epsilon^\gamma}^{X_1(t)\pm 2\epsilon^\gamma} \epsilon|u_{xx}|dx \le C\epsilon^{1+\gamma}. \tag{7.22}$$

Combining the estimates (7.17)–(7.22) leads to

$$\begin{aligned} I\!I &\le C\Big(\epsilon^{(1-\gamma)(1+1/\beta)} + \epsilon^{\gamma+(1-\gamma)/\beta)} + \epsilon^{\gamma(1+\beta)} + \epsilon|\ln\epsilon| + \epsilon^{1+\gamma}\Big) \\ &\le C\Big(\epsilon^{(1-\gamma)(1+1/\beta)} + \epsilon^{\gamma+(1-\gamma)/\beta)}\Big). \end{aligned} \tag{7.23}$$

Adding the estimates for I, $I\!I$, and $I\!I\!I$ gives

$$\begin{aligned} \|g(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})} &\le C\left(\epsilon^{\gamma+(1-\gamma)/\beta} + \epsilon^{(1-\gamma)(1+1/\beta)}\right) \\ &\le C\left(\epsilon^{\gamma+(1-\gamma)/\bar\beta} + \epsilon^{(1-\gamma)(1+1/\bar\beta)}\right). \end{aligned} \tag{7.24}$$

This completes the proof for Lemma 5.2.

8. Proof of Lemma 5.4. The main difference between Lemmas 5.2 and 5.4 is that Lemma 5.2 deals with only one shock, while the latter deals with two interacting shocks. The main tool for the proof of Lemma 5.4 is still the stability lemma, Lemma 3.1. Let $v^\epsilon=\widehat{u}^\epsilon$, which is defined by (5.15). Then v^ϵ satisfies (3.2) on its smooth region $\{(x,t): x\ne X_i(t),\ i=1,2\}$, with

$$\begin{aligned} g(x,t) \;=\; & \left(a(u+m_1(V_1-H_1)+m_2(V_2-H_2))-a(u)\right)u_x \\ & + m_1'\left(a(u+m_1(V_1-H_1)+m_2(V_2-H_2))-a(H_1)\right)(V_1-H_1)\epsilon^{-\gamma} \\ & + m_2'(a(u+m_1(V_1-H_1)+m_2(V_2-H_2))-a(H_2))(V_2-H_2)\epsilon^{-\gamma} \\ & + m_1(a(u+m_1(V_1-H_1)+m_2(V_2-H_2))-a(V_1))V_1'\epsilon^{-1} \\ & + m_2(a(u+m_1(V_1-H_1)+m_2(V_2-H_2))-a(V_2))V_2'\epsilon^{-1} \\ & - \epsilon u_{xx} - \left(m_1''(V_1-H_1)+m_2''(V_2-H_2)\right)\epsilon^{1-2\gamma} - 2\left(m_1'V_1'+m_2'V_2'\right)\epsilon^{-\gamma} \\ & + m_1(V_{1u_+}\dot u_{1+}+V_{1u_-}\dot u_{1-}-\dot H_1)+m_2(V_{2u_+}\dot u_{2+}+V_{2u_-}\dot u_{2-}-\dot H_2) \\ :=\; & g_1+g_2+g_3+g_4+g_5+g_6+g_7, \end{aligned}$$

where for $i = 1, 2$

$$m_i = m((x - X_i(t))/\epsilon^\gamma), \qquad u_{i\pm}(t) = \lim_{x \to X_i \pm 0} u(x,t),$$
$$\dot{u}_{i\pm}(t) = \frac{d}{dt} u_{i\pm}(t), \qquad V_i = V((x - X_i(t))/\epsilon; u_{i+}, u_{i-}),$$
$$H_i = H((x - X_i(t))/\epsilon; u_{i+}, u_{i-}), \qquad \dot{H}_i = H((x - X_i(t))/\epsilon; \dot{u}_{i+}, \dot{u}_{i-}).$$

Similarly, it follows from Lemma 3.1 that, for any $t \in [t_{m-1}, t_m]$,

$$\begin{aligned} &\|\widehat{u}^\epsilon(\cdot,t) - u^\epsilon(\cdot,t)\|_{L^1(\mathbf{R})} \\ &\le \|\widehat{u}^\epsilon(\cdot,t_{m-1}) - u^\epsilon(\cdot,t_{m-1})\|_{L^1(\mathbf{R})} \\ &\quad + \epsilon \sum_{i=1}^{2} \int_{t_{m-1}}^{t} \Big| \left[\partial_x \widehat{u}^\epsilon(x,t)\right] |_{x=X_i(t)} \Big| dt + \int_{t_{m-1}}^{t} \|g(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})} dt \\ &\le \|\widehat{u}^\epsilon(\cdot,t_{m-1}) - u^\epsilon(\cdot,t_{m-1})\|_{L^1(\mathbf{R})} + C\epsilon + \int_{t_{m-1}}^{t} \|g(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})} dt. \end{aligned} \tag{8.1}$$

In order to estimate the last term above we need to estimate $\int_{t_{m-1}}^{t} \|g_i(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})} dt$, $1 \le i \le 7$. It is noticed that the estimates for $\|g_6\|_{\mathrm{pis}(\mathbf{R})}$ and $\|g_7\|_{\mathrm{pis}(\mathbf{R})}$ are similar to those in one shock case, so they can be bounded above by the right-hand side of (7.24). Therefore, we just need to estimate $\|g_i\|_{\mathrm{pis}(\mathbf{R})}$ for $1 \le i \le 5$.

In what follows we assume that the two shock curves start at a same point, i.e.,

$$X_1(t_{m-1} + 0) = X_2(t_{m-1} + 0), \tag{8.2}$$

but do not become tangent to each other at this point:

$$\dot{X}_1(t_{m-1} + 0) < \dot{X}_2(t_{m-1} + 0). \tag{8.3}$$

This implies that there is a constant $c > 0$ such that

$$\delta(t) := X_2(t) - X_1(t) \ge c(t - t_{m-1}) \quad \text{for} \quad t \in [t_{m-1}, t_m]. \tag{8.4}$$

Note that the support of $m((\cdot - X_i(t))/\epsilon^\gamma)$ is $[X_i(t) - 2\epsilon^\gamma, X_i(t) + 2\epsilon^\gamma]$. Hence, if $X_2(t) - X_1(t) \ge 4\epsilon^\gamma$, then for each $x \in \mathbf{R}$ only one of m_1 and m_2 appears in $g(x,t)$. Therefore, when $X_2(t) - X_1(t) \ge 4\epsilon^\gamma$, the estimates for $\|g_i\|_{L^1}$, $1 \le i \le 5$, are similar to those for the one shock case. Let $\tau \in (t_{m-1}, t_m)$ such that

$$X_2(\tau) - X_1(\tau) = 4\epsilon^\gamma \quad \text{and} \quad X_2(t) - X_1(t) > 4\epsilon^\gamma \quad \text{for} \quad t \in (\tau, t_m]. \tag{8.5}$$

The above analysis implies that we only need to estimate $\|g_i\|_{L^1}$, $1 \le i \le 5$, for $t \in [t_{m-1}, \tau]$.

We first estimate $\|g_1\|_{L^1}$ for $t \in [t_{m-1}, \tau]$. Observe that

$$\begin{aligned} \|g_1(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})} &= \int_{\{x \le X_1(t) - 2\epsilon^\gamma\} \cup \{x \ge X_2(t) + 2\epsilon^\gamma\}} |g_1| dx \\ &\quad + \int_{X_1(t) - 2\epsilon^\gamma < x < X_2(t) + 2\epsilon^\gamma} |g_1| dx \\ &:= G_{11} + G_{12}. \end{aligned} \tag{8.6}$$

Since only one of m_1 and m_2 appears in G_{11}, the estimate of G_{11} is similar to that of $II_{\pm}^{(2)}$ in the one shock case, and it follows from (7.19) that

$$G_{11} \le C\Big(\epsilon^{\gamma(1+\beta)} + \epsilon|\ln\epsilon|\Big), \tag{8.7}$$

where $\beta = \max\{\beta_1, \beta_2\}$. The integrand of G_{12} is bounded, and therefore

$$G_{12} \le C\Big(X_2(t) - X_1(t) + \epsilon^{\gamma}\Big).$$

It follows from (8.5), (8.4), and (8.2) that $\tau - t_{m-1} \le C\epsilon^{\gamma}$ and $X_2(t) - X_1(t) \le C(t - t_{m-1})$ as $t \to t_{m-1} + 0$. Therefore,

$$\int_{t_{m-1}}^{\tau} G_{12}dt \le C\epsilon^{2\gamma},$$

which, together with (8.7), gives

$$\int_{t_{m-1}}^{\tau} \|g_1(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})}dt \le C\Big(\epsilon^{2\gamma} + \epsilon^{\gamma(2+\beta)} + \epsilon^{1+\gamma}|\ln\epsilon|\Big). \tag{8.8}$$

Next we estimate $\|g_2\|_{\mathrm{pis}(\mathbf{R})}$. Observe that

$$\begin{aligned}\|g_2\|_{\mathrm{pis}(\mathbf{R})} &= \int_{2\epsilon^{\gamma}>|x-X_1(t)|\ge\epsilon^{\gamma}} |g_2(x,t)|dx \le C\int_{2\epsilon^{\gamma}>|x-X_1(t)|\ge\epsilon^{\gamma}} |V_1 - H_1|\epsilon^{-\gamma}dx \\ &= \int_{2>|\xi|>1} |V_1(\xi/\epsilon^{1-\gamma}) - H_1(\xi/\epsilon^{1-\gamma})|d\xi \le C\epsilon^{(1-\gamma)/\beta_1}.\end{aligned}$$

Therefore,

$$\int_{t_{m-1}}^{\tau} \|g_2(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})}dt \le C\epsilon^{(1-\gamma)/\beta_1+\gamma}. \tag{8.9}$$

Similarly, it can be shown that

$$\int_{t_{m-1}}^{\tau} \|g_3(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})}dt \le C\epsilon^{(1-\gamma)/\beta_2+\gamma}. \tag{8.10}$$

We now estimate $\|g_4\|_{\mathrm{pis}(\mathbf{R})}$. Note that

$$\begin{aligned}\|g_4(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})} &= \int_{0<|x-X_1(t)|<\epsilon^{\gamma}} + \int_{|x-X_1(t)|\ge\epsilon^{\gamma}} |g_4(x,t)|dx \\ &:= G_{41} + G_{42}.\end{aligned} \tag{8.11}$$

The integrand in G_{41} satisfies the following inequality:

$$\begin{aligned}|g_4| \text{ in } G_{41} &= |a(u + (V_1 - H_1) + m_2(V_2 - H_2)) - a(V_1)|(-V_1')\epsilon^{-1} \\ &\le |a(u + (V_1 - H_1) + m_2(V_2 - H_2)) - a(V_1 + m_2(V_2 - H_2))|(-V_1')\epsilon^{-1} \\ &\quad + |a(V_1 + m_2(V_2 - H_2)) - a(V_1)|(-V_1')\epsilon^{-1} \\ &:= |g_{41}| + |g_{42}|.\end{aligned}$$

The integration of $|g_{41}|$ is similar to that of (7.13), so it follows from (7.13) that

$$\int_{\{0<|x-X_1|\le\epsilon^\gamma\}} |g_{41}|dx \le C\epsilon^{\gamma+(1-\gamma)/\beta_1}. \tag{8.12}$$

We estimate the integral of g_{42} by observing

$$\begin{aligned}
&\int_{\{0<|x-X_1|\le\epsilon^\gamma\}} |g_{42}|dx \\
&= \int_{\{|x-X_1|\le\epsilon^\gamma\}\cap\{|x-X_2|\le\delta(t)/2\}} |g_{42}|dx + \int_{\{|x-X_1|\le\epsilon^\gamma\}\cap\{|x-X_2|>\delta(t)/2\}} |g_{42}|dx \\
&:= G_{411} + G_{412},
\end{aligned}$$

where $\delta(t) := X_2(t) - X_1(t)$. Since $|g_{42}| \le -CV_1'\epsilon^{-1}$, we have

$$\begin{aligned}
G_{411} &\le C\epsilon^{-1}\int_{|x-X_2(t)|\le\delta(t)/2} -V_1'((x-X_1(t))/\epsilon)dx \\
&\le C\epsilon^{-1}\left(\frac{2\epsilon}{\delta(t)}\right)^{1/\beta_1+1}\delta(t) \le C\left(\frac{\epsilon}{\delta(t)}\right)^{1/\beta_1},
\end{aligned}$$

which leads to

$$\begin{aligned}
\int_{t_{m-1}}^{\tau} G_{411}dt &\le C\epsilon^{1/\max_t\beta_1}(t-t_{m-1})^{1-1/\max_t\beta_1}\Big|_{t_{m-1}}^{\tau} \\
&= C\epsilon^{1/\max_t\beta_1}(\tau-t_{m-1})^{1-1/\max_t\beta_1},
\end{aligned}$$

where we have applied the inequality (8.4) to the above integration. Since $\tau - t_{m-1} \le C\epsilon^\gamma$ we have

$$\int_{t_{m-1}}^{\tau} G_{411}dt \le C\epsilon^{\gamma+(1-\gamma)/\max_t\beta_1}. \tag{8.13}$$

On the other hand,

$$|g_{42}| \le C|V_2 - H_2|(-V_1')\epsilon^{-1}$$

by the mean value theorem. Substituting this inequality into G_{412} yields

$$\begin{aligned}
G_{412} &= \int_{\{|x-X_1|\le\epsilon^\gamma\}\cap\{|x-X_2|>\delta(t)/2\}} |g_{42}|dx \\
&\le C\left|V_2\left(\frac{\delta(t)}{2\epsilon}\right) - H_2\left(\frac{\delta(t)}{2\epsilon}\right)\right|\int_{-\infty}^{\infty}(-V_1'((x-X_1)/\epsilon))\epsilon^{-1}dx \\
&\le C\left(\frac{2\epsilon}{\delta(t)}\right)^{1/\beta_2}\int_{-\infty}^{\infty} -V_1'(\xi)d\xi \le C\left(\frac{2\epsilon}{\delta(t)}\right)^{1/\beta_2}.
\end{aligned}$$

Therefore, on account of $\delta(t) \ge c(t-t_{m-1})$ and $\tau - t_{m-1} \le C\epsilon^\gamma$, we obtain

$$\int_{t_{m-1}}^{\tau} G_{412}dt \le C\epsilon^{1/\max_t\beta_2}\epsilon^{\gamma(1-1/\max_t\beta_2)} = C\epsilon^{\gamma+(1-\gamma)/\max_t\beta_2}. \tag{8.14}$$

This result, together with (8.13) and (8.12), yields

$$\int_{t_{m-1}}^{\tau} G_{41}dt \le C\epsilon^{\gamma+(1-\gamma)/\bar{\beta}}, \tag{8.15}$$

where $\bar{\beta} = \max\{\max_t \beta_1, \max_t \beta_2\}$. In order to estimate G_{42} in (8.11), we first split it into two parts:

$$\begin{aligned} G_{42} &= \int_{\{\epsilon^\gamma<|x-X_1|\le 2\epsilon^\gamma\}\cap\{|x-X_2|\le\delta(t)/2\}} \\ &\qquad + \int_{\{\epsilon^\gamma<|x-X_1|\le 2\epsilon^\gamma\}\cap\{|x-X_2|>\delta(t)/2\}} |g_4|dx \\ &:= G_{421} + G_{422}. \end{aligned} \tag{8.16}$$

The integrand $|g_4|$ in G_{422} can be estimated as

$$\begin{aligned} |g_4| \text{ in } G_{422} &= m_1|a(u + m_1(V_1 - H_1) + m_2(V_2 - H_2)) - a(V_1)|(-V_1')\epsilon^{-1} \\ &\le m_1|a(u + m_1(V_1 - H_1) + m_2(V_2 - H_2)) \\ &\qquad - a(H_1 + m_1(V_1 - H_1) + m_2(V_2 - H_2))|(-V_1')\epsilon^{-1} \\ &\quad + m_1\big(|a(H_1 + m_1(V_1 - H_1) + m_2(V_2 - H_2)) - a(H_1)| \\ &\qquad + |a(H_1) - a(H_1 + V_1 - H_1)|\big)(-V_1')\epsilon^{-1} \\ &\le C\Big(|u - H_1| + |V_1 - H_1|^{\beta_1} + |V_2 - H_2|^{\beta_1}\Big)(-V_1')\epsilon^{-1}, \end{aligned}$$

where we have used the facts that $|V_1 - H_1| \to 0$ and $|V_2 - H_2| \to 0$ as $\epsilon \to 0$ for $\{\epsilon^\gamma < |x - X_1| \le 2\epsilon^\gamma\} \cap \{|x - X_2| > \delta(t)/2\}$. Therefore,

$$\begin{aligned} G_{422} &\le C\int_{\{\epsilon^\gamma<|x-X_1|\le 2\epsilon^\gamma\}\cap\{|x-X_2|>\delta(t)/2\}} |u - H_1|(-V_1')\epsilon^{-1}dx \\ &\quad + C\int_{\{\epsilon^\gamma<|x-X_1|\le 2\epsilon^\gamma\}\cap\{|x-X_2|>\delta(t)/2\}} |V_1 - H_1|^{\beta_1}(-V_1')\epsilon^{-1}dx \\ &\quad + C\int_{\{\epsilon^\gamma<|x-X_1|\le 2\epsilon^\gamma\}\cap\{|x-X_2|>\delta(t)/2\}} |V_2 - H_2|^{\beta_1}(-V_1')\epsilon^{-1}dx \\ &=: J_1 + J_2 + J_3. \end{aligned}$$

The estimate for J_1 is similar to that of I_1, with an upper bound the same as (7.13), namely,

$$J_1 \le C\epsilon^{\gamma+(1-\gamma)/\beta_1}. \tag{8.17}$$

It follows from (4.9) and (4.10) that

$$J_2 \le C\epsilon^{1-\gamma}\epsilon^{(1-\gamma)(1+1/\beta_1)}\epsilon^{-1}\epsilon^{\gamma} \le C\epsilon^{(1-\gamma)(1+1/\beta_1)}. \tag{8.18}$$

Observe that

$$\begin{aligned} J_3 &= C\int_{\{|x-X_1|\le 2\epsilon^\gamma\}\cap\{|x-X_2|>\delta(t)/2\}} |V_2-H_2|^{\beta_1}(-V_1')\epsilon^{-1}dx \\ &\le C\left|V_2\left(\frac{\delta(t)}{2\epsilon}\right)-H_2\left(\frac{\delta(t)}{2\epsilon}\right)\right|^{\beta_1}\int_{-\infty}^{\infty}(-V_1'((x-X_1)/\epsilon))\epsilon^{-1}dx \\ &\le C\left(\frac{2\epsilon}{\delta(t)}\right)^{\beta_1/\beta_2}\int_{-\infty}^{\infty}-V_1'(\xi)d\xi \\ &\le C\left(\frac{2\epsilon}{\delta(t)}\right)^{\beta_1/\beta_2}\le C\left(\frac{2\epsilon}{\delta(t)}\right)^{1/\beta_2}. \end{aligned}$$

Therefore, on account of $\delta(t)\ge c(t-t_{m-1})$ and $\tau-t_{m-1}\le C\epsilon^\gamma$, we have

$$\int_{t_{m-1}}^{\tau} J_3dt \le C\epsilon^{1/\max_t\beta_2}\epsilon^{\gamma(1-1/\max_t\beta_2)} = C\epsilon^{\gamma+(1-\gamma)/\max_t\beta_2}. \tag{8.19}$$

This, together with (8.17) and (8.18), yields

$$\int_{t_{m-1}}^{\tau} G_{422}dt \le C\Big(\epsilon^{\gamma+(1-\gamma)/\bar\beta}+\epsilon^{(1-\gamma)(1+1/\bar\beta)}\Big), \tag{8.20}$$

where $\bar\beta=\max\{\max_t\beta_1,\max_t\beta_2\}$. Since $|g_4|\le -CV_1'\epsilon^{-1}$, we have

$$\begin{aligned} G_{421} &\le C\epsilon^{-1}\int_{|x-X_2(t)|\le\delta(t)/2}-V_1'((x-X_1(t))/\epsilon)dx \\ &\le C\epsilon^{-1}\left(\frac{2\epsilon}{\delta(t)}\right)^{1/\beta_1+1}\delta(t)\le C\left(\frac{\epsilon}{\delta(t)}\right)^{1/\beta_1}. \end{aligned}$$

Thus

$$\begin{aligned} \int_{t_{m-1}}^{\tau} G_{421}dt &\le C\epsilon^{1/\max_t\beta_1}(t-t_{m-1})^{1-1/\max_t\beta_1}\Big|_{t_{m-1}}^{\tau} \\ &= C\epsilon^{1/\max_t\beta_1}(\tau-t_{m-1})^{1-1/\max_t\beta_1}, \end{aligned}$$

where we have used the inequality (8.4). Since $\tau-t_{m-1}\le C\epsilon^\gamma$, we have

$$\int_{t_{m-1}}^{\tau} G_{421}dt \le C\epsilon^{\gamma+(1-\gamma)/\max_t\beta_1}.$$

This result, together with (8.20), gives

$$\int_{t_{m-1}}^{\tau} G_{42}dt \le C\Big(\epsilon^{\gamma+(1-\gamma)/\bar\beta}+\epsilon^{(1-\gamma)(1+1/\bar\beta)}\Big).$$

Combining the above result and (8.15) gives

$$\int_{t_{m-1}}^{\tau}\|g_4(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})}dt \le C\Big(\epsilon^{\gamma+(1-\gamma)/\bar\beta}+\epsilon^{(1-\gamma)(1+1/\bar\beta)}\Big). \tag{8.21}$$

Similarly, it can be shown that

$$\int_{t_{m-1}}^{\tau}\|g_5(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})}dt \le C\Big(\epsilon^{\gamma+(1-\gamma)/\bar\beta}+\epsilon^{(1-\gamma)(1+1/\bar\beta)}\Big). \tag{8.22}$$

Therefore, we have proved that

$$\int_{t_{m-1}}^{\tau} \|g(\cdot,t)\|_{\mathrm{pis}(\mathbf{R})} dt \le C\Big(\epsilon^{\gamma+(1-\gamma)/\bar{\beta}} + \epsilon^{(1-\gamma)(1+1/\bar{\beta})}\Big). \tag{8.23}$$

This result and (8.1) yield Lemma 5.4.

9. Numerical experiments. To verify the theoretical results obtained in this work, we shall carry out a computational study in this section. The main purpose is to demonstrate the existence of the fractional rate of convergence. It is generally believed that monotone schemes have the same rate of convergence as that for the viscosity approximation. Therefore, to make the numerical verification available, we consider the (generalized) Lax–Friedrichs scheme

$$u_j^{n+1} = u_j^n - \frac{\lambda}{2}\left(f(u_{j+1}^n) - f(u_{j-1}^n)\right) + \frac{\mu}{2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) \tag{9.1}$$

to approximate the conservation law (1.1), where u_j^n is an approximation of $u(x_j, t_n)$, and $x_j = j\Delta x$, and $t_n = n\Delta t$, with Δx and Δt being the spatial and temporal grid sizes, respectively; μ is a constant satisfying $0 < \mu < 1$, and the temporal and spatial grid ratio $\lambda = \Delta t/\Delta x$ satisfies a Courant–Friedrichs–Levy condition,

$$\lambda \sup_{|u|\le\|u_0\|_\infty} |f'(u)| \le \mu.$$

The theoretical properties of the scheme (9.1) were investigated by Liu and Xin [14].

Example 9.1. In the first example, we approximate

$$\partial_t u + \partial_x f(u) = 0, \qquad f(u) = (1-u^2)^3,$$

with the initial data $u_0(x) = \mathrm{sgn}(x)$, by using the Lax–Friedrichs scheme (9.1).

The entropy solution for the above Riemann problem is $u(x,t) = u_0(x)$. It can be verified that $\max_{|u|\le 1} |f'(u)| \le 6/\sqrt{5}$. We then choose $\mu = 0.5$, $T = 1$, and $\lambda = \sqrt{5}\mu/6$. It follows from Theorem 3.1 that the rate of convergence should be $(1+\frac{1}{2})/2 = \frac{3}{4}$. It is observed from Table 1 that the numerical rate of convergence agrees very well with the theoretical prediction.

TABLE 1
The L^1-error and the convergence order for Example 9.1.

Mesh Δx	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
L^1-error	3.22×10^{-1}	2.01×10^{-1}	1.23×10^{-1}	7.44×10^{-2}	4.46×10^{-2}
Order		0.680	0.709	0.725	0.738

Example 9.2. In the second example, we approximate

$$\partial_t u + \partial_x f(u) = 0, \qquad f(u) = (1-u)^3(1+u)^4,$$

with the initial data $u_0(x) = \mathrm{sgn}(x)$, by using the Lax–Friedrichs scheme (9.1).

The entropy solution for the above Riemann problem is again $u(x,t) = u_0(x)$. It can be verified that $\max_{|u|\le 1} |f'(u)| \le 2$. We then choose $\mu = 0.5$, $T = 1$, and $\lambda = \mu/2$. It follows from Theorem 3.1 that the rate of convergence should be $(1+\frac{1}{3})/2 = \frac{2}{3}$. It is observed from Table 2 that the numerical result is again in excellent agreement with the theoretical prediction.

TABLE 2
The L^1-error and the convergence order for Example 9.2.

Mesh Δx	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
L^1-error	2.79×10^{-1}	1.89×10^{-1}	1.25×10^{-1}	8.05×10^{-2}	5.11×10^{-2}
Order		0.562	0.597	0.635	0.656

REFERENCES

[1] D. P. BALLOU, *Solutions to nonlinear hyperbolic Cauchy problems without convexity conditions*, Trans. Amer. Math. Soc., 152 (1970), pp. 441–460.

[2] C. M. DAFERMOS, *Generalized characteristics and the structure of solutions of hyperbolic conservation laws*, Indiana Univ. Math. J., 26 (1977), pp. 1097–1119.

[3] C. M. DAFERMOS, *Regularity and large time behavior of solutions of a conservation law without convexity*, Proc. Roy. Soc. Edinburgh Sect. A, 99 (1985), pp. 201–239.

[4] J. GOODMAN, *Nonlinear asymptotic stability of viscous shock profiles for conservation laws*, Arch. Ration. Mech. Anal., 95 (1986), pp. 325–344.

[5] J. GOODMAN AND Z. XIN, *Viscous limits for piecewise smooth solutions to systems of conservation laws*, Arch. Ration. Mech. Anal., 121 (1992), pp. 235–265.

[6] P. HOWARD AND K. ZUMBRUN, *Pointwise semigroup methods and stability of viscous shock waves*, Indiana Univ. J. Math., 47 (1998), pp. 741–871.

[7] S. N. KRUSHKOV, *First-order quasilinear equations with several space variables*, Mat. Sb., 123 (1970), pp. 228–255.

[8] N. N. KUZNETSOV, *Accuracy of some approximate methods for computing the weak solutions of a first-order quasi-linear equation*, U.S.S.R. Comput. Math. and Math. Phys., 16 (1976), pp. 105–119.

[9] G. Kreiss and H.-O. Kreiss, *Stability of systems of viscous conservation laws*, Comm. Pure Appl. Math., 51 (1998), pp. 1397–1424.

[10] P. D. LAX, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[11] T. P. LIU, *Admissible Solutions of Hyperbolic Conservation Laws*, Mem. Amer. Math. Soc. 240, AMS, Providence, RI, 1981.

[12] T. P. LIU, *Nonlinear Stability of Shock Waves for Viscous Conservation Laws*, Mem. Amer. Math. Soc. 328, AMS, Providence, RI, 1985.

[13] T.-P. LIU, *Pointwise convergence to shock waves for viscous conservation laws*, Comm. Pure Appl. Math., 50 (1997), pp. 1113–1182.

[14] J. LIU AND Z. XIN, *L^1-stability of stationary discrete shocks*, Math. Comp., 60 (1993), pp. 233–244.

[15] A. MATSUMURA AND K. NISHIHARA, *Asymptotic stability of traveling waves for scalar viscous conservation laws with non-convex nonlinearity*, Comm. Math. Phys., 165 (1994), pp. 83–96.

[16] A. MATSUMURA AND K. NISHIHARA, *On the stability of traveling waves of a one-dimensional model system for compressible viscous gas*, Japan J. Appl. Math., 2 (1985), pp. 17–25.

[17] K. NISHIHARA, *Stability of traveling waves with degenerate shock for systems of one-dimensional viscoelastic model*, J. Differential Equations, 120 (1995), pp. 304–318.

[18] O. A. OLEINIK, *Discontinuous solutions of non-linear differential equations*, Amer. Math. Soc. Transl., 26 (1963), pp. 95–172.

[19] F. SABAC, *The optimal convergence rate of monotone finite difference methods for hyperbolic conservation laws*, SIAM J. Numer. Anal., 34 (1997), pp. 2306–2318.

[20] J. A. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1984.

[21] A. SZEPESSY AND Z. P. XIN, *Nonlinear stability of viscous shock waves*, Arch. Ration. Mech. Anal., 122 (1993), pp. 53–103.

[22] E. TADMOR AND T. TANG, *Pointwise error estimates for scalar conservation laws with piecewise smooth solutions*, SIAM J. Numer. Anal., 36 (1999), pp. 1739–1758.

[23] E. TADMOR AND T. TANG, *Pointwise error estimates for relaxation approximations to conservation laws*, SIAM J. Math. Anal., 32 (2000), pp. 870–886.

[24] T. TANG AND Z. H. TENG, *Viscosity methods for piecewise smooth solutions to scalar conservation laws*, Math. Comp., 66 (1997), pp. 495–526.

[25] T. Tang and Z. H. Teng, *The sharpness of Kuznetsov's $O(\sqrt{\Delta x})$ L^1-error estimate for monotone difference schemes*, Math. Comp., 64 (1995), pp. 581–589.
[26] Z. H. Teng, *First-order L^1-convergence for relaxation approximations to conservation laws*, Comm. Pure Appl. Math., 51 (1998), pp. 857–895.
[27] Z. H. Teng and P. Zhang, *Optimal L^1-rate of convergence for the viscosity method and monotone scheme to piecewise constant solutions with shocks*, SIAM J. Numer. Anal., 34 (1997), pp. 959–978.
[28] W.-C. Wang, *On L^1 convergence rate of viscous and numerical approximate solutions of genuinely nonlinear scalar conservation laws*, SIAM J. Math. Anal., 30 (1998), pp. 38–52.
[29] Z. P. Xin, *Theory of viscous conservation laws*, in Some Current Topics on Nonlinear Conservation Laws, L. Hsiao and Z. P. Xin, eds., AMS/IP Stud. Adv. Math. 15, AMS, Providence, RI, 2000, pp. 141–194.
[30] Z. P. Xin, *Viscous boundary layers and their stability* (I), J. Partial Differential Equations, 11 (1998), pp. 97–124.
[31] S.-H. Yu, *Zero dissipation limit to solutions with shocks for systems of hyperbolic conservation laws*, Arch. Ration. Mech. Anal., 146 (1999), pp. 275–370.
[32] K. Zumbrun, *Asymptotic Behavior for Systems of Nonconvex Conservation Laws*, Ph.D. Dissertation, New York University, New York, 1990.

SIAM J. MATH. ANAL.
Vol. 35, No. 1, pp. 123–134

LONG TIME BEHAVIOR OF SOLUTIONS TO P-LAPLACIAN EQUATION WITH ABSORPTION*

XINFU CHEN†, Y. W. QI‡, AND MINGXIN WANG§

Abstract. In this paper, we study the long time behavior of solutions to the Cauchy problem of $u_t = \mathrm{div}(|\nabla u|^{p-2}\nabla u) - u^q$ in $R^n \times (0,\infty)$, with nonnegative initial value $u(x,0) = \phi(x)$ in R^n, where $(2n)/(n+1) < p < 2$ and $q > 1$. For initial data of various decay rates, especially the critical decay $\phi = O(|x|^{-\mu})$ with $\mu = p/(q+1-p)$, we show that the solution converges as $t \to \infty$ to a self-similar solution. This extends the recent result of Escobedo, Kavian, and Matano for the semilinear case of $p = 2$. Here an essential role is played by singular and very singular self-similar solutions established in our previous works [X. Chen, Y. Qi, and M. Wang, *J. Differential Equations*, 190 (2003), pp. 1–15; X. Chen, Y. Qi, and M. Wang, preprint, Department of Mathematics, HKUST, Hong Kong, 1998].

Key words. long time behavior, P-Laplacian equation, Cauchy problem, self-similar solutions

AMS subject classifications. 35K65, 35K15

DOI. 10.1137/S0036141002407727

1. Introduction.

In this paper, we study the long time behavior of solutions to the Cauchy problem

$$\text{(I)} \qquad \begin{cases} \mathcal{L}u := u_t - \mathrm{div}(|\nabla u|^{p-2}\nabla u) = -u^q & \text{in } R^n \times (0,\infty), \\ u(x,0) = \phi(x) & \text{in } R^n, \end{cases}$$

where $2n/(n+1) < p < 2$, $q > 1$, and ϕ is a nonnegative continuous function which decays to 0 as $|x| \to \infty$. The existence, uniqueness, and Hölder continuity of nonnegative weak solutions of (I) are well established by Chen and DiBenedetto [1, 2] and DiBenedetto and Herrero [7] under much weaker conditions, say, if $\phi \in L^1_{loc}(R^n)$. Therefore, our continuity assumption on ϕ is more than necessary and is used only for the convenience of stating our results.

Our major concern is the behavior of $u(\cdot,t)$ as $t \to \infty$ and how it is influenced by

(a) the decay rate of $\phi(x)$ as $|x| \to \infty$, and

(b) the competition between the diffusion $\mathrm{div}(|\nabla u|^{p-2}\nabla u)$ and the absorption u^q.

It turns out that a critical decay rate for ϕ is $O(|x|^{-\mu})$, and a critical exponent of absorption is q^*, where μ and q^* are defined by

$$q^* := p - 1 + \frac{p}{n}, \qquad \mu := \frac{p}{q+1-p} = \frac{n}{1 + n(q-q^*)/p}. \tag{1.1}$$

The exponents μ and q^* arise naturally from finding radially symmetric and scaling invariant solutions, i.e., self-similar solutions of the form $t^{-\beta}w(|x|t^{-\gamma})$. For later

*Received by the editors May 16, 2002; accepted for publication (in revised form) December 23, 2002; published electronically June 10, 2003.
http://www.siam.org/journals/sima/35-1/40772.html

†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (xinfu+@pitt.edu). This author wishes to acknowledge the financial support of the National Science Foundation grant DMS-9971043, USA.

‡Department of Mathematics, University of California, Santa Barbara, CA 93106 (maqi@math.ucsb.edu). This author was partially supported by HK RGC grant HKUST630/95P.

§Department of Applied Mathematics, Southeast University, Nanjing 210018, P.R. China (mxwang@seu.edu.cn). This author is grateful for the support of PRC grant NSFC-19831060 and HK RGC grant HKUST630/95P.

references, we cite several relevant results on self-similar solutions [3, 5, 6, 8, 9, 14, 15, 17, 20, 21, 22].

PROPOSITION 1.1. *Let $p > 2n/(n+1)$ and $\mathcal{L}u = u_t - (|\nabla u|^{p-2}\nabla u)$.*

(1) *A self-similar solution to $\mathcal{L}u = 0$ has the form $t^{\gamma\nu}w(xt^{\gamma})$ with $\gamma = -1/[p + (p-2)\nu]$ and some $\nu \in \mathbb{R}$. Self-similar solutions to $\mathcal{L}u = -u^q$ have the form $t^{\gamma\mu}w(xt^{\gamma})$ with $\gamma = -1/[p + (p-2)\mu]$.*

(2) *For every $\nu \in (0, n)$ and $B \geq 0$, the solution G_B^{ν} to $\mathcal{L}u = 0$ with initial value $B|x|^{-\nu}$ is self-similar of the form $t^{-\gamma\nu}w(|x|t^{-\gamma})$ with $\gamma = 1/[p + (p-2)\nu]$.*

For each $c > 0$, there is a unique fundamental solution E_c of $\mathcal{L}u = 0$ with initial mass c. It is self-similar of the form $t^{-\gamma n}w(|x|t^{-\gamma})$ with $\gamma = 1/[p + (p-2)n]$.

(3) *For each $A > 0$, there is a unique radially symmetric self-similar solution W_A to $\mathcal{L}u = -u^q$ such that $W(x, 0+) = A|x|^{-\mu}$ for all $x \neq 0$.*

When $1 < q < q^*$ (i.e., $\mu > n$), W_A is monotonic in A, and $W_0 := \lim_{A \searrow 0} W_A$ is the unique very singular solution (VSS) of $\mathcal{L}u = -u^q$ in the sense that

$$\lim_{t \searrow 0} \sup_{|x| > \varepsilon} W_0(x,t) = 0 \quad \text{and} \quad \lim_{t \searrow 0} \int_{|x| \leq \varepsilon} W_0 dx = \infty \qquad \forall\, \varepsilon > 0.$$

Also, for every $c > 0$, $\mathcal{L}u = -u^q$ admits a unique fundamental solution u^c with initial mass c. Fundamental solutions are not self-similar, but $\lim_{c \to \infty} u^c = W_0$ is the self-similar VSS.

When $q = q^*$ (i.e., $\mu = n$), W_A is monotonic in A and $\lim_{A \searrow 0} W_A = 0$ in $\mathbb{R}^n \times (0, \infty)$. There are neither fundamental solutions nor VSSs to $\mathcal{L}u = -u^{q^*}$.

When $q > q^*$ (i.e., $\mu < n$), W_A is the solution with the initial data $A|x|^{-\mu} \in L^1_{loc}(\mathbb{R}^n)$. There are neither fundamental solutions nor VSSs to $\mathcal{L}u = -u^q$.

In what follows, $W_0 \equiv 0$ when $q \geq q^*$, and W_0 is the VSS when $1 < q < q^*$.

To study the long time behavior of solutions of (I), it is convenient to divide the decay rates of ϕ ($\phi \not\equiv 0$) as follows:

(A1) $\lim_{|x| \to \infty} |x|^{\mu}\phi(x) = \infty$;

(A2) $\lim_{|x| \to \infty} |x|^{\mu}\phi(x) = A \in [0, \infty)$;

(B1) $q > q^*$ and $\lim_{|x| \to \infty} |x|^{\nu}\phi(x) = B \in [0, \infty)$ for some $\nu \in (\mu, n)$;

(B2) $q > q^*$ and $\phi \in L^1(\mathbb{R}^n)$.

(B3) $q = q^*$ and $\lim_{x \to \infty} |x|^{\alpha}\phi(x) = A > 0$ for some $\alpha > n$.

Note that the analogues of (B1), (B2) are not needed for $1 < q < q^*$ since $\mu > n$ then.

For $p = 2$, questions (a) and (b) have been discussed by Gmira and Veron [12], Kamin and Peletier [14, 15], Escobedo and Kavian [9], and recently by Escobedo, Kavian, and Matano [10] and Herraiz [13].

PROPOSITION 1.2. *Let $p = 2, q > 1$, and u solve* (I). *Set $q^* = 1 + 2/n$, $\mu = 2/(q-1)$, and $\gamma = 1/2$.*

(1) (See [12].) *If* (A1) *holds, then*

$$\lim_{t \to \infty} \sup_{|x| \leq at^{\gamma}} |t^{1/(q-1)}u(x,t) - (q-1)^{-1/(1-q)}| = 0 \quad \forall\, a > 0. \tag{1.2}$$

(2) (See [10].) *If $1 < q < q^*$ and* (A2) *holds, then*

$$\lim_{t \to \infty} \sup_{|x| < at^{\gamma}} t^{1/(q-1)}|u(x,t) - W_A(x,t)| = 0 \quad \forall\, a > 0. \tag{1.3}$$

(3) (See [15].) *If* (B1) *holds, then*

$$\lim_{t \to \infty} \sup_{|x| \leq at^{\gamma}} t^{\gamma\nu}|u(x,t) - G_B^{\nu}(x,t)| = 0 \quad \forall\, a > 0. \tag{1.4}$$

(4) (See [12].) *If* (B2) *holds, then for some* $c \in [0, \infty)$

$$\lim_{t\to\infty} \sup_{|x|<at^\gamma} t^{\gamma n}|u(x,t) - E_c(x,t)| = 0 \quad \forall\, a > 0. \tag{1.5}$$

(5) (See [13].) *If* $q > q^*$ *and* $\lim_{|x|\to\infty} |x|^n \phi(x) = B > 0$, *then*

$$\lim_{t\to\infty} \sup_{|x|<at^\gamma} t^{\gamma n}|u/\log t - E_c(x,t)| = 0 \quad \forall\, a > 0, \tag{1.6}$$

where $c = \omega_n B/2$ *with* ω_n *the area of unit sphere* S^{n-1}.

(6) (See [13].) *If* (B3) *holds, then*

$$\lim_{t\to\infty} \sup_{|x|<at^\gamma} t^{n/2}|u(x,t)(\log t)^{n/2} - E_{C_n}(x,t)| = 0 \quad \forall\, a > 0 \tag{1.7}$$

for a unique constant $C_n > 0$.

Remark. It is of great interest to know whether the results in Propositions 1.1 and 1.2 are still valid if the precise pointwise assumptions on initial values are replaced by more general integral estimates. However, we do not have a clear idea of how to approach a situation like that. One possible difficulty is the loss of scaling laws.

For $p \neq 2$, similar but not as complete results were obtained by Zhao [23].

PROPOSITION 1.3. *Let* $p > 2n/(n+1)$, $q > 1$, *and* u *solve* (I).

(1) *If* $\lim_{|x|\to\infty} |x|^{\mu-\varepsilon}\phi(x) = \infty$ *for some* $\varepsilon > 0$, *then* (1.2) *holds with* $\gamma = 1/[p + (p-2)\mu]$.

(2) *If* $1 < q < q^*$ *and* $\lim_{|x|\to\infty} |x|^{\mu+\varepsilon}\phi(x) = 0$ *for some* $\varepsilon > 0$, *then* (1.3) *holds with* $A = 0$ *and* $\gamma = 1/[p + (p-2)\mu]$. *(The result for when* $p > 2$ *and* ϕ *has compact support was obtained by Kamin and Vasquez* [17].*)*

(3) *If* (B1) *holds, so does* (1.4) *with* $\gamma = 1/[p + (p-2)\nu]$.

(4) *If* $\lim_{|x|\to\infty} |x|^{n+\varepsilon}\phi(x) = 0$ *for some* $\varepsilon > 0$, *then* (1.5) *holds with* $\gamma = 1/[p + (p-2)n]$ *and some* $c \in [0, \infty)$.

There are also analogous results for the porous medium equation

$$\text{(II)} \qquad \begin{cases} u_t = \triangle u^m - u^q & \text{in } R^n \times (0,\infty), \\ u(x,0) = \phi(x) & \text{in } R^n. \end{cases}$$

In this case, the critical exponents should be defined as $\mu = 2/(q-m)$ and $q^* = m + 2/n$. For radially symmetric self-similar solutions and for $m > (1-2/n)_+$ and $q > \max\{1, m\}$, Proposition 1.1 holds with $\mathcal{L}$ defined as $\mathcal{L}u = u_t - \Delta u^m$; see [4, 12, 16, 17, 19, 21]. The results on noncritical decay rates were obtained by Kamin and Peletier [16] for $m > 1$ and Peletier and Zhao [21] for $(1-2/n)_+ < m < 1$, and the result on the critical decay rate was obtained by Kwak [18].

PROPOSITION 1.4. *Let* $m > (1-2/n)_+$, $q > \max\{1, m\}$, *and* u *be a solution to* (II).

(1) *If* (A1) *holds, so does* (1.2) *with* $\gamma = 1/[2 + (m-1)\mu]$.

(2) *Suppose* $1 < q < q^*$. *If* (A2) *holds with* $A > 0$ *or* $A = 0$ *and in addition* $\lim_{|x|\to\infty} |x|^{\mu+\varepsilon}\phi(x) = 0$ *for some* $\varepsilon > 0$, *then* (1.3) *holds with* $\gamma = 1/[2 + (m-1)\mu]$.

(3) *The condition* (B1) *implies* (1.4) *with* $\gamma = 1/[2 + (m-1)\nu]$.

(4) *The condition* (B2) *implies* (1.5) *with* $\gamma = 1/[2 + (m-1)n]$.

(5) *If* $q > q^*$ *and* $\lim_{|x|\to\infty} |x|^n \phi(x) = A > 0$, *then*

$$\lim_{t\to\infty} \sup_{|x|<at^\gamma} \tau^{\gamma n}|u(x,t)/\log\tau - E_c(x,\tau)| = 0 \quad \forall\, a > 0, \tag{1.8}$$

where $c = c(m, n, a) > 0$, $t = c^{1-m}\tau(\log\tau)^{1-m}$, *and* $\gamma = 1/[2 + (m-1)n]$.

One may notice that the recent results of Escobedo, Kavian, and Matano [10] listed in Proposition 1.2 (2) and that of Kwak [18] in Proposition 1.4 (2) address the critical decay of ϕ with sharp results for semilinear and porous media cases, respectively. We would like to mention that though some ideas of Escobedo, Kavian, and Matano [10] and Kwak [18] can be adopted to study (I), their approach cannot be extended easily to the present case.

In this paper, we shall extend the results of Propositions 1.2, 1.3, and 1.4 to the case in which $p \in (2n/(n+1), 2)$. Our main results are the following.

Assume that $2n/(n+1) < p < 2$ *and* $q > 1$. *Let* μ *and* q^* *be defined as in* (1.1), *and let* u *be a solution to* (I). *Let* E_c, G_B^ν, *and* W_A *be defined as in Proposition* 1.1. *Denote* $L^\infty = L^\infty(\mathbb{R}^n)$.

THEOREM 1.5. *Assume* (A1). *Then* (1.2) *holds with* $\gamma = 1/[p + (p-2)\mu]$.

THEOREM 1.6. *Assume* (A2). *Then* (1.3) *holds with* $\gamma = 1/[p + (2-p)\mu]$; *more precisely,*

$$\lim_{t\to\infty} t^{\gamma\mu}\|u(\cdot,t) - W_A(\cdot,t)\|_{L^\infty} = \lim_{t\to\infty} \|t^{\gamma\mu}u(yt^\gamma,t) - W_A(y,1)\|_{L^\infty} = 0. \tag{1.9}$$

In particular, when $A = 0$ *and* $q = q^*$,

$$\lim_{t\to\infty} t^{1/(q^*-1)}\|u(\cdot,t)\|_{L^\infty} = 0 \quad \textit{and} \quad \lim_{t\to\infty} t^{1/(q^*-1-\varepsilon)}u(0,t) = \infty \quad \forall \varepsilon > 0. \tag{1.10}$$

THEOREM 1.7. *Assume* (B1). *Then* (1.4) *holds with* $\gamma = 1/[p + (p-2)\nu]$; *more precisely,*

$$\lim_{t\to\infty} t^{\gamma\nu}\|u(\cdot,t) - G_B^\nu(\cdot,t)\|_{L^\infty} = \lim_{t\to\infty} \|t^{\gamma\nu}u(yt^\gamma,t) - G_B^\nu(y,1)\|_{L^\infty} = 0. \tag{1.11}$$

THEOREM 1.8. *Assume* (B2). *Then for* $\gamma = 1/[p + (p-2)n]$ *and some constant* $c > 0$,

$$\lim_{t\to\infty} t^{\gamma n}\|u(\cdot,t) - E_c(\cdot,t)\|_{L^\infty} = \lim_{t\to\infty} \|t^{\gamma n}u(yt^\gamma,t) - E_c(y,1)\|_{L^\infty} = 0. \tag{1.12}$$

Remark. We make some interesting observations.

1. Theorem 1.5 says that when ϕ decays too slowly, absorption dominates diffusion.

2. When $1 < q < q^*$, $W_0 = VSS > 0$, so Theorem 1.6 indicates that the absorption is still strong, but diffusion plays a role to balance it.

When $q \geq q^*$, $W_0 \equiv 0$, so the conclusion of Theorem 1.6 for $A = 0$ is insufficient. Theorems 1.7 and 1.8 exactly compensate for this insufficiency for the case in which $q > q^*$ but not for $q = q^*$.

3. Here we show explicitly that the constant c in Theorem 1.8 is positive, which is not mentioned in [12, 16, 21]. With this improvement, we can conclude that when ϕ decays quickly enough (say, of compact support), then in large time, the solution approaches the VSS when $q \in (1, q^*)$ or the fundamental solution E_c of the pure diffusion equation when $q > q^*$.

4. For the case in which $q = q^*$ and ϕ decays quickly (say, of compact support or L^1), except the two limits in (1.10), we do not have a more precise description. Here, different scaling is needed. For the case in which $p = 2$, we refer interested readers to the work of Herraiz [13] and some insightful observation of Galaktionov, Kurdyumov, and Samarskii [11].

5. The most striking fact, of course, is that in each and every case being discussed, the limiting behavior of $u(x,t)$ as $t \to \infty$ is characterized by a self-similar solution of the same equation or the corresponding equation without absorption.

6. We intend to answer questions (a) and (b) as follows: The absorption dynamics ($u_t = -u^{-q}$) decreases the solution (in time) with a $O(t^{-1/(q-1)})$ rate. The diffusion process ($u_t = \operatorname{div}(|\nabla u|^{p-2}\nabla u)$) makes mass diffuse, thereby decreasing u at a maximum rate of $O(t^{-n/(p+[p-2]n)}) = O(t^{-1/(q^*-1)})$, as can be seen from the fundamental solutions. Hence, when $1 < q < q^*$, the diffusion-related decay is slower than that of the absorption, so absorption is strong and will dominate, or at least not subordinate to, the diffusion. When $q > q^*$, as diffusion can diminish u at a rate ranging from $O(1)$ (for constant initial data) to $O(t^{-1/(q^*-1)})$ (for L^1 initial data), the relative strengths of absorption and diffusion then heavily depend on the decay rate of the initial data. For fast decay, diffusion dominates; for slow decay, absorption prevails; the critical decay rate is $O(|x|^{-\mu})$.

We shall first prove Theorem 1.8 in the next section since, with the results we established in [3, 5], the proof becomes very easy. Although Theorem 1.7 had already been proven by Zhao [23] (cf. Proposition 1.3 (3)), we provide a different proof in section 2 for completeness. In section 3, we first prove Theorem 1.6 for the case in which $A \in (0,\infty)$. Then the cases in which $A = \infty$ and $A = 0$ can be handled by taking appropriate limits.

2. Proof of Theorems 1.7 and 1.8. For each $\nu \in [\mu, p/(2-p))$ and $\lambda \geq 1$, we define

$$\gamma = 1/[p + (p-2)\nu], \quad \sigma = \gamma(q+1-p)[\nu-\mu], \quad u_\lambda = \lambda^{\gamma\nu} u(\lambda^\gamma x, \lambda t). \tag{2.1}$$

Then

$$\begin{cases} \mathcal{L}u_\lambda = -\lambda^{-\sigma} u^q & \text{in } \mathbb{R}^n \times (0,\infty), \\ u_\lambda(x,0) = \phi_\lambda(x) := \lambda^{\gamma\nu}\phi(\lambda^\gamma x) & \text{on } \mathbb{R}^n \times \{0\}. \end{cases} \tag{2.2}$$

Proof of Theorem 1.8. Set $\nu = n$. Let ψ_λ be the solution to $\mathcal{L}\psi_\lambda = 0$ with initial value ϕ_λ. Define $c_0 = \int_{\mathbb{R}^n} \phi$. Then $\int_{R^n} \phi_\lambda(x)\, dx = c_0$ for every $\lambda > 1$, and $\lim_{\lambda\to\infty} \int_{|x|>\varepsilon} \phi_\lambda(x)\, dx = 0$ for all $\varepsilon > 0$. Hence $\{\frac{1}{c_0}\phi_\lambda\}$ is a δ-sequence (i.e., a sequence approaching the δ function). It then follows from Theorem 3.1 of [5] that

$$\lim_{\lambda\to\infty} \psi_\lambda = E_{c_0} \qquad \text{in} \quad L^\infty((\varepsilon,\varepsilon^{-1});\ L^1(R^n) \cap C(\mathbb{R}^n)) \qquad \forall\ \varepsilon > 0.$$

Now we consider $\{u_\lambda\}_{\lambda>1}$. First by applying an L^∞ estimate (cf. [6, p. 127] or [7]) for ψ_λ and then by comparison, we obtain $u_\lambda \leq \psi_\lambda \leq M(p,n,c_0)t^{-\gamma n}$ for all $\lambda > 0$ and $t > 0$. By the regularity results in [1, 2, 7], $\{u_\lambda\}_{\lambda>1}$ is an equicontinuous family in any compact subset of $\mathbb{R}^n \times (0,\infty)$. Hence there exist a function U and a sequence $\{\lambda_j\}$ with $\lim_{j\to\infty} \lambda_j = \infty$ such that $\lim_{j\to\infty} u_{\lambda_j} = U(x,t)$ uniformly in any compact subset of $R^n \times (0,\infty)$. As σ is positive, (2.2) implies $\mathcal{L}U = 0$ in $\mathbb{R}^n \times (0,\infty)$. In addition, from $u_\lambda \leq \psi_\lambda$, there holds $U \leq E_{c_0}$ in $R^n \times (0,\infty)$ so that $\lim_{t\searrow 0} \sup_{|x|>\varepsilon} U(x,t) = 0$ for all $\varepsilon > 0$. Furthermore, for every fixed $\tau > 0$,

$$\begin{aligned} \limsup_{j\to\infty} \int_{\mathbb{R}^n} |U - u_{\lambda_j}|(x,\tau)dx &= \lim_{M\to\infty} \limsup_{j\to\infty} \int_{|x|>M} |U - u_{\lambda_j}|(x,\tau)|dx \\ &\leq \lim_{M\to\infty} \limsup_{j\to\infty} \int_{|x|\geq M} (E_c + \psi_{\lambda_j})(x,\tau)dx = 0. \end{aligned} \tag{2.3}$$

Note that $\int_{\mathbb{R}^n} u(x,t)dx$ is a nonincreasing function of t, so $c := \lim_{t\to\infty}\int_{\mathbb{R}^n} u(x,t)dx$ exists. Consequently, for every $\tau > 0$,

$$\int_{\mathbb{R}^n} U(x,\tau)dx = \lim_{j\to\infty}\int_{\mathbb{R}^n} u_{\lambda_j}(x,\tau)dx = \lim_{j\to\infty}\int_{\mathbb{R}^n} u(y,\lambda_j\tau)dy$$
$$= \lim_{t\to\infty}\int_{\mathbb{R}^n} u(x,t)dx = c.$$

Thus U is a fundamental solution with mass c; i.e., $U = E_c$. As c is independent of $\{\lambda_j\}$, the whole sequence $\{u_\lambda\}$ must converge to E_c in any compact subset of $\mathbb{R}^n \times (0,\infty)$. A similar argument to that in (2.3) yields $u_\lambda(\cdot,\tau) \to E_c(\cdot,\tau)$ in $L^1(\mathbb{R}^n)$ for every $\tau > 0$. Local regularity estimates then imply that $u_\lambda(\cdot,1) \to E_c(\cdot,1)$ in $L^\infty(\mathbb{R}^n)$. The assertion (1.12) thus follows from the definition of u_λ and the scaling invariance of E_c.

It remains to show $c > 0$. As $u \le \psi_1 \le Mt^{-\gamma n}$ for all $t > 0$,

$$-\frac{d}{dt}\int_{R^n} u(x,t)\ dx = \int_{R^n} u^q(x,t)\ dx \le Mt^{-(q-1)\gamma n}\int_{R^n} u(x,t)\ dx.$$

An integration over $[1,t]$ then gives

$$\log\int_{R^n} u(x,t)\ dx \ge \log\int_{R^n} u(\cdot,1)\ dx + \frac{M(1-t^{1-(q-1)\gamma n})}{1-(q-1)\gamma n} \qquad \forall\ t > 1.$$

Simple calculation shows $1-(q-1)\gamma n = \gamma n[q^*-q] < 0$. Thus $\lim_{t\to\infty}\log\int_{R^n} u(\cdot,t) > -\infty$; i.e., $c > 0$. This completes the proof. □

Proof of Theorem 1.7. Let $\nu \in (\mu, n)$ be as in the statement of the theorem, and let γ, σ, u_λ, and ϕ_λ be as in (2.1). Let G_B^ν be the solution to $\mathcal{L}G = 0$ with initial data $B|x|^{-\nu}$. Fix an arbitrary $\delta \in (0, (n/\nu - 1)/q)$ (so that $\nu(1+q\delta) < n$). To prove the theorem, it suffices to show

$$\lim_{\lambda\to\infty}\sup_{0<t<2, x_0\in\mathbb{R}^n}\int_{|x-x_0|<1}|u_\lambda(x,t) - G_B^\nu(x,t)|^{1+\delta}dx = 0. \tag{2.4}$$

Indeed, by local regularities of solutions, the above limit implies that $\|u_\lambda(\cdot,1) - G_B^\nu(\cdot,1)\|_{L^\infty(\mathbb{R}^n)} \to 0$ as $\lambda\to\infty$, and therefore (1.11) follows from the definition of u_λ and the scaling invariance of G_B^ν. We now prove (2.4).

Let $x_0 \in \mathbb{R}^n$ and $R > 2$ be arbitrarily given. We can find a smooth cut-off function $\zeta(x)$ satisfying $0 \le \zeta \le 1$ in $\mathbb{R}^n$, $\zeta = 1$ when $|x-x_0| < 1$, $\zeta = 0$ when $|x-x_0| > R$, and $|\nabla\zeta| \le 3/R$ in $\mathbb{R}^n$. Let $s = (1+q\delta)p/(2-p)$. In what follows, all positive constants depending only on p, n, q, δ, and ν will be denoted by the same letter C.

Integrating the identity $0 = (1+q\delta)u_\lambda^{q\delta}\zeta^s(\mathcal{L}u_\lambda + \lambda^{-\sigma}u_\lambda^q)$ over $\mathbb{R}^n$ and using integration by parts, the assumption $|\nabla\zeta| \le 3/R$, and the Cauchy inequality

$$su^{q\delta}\zeta^{s-1}|\nabla u_\lambda|^{p-2}\nabla u_\lambda\cdot\nabla\zeta \le mu_\lambda^{q\delta-1}|\nabla u_\lambda|^p\zeta^s + \frac{1}{q\delta+1}u_\lambda^{q\delta+1}\zeta^s$$
$$+ C|\nabla\zeta|^{(q\delta+1)p/(2-p)},$$

we obtain

$$\frac{d}{dt}\int_{R^n} u_\lambda^{1+q\delta}\zeta^s dx + \int_{R^n}\{(1+q\delta)\lambda^{-\sigma}u_\lambda^{q+q\delta}\zeta^s - u_\lambda^{1+q\delta}\zeta^s\}dx \tag{2.5}$$
$$\le CR^{n-(q\delta+1)p/(2-p)}.$$

Before proceeding further, we first establish an elementary algebraic inequality. Since $p \in (1,2)$, there are positive constants $C(p,n)$ and $c(p,n)$ such that for every $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$,

$$\begin{aligned} |\ |\mathbf{a}|^{p-2}\mathbf{a} - |\mathbf{b}|^{p-2}\mathbf{b}\ | &\le C(p,n)\{|\mathbf{a}-\mathbf{b}|^2(|\mathbf{a}|+|\mathbf{b}|)^{p-2}\}^{1-1/p}, \\ (|\mathbf{a}|^{p-2}\mathbf{a} - |\mathbf{b}|^{p-2}\mathbf{b})\cdot(\mathbf{a}-\mathbf{b}) &\ge c(p,n)|\mathbf{a}-\mathbf{b}|^2(|\mathbf{a}|+|\mathbf{b}|)^{p-2}. \end{aligned}$$

(They can be proven first for $1 = |\mathbf{a}| \ge |\mathbf{b}|$, and then for the general case by scaling.) It then follows that for any $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^n$,

$$(|\mathbf{a}|^{p-2}\mathbf{a} - |\mathbf{b}|^{p-2}\mathbf{b})\cdot((\mathbf{a}-\mathbf{b})+\mathbf{c}) \ge -\hat{C}(p,n)|\mathbf{c}|^p. \tag{2.6}$$

Denote $v = u_\lambda - G_B^\nu$. Multiplying the difference of the equations for u_λ and G_B^ν by $|v|^{\delta-1}v\zeta^s$, integrating the resulting equation over $\mathbb{R}^n$, and using integration by parts and the inequality (2.6) with $\mathbf{a} = \nabla u_\lambda, \mathbf{b} = \nabla G_B^\nu$, and $\mathbf{c} = sv\delta^{-1}\zeta^{-1}\nabla\zeta$, we then obtain

$$\frac{d}{dt}\int_{R^n}|v|^{1+\delta}\zeta^s dx \le \int_{R^n}\{(1+\delta)\lambda^{-\sigma}|v|^\delta u_\lambda^q\zeta^s dx + C|v|^{\delta-1+p}\zeta^{s-p}|\nabla\zeta|^p\}.$$

Cauchy's inequality and the assumptions $|\nabla\zeta| < 3/R$ and $s = (1+q\delta)p/(2-p)$ then yield

$$\frac{d}{dt}\int_{R^n}|v|^{1+\delta}\zeta^s dx - \int_{R^n}\{|v|^{1+\delta}\zeta^s + C\lambda^{-\sigma-\sigma\delta}u_\lambda^{q+q\delta}\zeta^s\}dx \le CR^{n-(1+\delta)p/(2-p)}.$$

Now adding to it a $C\lambda^{-\sigma\delta}$ multiple of (2.6), we obtain

$$\begin{aligned} \frac{d}{dt}\int_{\mathbb{R}^n}\{|v|^{1+\delta} + C\lambda^{-\sigma\delta}u_\lambda^{1+q\delta}\}\zeta^s dx - \int_{\mathbb{R}^n}\{|v|^{1+\delta} + C\lambda^{-\sigma\delta}u_\lambda^{1+q\delta}\}\zeta^s dx \\ \le CR^{n-(1+\delta)p/(2-p)} \end{aligned}$$

for all $\lambda > 1$. Gronwall's inequality then yields

$$\sup_{0<t<2} e^{-t}\int_{\mathbb{R}^n}|v|^{1+\delta}\zeta^s dx \le CR^{n-(1+\delta)p/(2-p)} + \int_{\mathbb{R}^n}\zeta^s\{|v(\cdot,0)|^{1+\delta} + C\lambda^{-\sigma\delta}\phi_\lambda^{1+q\delta}\}dx.$$

Denote by $B_r(x_0)$ the ball of radius r and center x_0. That ζ is a cut-off function then gives, after replacing v by $u_\lambda - G_B^\nu$,

$$\begin{aligned} &\sup_{x_0\in\mathbb{R}^n}\sup_{0<t<2}\int_{B_1(x_0)}|u_\lambda - G_B^\nu|^{1+\delta}dx \le CR^{n-(1+\delta)p/(2-p)} \\ &+ C\sup_{x_0\in\mathbb{R}^n}\int_{B_R(x_0)}|\phi_\lambda - B|x|^{-\nu}|^{1+\delta}dx + C\sup_{x_0\in\mathbb{R}^n}\lambda^{-\sigma\delta}\int_{B_R(x_0)}\phi_\lambda^{1+q\delta}dx. \end{aligned} \tag{2.7}$$

Set $M = \sup_{x\in\mathbb{R}^n}|x|^\nu\phi(x)$. Then the definition of ϕ_λ implies $\phi_\lambda(x) \le M|x|^{-\nu}$. Consequently, the last term in (2.7) converges to zero as $\lambda \to \infty$.

To estimate the first integral in (2.7), one notices that $\phi_\lambda(x) \to B|x|^{-\nu}$ uniformly in $\mathbb{R}^n \setminus B_\varepsilon(0)$ for any $\varepsilon > 0$. As $\int_{B_\varepsilon(0)}|x|^{-\nu(1+\delta)}dx = O(\varepsilon^{n-(1+\delta)\nu})$, we see also that, as $\lambda \to \infty$, the first term in (2.7) converges to zero. Hence, sending $\lambda \to \infty$, we obtain from (2.7) that

$$\limsup_{\lambda\to\infty}\sup_{x_0\in\mathbb{R}^n,0<t<2}\int_{B_1(x_0)}|u_\lambda - G_B^\nu|^{1+\delta}dx \le CR^{n-(1+\delta)p/(2-p)}.$$

Sending $R \to \infty$, we then obtain (2.4), thereby completing the proof. □

3. Proof of Theorems 1.5 and 1.6. Set $\gamma = 1/[p+(p-2)\mu]$. Then $(q-1)\gamma\mu = 1$. For some $\delta > 0$ to be chosen later, we make a self-similar transformation:

$$u(x,t) = [(t+\delta)/\gamma\mu]^{-\gamma\mu} w(y,\tau), \quad y = [(t+\delta)/\gamma\mu]^{-\gamma} x, \quad \tau = \gamma\mu \, \log[(t+\delta)/\delta]. \tag{3.1}$$

Then w satisfies

$$\begin{cases} w_\tau = \operatorname{div}(|\nabla w|^{p-2}\nabla w) + \mu^{-1} y \cdot \nabla w + w - w^q, & y \in R^n, \quad \tau > 0, \\ w(y,0) = [\delta/\gamma\mu]^{\gamma\mu} \phi([\delta/\gamma\mu]^\gamma y), & y \in \mathbb{R}^n. \end{cases} \tag{3.2}$$

Now we fix $\delta = \gamma\mu \|\phi\|^{-1/\gamma\mu}_{L^\infty(\mathbb{R}^n)}$. Then

(1) $0 \le w(y,0) \le 1$ in R^n;

(2) $\lim_{|y|\to\infty} |y|^\mu w(y,0) = \lim_{|x|\to\infty} |x|^\mu \phi(x) = A$.

To prove Theorem 1.6 for $A \in (0,\infty)$, we need only to show that

$$\lim_{\tau\to\infty} w(y,\tau) = W_A(y,1) = w_A(|y|) \quad \text{uniformly in } y \in \mathbb{R}^n; \tag{3.3}$$

here $w_A(r)$, $r \in [0,\infty)$, is the unique solution to the following boundary value problem:

$$\begin{cases} \mathcal{L}_r w := |w'|^{p-2}[(p-1)w'' + (n-1)r^{-1}w'] + w + \mu^{-1} r w' - w^q = 0, \quad r > 0, \\ w_A'(0) = 0, \quad w(r) > 0 \quad \text{on } [0,\infty), \quad \text{and} \quad \lim_{r\to\infty} r^\mu w(r) = A. \end{cases} \tag{3.4}$$

We prove (3.3) by using sub- and supersolutions. To ease the computation, it is convenient to introduce new dependent and independent variables

$$w(r) = r^{-\mu} J(\theta), \qquad r = Re^\theta, \tag{3.5}$$

where R is a parameter of our choice. Then $\mathcal{L}_r w = r^{-\mu q}\mathcal{L}_\theta J$, where, denoting $\dot{} = d/d\theta$,

$$\mathcal{L}_\theta J := |\dot J - \mu J|^{p-2}\{a(\ddot J - \mu \dot J) + b(\dot J - \mu J)\} + \mu^{-1} R^l e^{l\theta} \dot J - J^q,$$
$$a = (p-1), \quad b = (n-1) - (\mu+1)a, \quad l = \mu(q-1).$$

LEMMA 3.1. *For every $A \in (0,\infty)$, there exists $R_0(A) > 0$ such that for every $R > R_0$, there are two functions $w^+_{A,R}(r)$ and $w^-_{A,R}(r)$ with the following properties:*

(a) $w^+_{A,R}(r) = 1$ *in* $[0,R]$, $Ar^{-\mu} \le w^+_{A,R}(r) \le 1$ *in* $[R,\infty)$;

(b) $w^-_{A,R}(r) = 0$ *in* $[0,R]$, $0 \le w^-_{A,R}(r) \le Ar^{-\mu}$ *in* $[R,\infty)$;

(c) $\mathcal{L}_r w^+_{A,R} \le 0$ *and* $\mathcal{L}_r w^-_{A,R} \ge 0$ *in* $(0,\infty)$ *in the distributional sense;*

(d) $\lim_{r\to\infty} r^\mu w^+_{A,R}(r) = \lim_{r\to\infty} r^\mu w^-_{A,R}(r) = A$.

Proof. (I) Construction of $w^+_{A,R}(r)$. We define $w^+_{A,R} = 1$ in $[0,R]$ and $w^+_{A,R} = r^{-\mu}[A + (R^\mu - A)(R/r)^l]$ in $[R,\infty)$. Then $w^+_{A,R}$ satisfies (a) and (d). In addition, $(w^+_{A,R})'' \le 0$ at $r = R$ in the distributional sense. It remains to check $\mathcal{L}_r w^+_{A,R} \le 0$ in (R,∞) or, equivalently, $\mathcal{L}_\theta J \le 0$ in $(0,\infty)$, where

$$J = A + (R^\mu - A)e^{-l\theta}.$$

Note that $\mu J - \dot J = \mu A + (R^\mu - A)(l+\mu)e^{-l\theta} > \mu A$ in $(0,\infty)$. Also, assuming $R > 2A$, then $R^\mu - A > R^\mu/2$. Hence, using $p - 2 < 0$, we can calculate, for all $\theta \in (0,\infty)$,

$$\mathcal{L}_\theta J = |rJ' - \mu J|^{p-2}\{(l+\mu)(al-b)(R^\mu - A)e^{-l\theta} - \mu b A\}$$
$$-l\mu^{-1}(R^\mu - A)R^l - J^q \le (\mu A)^{p-2}(l+\mu)(al+|b|)R^\mu - lR^\mu R^l/(2\mu) \le 0$$

if R is sufficiently large (depending on A). The construction of $w^+_{A,R}(r)$ is now complete.

(II) Construction of $w^-_{A,R}(r)$. We need only construct $J(\theta) := r^\mu w^-_{A,R}(r)|_{r=Re^\theta}$. Let $m \geq 1$ be the smallest integer such that $(p-1)(2m+1) \geq 1$. Define

$$\Theta(J) = \begin{cases} \mu[J - (J-\varepsilon)^{2m+1}], & 0 < J < 2\varepsilon, \\ (A-J)D, & 2\varepsilon \leq J < A, \end{cases}$$
$$D = \mu[2\varepsilon - \varepsilon^{2m+1}]/[A - 2\varepsilon],$$
$$\varepsilon = \min\{1/(2m+1),\ 1/(4A),\ l/(4\mu A)\}.$$

Clearly Θ is Lipschitz continuous and positive in $[0, A)$. In addition, $0 < D \leq l$.

Let $J(\theta), \theta \in [0, \infty)$, be the solution to

$$\dot{J} = \Theta(J) \quad \text{in } (0,\infty), \qquad J(0) = 0.$$

Since $\Theta > 0$ in $[0, A)$ and $\Theta(A) = 0$, J is strictly increasing, and $\lim_{\theta\to\infty} J(\theta) = A$.

Now we define $w^-_{A,R} = 0$ for $r \leq R$ and $w^-_{A,R} = r^{-\mu} J(\theta)|_{\theta=\log(r/R)}$ for $r > R$. Then $w^-_{A,R}$ satisfies conditions (b) and (d) in the lemma. In addition, $(w^-_{A,R})'' \geq 0$ at $r = R$ in the distributional sense. It remains to check that $\mathcal{L}_r w^-_{A,R} \geq 0$ in (R, ∞), which is equivalent to checking that $\mathcal{L}_\theta J(\theta) \geq 0$ for all $\theta > 0$.

Let $\theta_1 := \int_0^{2\varepsilon} dJ/\Theta(J)$ be the unique point such that $J = 2\varepsilon$. We consider two cases: (i) $\theta \in (\theta_1, \infty)$ and (ii) $\theta \in (0, \theta_1]$.

Case (i). $\theta \in (\theta_1, \infty)$. Then $J = A - (A-2\varepsilon)e^{D(\theta_1-\theta)}$ and $\dot{J} = D(A-2\varepsilon)e^{D(\theta_1-\theta)}$. Also, $\mu J - \dot{J} = (\mu + D)J - DA \geq 2\varepsilon(\mu + D) - AD = \mu\varepsilon^{2m+1}$. Thus, as $p < 2$ and $D \leq l$,

$$\mathcal{L}_\theta J \geq -(\mu\varepsilon^{2m+1})^{p-2} C(A) + \mu^{-1} R^l D(A - 2\varepsilon) e^{D\theta_1 + (l-D)\theta} - A^q \geq 0$$

provided that R is large enough.

Case (ii). $\theta \in (0, \theta_1]$. Then $J \in (0, 2\varepsilon]$ and $\dot{J} - \mu J = -\mu(J-\varepsilon)^{2m+1}$ so that

$$\begin{aligned} &|\dot{J} - \mu J|^{p-2}\{a(\ddot{J} - \mu\dot{J}) + b(\dot{J} - \mu J)\} \\ &= -\mu^{p-1}|J-\varepsilon|^{(p-2)(2m+1)+2m}\{(2m+1)a\dot{J} + b(J-\varepsilon)\} \geq -C(A) \end{aligned}$$

since $(p-2)(2m+1) + 2m \geq 0$. Noting that $\dot{J} = \mu[J - (J-\varepsilon)^{2m+1}] \geq \mu\varepsilon^{2m+1}$, we have

$$\mathcal{L}_\theta J \geq -C(A) + R^l e^{l\theta} \varepsilon^{2m+1} - (2\varepsilon)^q \geq 0$$

if R is large enough. This completes the proof of the lemma. □

LEMMA 3.2. *For every $A \in (0, \infty)$, $R > R_0(A)$, let $W^\pm_{A,R}(y, \tau)$ be the solution to the PDE in* (3.2) *with initial value $w^\pm_{A,R}(|y|)$. Then*

$$\lim_{\tau\to\infty} \|W^\pm_{A,R}(\cdot, \tau) - w_A(|\cdot|)\|_{L^\infty(R^n)} = 0.$$

Proof. Since $w^+_{A,R}(|y|)$, as a function of (y, τ), is a supersolution to the PDE in (3.2), by comparison, $W^+_{A,R}(\cdot, \tau) \leq w^+_{A,R}(|\cdot|)$ for all $\tau > 0$. In turn, this implies that $W^+_{A,R}(\cdot, \tau + \tau_1) \leq W^+_{A,R}(\cdot, \tau_1)$ for all $\tau > 0$ and $\tau_1 \geq 0$. It then follows that $W^+_{A,R}(\cdot, \tau)$

is monotone decreasing in τ. Similarly, $W^-_{A,R}(\cdot,\tau)$ is monotone increasing in τ. Hence there exist $w^{\infty,\pm}_{A,R}$ such that

$$W^{\pm}_{A,R}(\cdot,\tau) \to w^{\infty,\pm}_{A,R}(\cdot) \qquad \text{as } \tau\to\infty \text{ pointwise in } R^n. \tag{3.6}$$

It is clear that $w^{\infty,\pm}_{A,R}$ solve $\mathcal{L}_r w^{\infty,\pm}_{A,R}=0$ and are radially symmetric and that

$$w^-_{A,R}(|y|) \le w^{\infty,-}_{A,R}(y) \le w^{\infty,+}_{A,R}(y) \le w^+_{A,R}(|y|).$$

It follows that $\lim_{|y|\to\infty}|y|^\mu w^{\infty,\pm}_{A,R} = A$. By the uniqueness of solution to (3.4) (cf. [3]), $w^{\infty,\pm}_{A,R}(y) = w_A(|y|)$. Hence $\lim_{\tau\to\infty} W^{\pm}_{A,R}(\cdot,\tau) = w_A(|\cdot|)$. Since $W^{\pm}_{A,R}(y,\tau) \le w^+_{A,R}(|y|)$, by local regularity, the convergence is uniform in any compact subset of $\mathbb{R}^n$. Further, as $w^+_{A,R}(|y|) = O(|y|^{-\mu})$, the convergence is also in L^∞. This completes the proof. □

Proof of Theorems 1.5 *and* 1.6.

Case 1. $0<A<\infty$. Let $w(y,\tau)$ be the solution to (3.2). We need to prove (3.3). Since we have the limit $\lim_{|y|\to\infty}|y|^\mu w_0(y) = A$, for every $\varepsilon>0$, there exists $R_\varepsilon>0$ such that

$$A-\varepsilon < |y|^\mu w(y,0) < A+\varepsilon \qquad \forall\ |y|\ge R_\varepsilon.$$

It follows from Lemma 3.1 (a), (b) and $0\le w(y,0)\le 1$ that

$$w^-_{A-\varepsilon,R_\varepsilon}(|\cdot|) \le w(\cdot,0) \le w^+_{A+\varepsilon,R_\varepsilon}(|\cdot|).$$

Consequently, by the comparison principle,

$$W^-_{A-\varepsilon,R_\varepsilon}(y,\tau) \le w(y,\tau) \le W^+_{A+\varepsilon,R_\varepsilon}(y,\tau).$$

Thus, by Lemma 3.2,

$$\limsup_{\tau\to\infty} \|w(y,\tau) - w_A(|y|)\|_{L^\infty} \le \|w_{A+\varepsilon} - w_A\|_{L^\infty} + \|w_{A-\varepsilon} - w_A\|_{L^\infty}.$$

However, since $\varepsilon>0$ is arbitrary, we obtain (3.3).

Case 2. $\lim_{|x|\to\infty}|x|^{-\nu}\phi(x) = A = \infty$. By comparison,

$$\liminf_{\tau\to\infty} w(y,\tau) \ge \lim_{A\to\infty} w_A(|y|) = 1$$

uniformly in any compact subset of $y\in\mathbb{R}^n$. Since $w\le 1$ for all y and $\tau\ge 0$, we conclude that as $\tau\to\infty$, $w(y,\tau)\to 1$ uniformly in any compact subset of $\mathbb{R}^n$. Using the definition of $w(y,\tau)$ in (3.1) and the fact that $\gamma\mu = 1/(q-1)$, the assertion of Theorem 1.5 thus follows.

Case 3. $A=0$. If $1<q<q^*$, then

$$\limsup_{\tau\to\infty} w(y,\tau) \le \lim_{A\to 0} w_A(|y|) = w_0(|y|) = W_0(y,1),$$

where w_0 is the unique positive solution of (3.4) with $A=0$ and W_0 is the unique VSS. It remains to show that

$$\liminf_{\tau\to\infty} w(y,\tau) \ge w_0(|y|). \tag{3.7}$$

To do this, we go back to the PDE (I). Set $\nu = \mu$, and define γ and u_λ as in (2.1). Then (2.2) holds with $\sigma = 0$. In addition,

$$\lim_{\lambda\to\infty}\int_{|x|\le\varepsilon}\phi_\lambda(x)dx = \lim_{\lambda\to\infty}\int_{|x|\le\varepsilon\lambda^\gamma}\lambda^{\gamma(\mu-n)}\phi(x)dx = \infty \quad \forall\,\varepsilon > 0.$$

Hence, for each $c > 0$, there exists a sequence $\{\psi^c_\lambda\}_{\lambda\ge1}$ such that $0 \le \psi_\lambda \le \phi_\lambda$ for all $\lambda \ge 1$ and $\{\frac{1}{c}\psi_\lambda\}_{\lambda>1}$ is a δ-sequence. Consequently, the solution u^c_λ of $\mathcal{L}u_\lambda = -u^q_\lambda$ with initial value $u^c_\lambda(x,0) = \psi^c_\lambda$ tends to the fundamental solution u^c of $\mathcal{L}u_\lambda = -u^q_\lambda$ with initial mass c as $\lambda \to \infty$; see [5]. Since $u_\lambda \ge u^c_\lambda$, we have

$$\liminf_{\lambda\to\infty} u_\lambda \ge \lim_{\lambda\to\infty} u^c_\lambda = u^c.$$

It follows that

$$\liminf_{\lambda\to\infty} u_\lambda(x,t) \ge \lim_{c\to\infty} u^c(x,t) = W_0(x,t).$$

A direct translation of the above relation in terms of $w(y,\tau)$ is exactly (3.7). Thus

$$\lim_{\tau\to\infty} w(y,\tau) = w_0(|y|) = W_0(y,1)$$

uniformly in any compact subset of $\mathbb{R}^n$. Since for $A = 1$ and some large R, $w(y,\tau) \le w^+_{A,R}(|y|)$ for all $\tau > 0$ and $w^+_{1,R}$ decays to zero as $|y| \to \infty$, the above limit is uniform in $\mathbb{R}^n$. This completes the proof of Theorem 1.6 for the case in which $A = 0$ and $1 < q < q^*$.

If $q \ge q^*$, then

$$\limsup_{\tau\to\infty} \|w(y,\tau)\|_{L^\infty(\mathbb{R}^n)} \le \lim_{A\to0+} \|w_A(|y|)\|_{L^\infty(\mathbb{R}^n)} = 0.$$

Finally, we prove the second limit in (1.10), where $q = q^*$. By taking smaller ϕ if necessary, we can assume that $\phi \le 1$ and has compact support. Then $u \le 1$ in $\mathbb{R}^n \times (0,\infty)$. Consequently, $\mathcal{L}u = -u^{q^*} \ge -u^{q^*-\varepsilon}$. Now let u^ε be the solution to $\mathcal{L}u^\varepsilon = -u^{q^*-\varepsilon}$ and initial data $u^\varepsilon(x,0) = \phi$. Then, from what we just proved, we have $\lim_{t\to\infty} t^{1/(q-\varepsilon-1)}u^\varepsilon(0,t) = W_0^{q^*-\varepsilon}(0,1)$, where $W_0^{q^*-\varepsilon}$ is the VSS for $q = q^* - \varepsilon$. Thus $\liminf_{t\to\infty} t^{1/(q^*-\varepsilon-1)}u(0,t) > 0$. As ε is arbitrary, the second limit in (1.10) follows. This completes the proof of Theorem 1.6. □

REFERENCES

[1] Y. Z. Chen and E. DiBenedetto, *On the local behavior of solutions of singular parabolic equations*, Arch. Ration. Mech. Anal., 103 (1988), pp. 319–346.

[2] Y. Z. Chen and E. DiBenedetto, *Holder estimates for solutions of singular parabolic equations with measurable coefficients*, Arch. Ration. Mech. Anal., 118 (1992), pp. 257–271.

[3] X. Chen, Y. Qi, and M. Wang, *Self-similar very singular solutions of the parabolic p-Laplacian*, J. Differential Equations, 190 (2003), pp. 1–15.

[4] X. Chen, Y. Qi, and M. Wang, *Existence and Uniqueness of Singular Solutions of a Fast Diffusion Porous Medium Equation*, preprint, Department of Mathematics, HKUST, Hong Kong, 1998.

[5] X. Chen, Y. Qi, and M. Wang, *Singular Solutions of Parabolic p-Laplacian with Absorption*, preprint, Department of Mathematics, HKUST, Hong Kong, 1998.

[6] E. DiBenedetto, *Degenerate Parabolic Equations*, Springer-Verlag, New York, 1993.

[7] E. DiBenedetto and M. Herrero, *Non-negative solutions of the evolution p-Laplacian equation. Initial traces and Cauchy problem when* $1 < p < 2$, Arch. Ration. Mech. Anal., 111 (1990), pp. 225–290.

[8] J. I. Diaz and J. E. Saa, *Uniqueness of very singular self-similar solution of a quasilinear degenerate parabolic with absorption*, Publ. Mat., 36 (1992), pp. 19–38.
[9] M. Escobedo and O. Kavian, *Variational problems related to self-similar solutions of the heat equation*, Nonlinear Anal., 11 (1987), pp. 1103–1133.
[10] M. Escobedo, O. Kavian, and H. Matano, *Large time behavior of solutions of a dissipative semilinear heat equation*, Comm. Partial Differential Equations, 20 (1995), pp. 1427–1452.
[11] V. A. Galaktionov, S. P. Kurdyumov, and A. A. Samarskii, *On asymptotic "eigenfunctions" of the Cauchy problem for a non-linear parabolic equation*, Math. USSR-Sb., 54 (1986), pp. 421–455.
[12] A. Gmira and L. Veron, *Large time behavior of the solutions of a semilinear parabolic equation in R^n*, J. Differential Equations, 53 (1984), pp. 258–276.
[13] L. Herraiz, *Asymptotic behaviour of solutions of some semilinear parabolic problems*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 16 (1999), pp. 49–105.
[14] S. Kamin and L.A. Peletier, *Singular solutions of the heat equation with absorption*, Proc. Amer. Math. Soc., 95 (1985), pp. 205–210.
[15] S. Kamin and L. A. Peletier, *Large time behavior of the solutions of the heat equation with absorption*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 12 (1985), pp. 393–408.
[16] S. Kamin and L. A. Peletier, *Large time behavior of the solutions of the porous media equation with absorption*, Israel J. Math., 55 (1986), pp. 129–146.
[17] S. Kamin and J. Vasquez, *Singular solutions of some nonlinear parabolic equations*, J. Anal. Math., 59 (1992), pp. 51–74.
[18] M. Kwak, *A porous media equation with absorption* I. *Long time behaviour*, J. Math. Anal. Appl., 223 (1998), pp. 96–110.
[19] G. Leoni, *A very singular solution for the porous media equation $u_t = \triangle(u^m) - u^p$ when $0 < m < 1$*, J. Differential Equations, 132 (1996), pp. 353–376.
[20] L. A. Peletier and J. Wang, *A very singular solution of a quasilinear degenerate diffusion equation with absorption*, Trans. Amer. Math. Soc., 307 (1988), pp. 813–826.
[21] L. A. Peletier and J. Zhao, *Large time behavior of the solutions of the porous media equation with absorption: The fast diffusion case*, Nonlinear Anal., 17 (1991), pp. 991–1009.
[22] Y. Qi and M. Wang, *The global existence and finite time extinction of a quasi-linear parabolic equation*, Adv. Differential Equations, 4 (1999), pp. 731–753.
[23] J. Zhao, *The Asymptotic Behavior of solutions of a quasilinear degenerate parabolic equation*, J. Differential Equations, 102 (1993), pp. 33–52.

SIAM J. MATH. ANAL.
Vol. 35, No. 1, pp. 135–147

PERTURBATION THEORY FOR VISCOSITY SOLUTIONS OF HAMILTON–JACOBI EQUATIONS AND STABILITY OF AUBRY–MATHER SETS*

DIOGO AGUIAR GOMES†

Abstract. In this paper we study the stability of integrable Hamiltonian systems under small perturbations, proving a weak form of the KAM/Nekhoroshev theory for viscosity solutions of Hamilton–Jacobi equations. The main advantage of our approach is that only a finite number of terms in an asymptotic expansion are needed in order to obtain uniform control. Therefore there are no convergence issues involved. An application of these results is to show that Diophantine invariant tori and Aubry–Mather sets are stable under small perturbations.

Key words. Hamiltonian dynamics, KAM theory, Aubry–Mather sets, viscosity solutions

AMS subject classifications. 37J40, 49L25, 35C20

DOI. 10.1137/S0036141002405960

1. Introduction.

We consider Hamiltonians of the form

$$H_\epsilon(p,x) = H_0(p) + \epsilon H_1(p,x), \tag{1}$$

with H_0, H_1 smooth, $H_0(p)$ strictly convex and $H_1(p,x)$ bounded with bounded derivatives, and $\mathbb{Z}^n$ periodic in x. The objective of this paper is to understand the dependence on ϵ of periodic viscosity solutions (for the definition of viscosity solution, see section 3) of

$$H_\epsilon(P + D_x u^\epsilon, x) = \overline{H}_\epsilon(P), \tag{2}$$

and prove stability of Aubry–Mather sets [Mat89a], [Mat89b], [Mat91], [Mn92], [Mn96] under small perturbations. Equation (2) has two unknowns, the function u^ϵ and the value of the effective Hamiltonian $\overline{H}_\epsilon(P)$. Given a viscosity solution of (2), the Mather set is a set invariant under the Hamiltonian dynamics

$$\dot{x} = -D_p H_\epsilon(p,x), \qquad \dot{p} = D_x H_\epsilon(p,x)$$

that is contained on the graph $(x,p) = (x, P + D_x u^\epsilon(x))$. For smooth solutions of (2) it corresponds to KAM tori.

We are given a reference value $P = P_0$, and we assume that for $\epsilon = 0$ the rotation vector $\omega_0 = D_P \overline{H}_0(P_0)$ satisfies Diophantine nonresonance conditions

$$|\omega_0 \cdot k| \geq \frac{C}{|k|^s} \tag{3}$$

for some positive constant C and some real $s > 0$. Is well known that the KAM theory applies to all nonresonant vectors ω_0. In particular it implies that for sufficiently

*Received by the editors April 19, 2002; accepted for publication (in revised form) December 9, 2002; published electronically June 10, 2003. This work was supported in part by FCT (Portugal) through programs POCTI, POCTI/32931/MAT/2000, and BPD 1531/2000.
http://www.siam.org/journals/sima/35-1/40596.html
†Departamento de Matemática, Instituto Superior Técnico, Avenida Rovisco Pais, 1049-001 Lisboa, Portugal (dgomes@math.ist.utl.pt).

small perturbations the solution to (2) is smooth. However, our results hold even if the solution to (2) fails to be smooth.

In general, if one keeps the momentum P_0 fixed, for $\epsilon > 0$ the new rotation vector $D_P\overline{H}_\epsilon(P_0)$ may fail to be Diophantine. In particular an ergodic flow for $\epsilon = 0$ may give rise to periodic orbits when $\epsilon > 0$. Therefore it is convenient to let the momentum P change with ϵ while keeping the rotation vector fixed. To that effect, given $N > 0$ we construct an approximate solution $\tilde{u}^\epsilon_N(x, P)$ of (2); that is, $\tilde{u}^\epsilon_N(x, P)$ solves

$$H_\epsilon(P + D_x\tilde{u}^\epsilon, x) = \tilde{H}^N_\epsilon(P) + O(\epsilon^N + |P - P_0|^N)$$

for all P close to P_0. Using the implicit function theorem, we can prove that there exists a new momentum $P_\epsilon = P_0 + O(\epsilon)$ such that $D_P\tilde{H}^N_\epsilon(P_\epsilon) = D_P\overline{H}_0(P_0)$. Therefore we are able to keep the rotation vector fixed (up to order N) under small perturbations.

The two main results of this paper are stated in Theorems 3 and 4. In Theorem 3, we show that for any $M > 0$ there exists $N(M)$ such that

$$|u^\epsilon(x, P_\epsilon) - \tilde{u}^\epsilon_N(x, P_\epsilon)| = O(\epsilon^M).$$

To show this, we first prove estimates along trajectories of the Hamiltonian flow. Then we extend these estimates to nearby points using a priori Lipschitz continuity of viscosity solutions. To that effect, we use the fact that the rotation vector is kept unchanged so that we can take advantage of the Diophantine properties. These properties imply that the Hamiltonian flow takes at most time $O\left(\frac{1}{\epsilon^{Ms}}\right)$ to visit an arbitrary ϵ^M neighborhood of any point of the torus [BGW98] (similar estimates were proved originally in [Dum91], [DDG96]). From this we can extend the estimates along trajectories to ϵ^M neighborhoods, using the Lipschitz continuity of viscosity solutions, and therefore obtain uniform control. The technique of the proof would break down if we choose $P_\epsilon = P_0$ because we do not have any control over the number of theoretical properties of $D_P\overline{H}_\epsilon(P_0)$.

Finally, in Theorem 4, we prove estimates for the derivatives of viscosity solutions, showing that

$$\operatorname{esssup} |D_x u^\epsilon - D_x\tilde{u}^\epsilon_N| = O(\epsilon^{M/2}).$$

Note that these estimates do not imply that u^ϵ is differentiable; therefore they apply even when the invariant tori that exist when $\epsilon = 0$ cease to exist and are replaced by Mather sets. Therefore this result implies the stability of Mather sets under small perturbations.

Our results should be compared with previous results. One result is KAM theory, which states that, for all Diophantine rotation vectors, u^ϵ is smooth and can be computed by summing a convergent series for ϵ small enough. A result that is somewhat close to ours is the following (see [BGGM98], [MG95]): Suppose P_0 is such that the corresponding frequency is Diophantine. Then there exists a canonical transformation T_ϵ, defined in the domain $|P - P_0| \leq C\epsilon$, such that in the new coordinates, the new Hamiltonian can be written as

$$H_0(P) + \tilde{H}_\epsilon(P) + e^{-C/\epsilon 1/n}R(x, P);$$

that is, the system is very close to integrable in an appropriate coordinate system. At the other extreme one has that the viscosity solution u^ϵ, which is Lipschitz, converges uniformly to a constant, provided the rotation number is nonresonant ($\omega_0 \cdot k \neq 0$ for all

$k \in \mathbb{Z}^n$) [Gom02]. The results in this paper apply to all Diophantine rotation vectors, provided the Hamiltonian is smooth enough so that one can construct $\tilde{u}_N^\epsilon$ to all orders, and yields an asymptotic representation, for small ϵ, of u^ϵ and its derivatives, even when there is not smoothness or no convergence of the approximations.

This paper is organized as follows: In section 2 we review the classical Lindstedt series expansion for perturbations of nonresonant integrable Hamiltonians. In section 3 we recall the necessary background from the theory of viscosity solutions and its relation to Aubry–Mather theory. In section 4 we study the expansions for $\overline{H}_\epsilon$. Uniform estimates for viscosity solutions are discussed in section 5. In the last section we present some applications to the stability of viscosity solutions and Aubry–Mather sets, bootstrapping from L^∞ estimates for viscosity solutions to $W^{1,\infty}$ estimates.

Other examples of asymptotic expansions for Hamilton–Jacobi equations in the case in which the perturbation is a second order elliptic operator can be found in [FS86a], [FS86b], and [Bes].

2. Classical perturbation theory. In this section we review the classical perturbation theory for Hamiltonian systems using a construction equivalent to the Poincaré normal form near an invariant torus. Somewhat incorrectly, but following [AKN97], we call it the Lindstedt series method. Although these results are fairly standard (see [AKN97], for instance), we present them in a more convenient form for our purposes.

Consider the Hamiltonian dynamics

$$\begin{cases} \dot{\mathbf{x}} = -D_p H_\epsilon(\mathbf{p}, \mathbf{x}), \\ \dot{\mathbf{p}} = D_x H_\epsilon(\mathbf{p}, \mathbf{x}). \end{cases} \tag{4}$$

We use the convention that boldface $(\mathbf{x}, \mathbf{p})$ are trajectories of the Hamiltonian flow and not the coordinates (x, p). The Hamilton–Jacobi integrability theory suggests that we should look for functions $\overline{H}_\epsilon(P)$ and $u^\epsilon(x, P)$, periodic in x, solving the Hamilton–Jacobi equation

$$H_\epsilon(P + D_x u^\epsilon, x) = \overline{H}_\epsilon(P). \tag{5}$$

Then, by performing the change of coordinates $(x, p) \leftrightarrow (X, P)$ determined by

$$\begin{cases} X = x + D_P u^\epsilon, \\ p = P + D_x u^\epsilon, \end{cases} \tag{6}$$

the dynamics (4) is simplified to

$$\begin{cases} \dot{\mathbf{X}} = -D_P \overline{H}(\mathbf{P}), \\ \dot{\mathbf{P}} = 0. \end{cases}$$

We again use the convention that boldface $(\mathbf{X}, \mathbf{P})$ are trajectories of the Hamiltonian flow and not the new coordinates (X, P).

If $\tilde{u}$ is an approximate solution to (5) satisfying

$$H_\epsilon(P + D_x \tilde{u}, x) = \overline{H}_\epsilon(P) + f(x, P), \tag{7}$$

then the change of coordinates (6) transforms (4) into

$$\begin{cases} \dot{\mathbf{X}} = -D_P \overline{H}_\epsilon(\mathbf{P}) - D_P f(\mathbf{X}, \mathbf{P}), \\ \dot{\mathbf{P}} = D_X f(\mathbf{X}, \mathbf{P}), \end{cases} \tag{8}$$

with the convention that $f(X,P) = f(x(X,P),P)$.

The KAM theory deals with constructing solutions of (5) by using an iterative procedure, a modified Newton's method, that yields an expansion

$$u^\epsilon = u_0 + \epsilon v_1 + \epsilon^2 v_2 \cdots .$$

The main technical point in KAM theory is to prove the convergence of these expansions. An alternate method that yields such an expansion is the Lindstedt series [AKN97]. However, we should point out that whereas the KAM expansion is a convergent one, the Lindstedt series may fail to converge. Nevertheless, since we will need only finitely many terms, we will use a variation of the Lindstedt series that we describe next.

We say that a vector $\omega \in \mathbb{R}^n$ is Diophantine if for all $k \in \mathbb{Z}^n \backslash \{0\}$, $|\omega \cdot k| \geq \frac{C}{|k|^s}$ for some $C, s > 0$. Let P_0 be such that $\omega_0 = D_P \overline{H}_0(P_0)$ is Diophantine. We look for an approximate solution of

$$H_\epsilon(P + D_x u^\epsilon(x,P), x) = \overline{H}_\epsilon(P),$$

valid for $P = P_0 + O(\epsilon)$. When $\epsilon = 0$, $\overline{H}_0(P) = H_0(P)$ and the solution u^0 is constant; for instance, we may take $u^0 \equiv 0$. For $\epsilon > 0$ we have, formally, $u^\epsilon = O(\epsilon)$, and so we suggest the following approximation $\tilde{u}^\epsilon_N$ to u^ϵ:

$$\begin{aligned} \tilde{u}^\epsilon_N = {} & \epsilon v_1(x,P_0) + \epsilon (P-P_0) D_P v_1(x,P_0) + \epsilon^2 v_2(x,P_0) \\ & + \frac{1}{2}\epsilon (P-P_0)^2 D^2_{PP} v_1(x,P_0) + \epsilon^2 (P-P_0) D_P v_2(x,P_0) \\ & + \epsilon^3 v_3(x,P_0) + \cdots . \end{aligned} \tag{9}$$

This expansion is carried out up to order $N-1$ in such a way that, formally, $u^\epsilon - \tilde{u}^\epsilon_N = O(\epsilon^N)$. For example,

$$\tilde{u}^\epsilon_1 = 0, \quad \tilde{u}^\epsilon_2 = \epsilon v_1, \quad \tilde{u}^\epsilon_3 = \epsilon v_1 + \epsilon^2 v_2 + \epsilon (P-P_0) D_P v_1.$$

The functions v_i and $D^k_{P^k} v_i$ satisfy transport equations

$$D_p H_0(P_0) D_x w = f(\cdots)$$

for some suitable f and can be solved inductively. For instance,

$$\overline{H}_1(P_0) = D_p H_0(P_0) D_x v_1 + H_1(P_0, x),$$

$$D_P \overline{H}_1(P_0) = D_p H_0(P_0) D_x (D_P v_1) + D^2_{pp} H_0(P_0) D_x v_1 + D_p H_1(P_0, x),$$

and

$$\overline{H}_2(P_0) = D_p H_0(P_0) D_x v_2 + \frac{1}{2} D^2_{pp} H_0(P_0) D_x v_1 D_x v_1 + D_p H_1(P_0, x) D_x v_1.$$

Note that the derivatives of v_i with respect to P and $D^k_{P^k} v_i$ are computed by solving appropriate transport equations, as illustrated above for $D_P v_1$, and not by differentiating v_i. In fact v_i may not be defined for $P \neq P_0$. However, if its derivative exists, it satisfies a transport equation.

The constants $\overline{H}_1(P_0), D_P\overline{H}_1(P_0), \overline{H}_2(P_0), \dots$ are uniquely determined by integral compatibility conditions; for example,

$$\overline{H}_1(P_0) = \int H_1(P_0, x)dx,$$

$$D_P\overline{H}(P_0) = \int D_p H_1(P_0, x)dx,$$

and

$$\overline{H}_2(P_0) = \int \frac{1}{2} D^2_{pp} H_0(P_0) D_x v_1 D_x v_1 + D_p H_1(P_0, x) D_x v_1 dx.$$

If H is sufficiently smooth and ω_0 is nonresonant, then these equations have smooth solutions that are unique up to constants. Finally one can check that

$$H_\epsilon(P + D_x\tilde{u}^\epsilon_N, x) = \tilde{H}^N_\epsilon(P) + O(\epsilon^N + |P - P_0|^N), \tag{10}$$

with

$$\tilde{H}^N_\epsilon(P) = \overline{H}_0(P_0) + \epsilon\overline{H}_1(P_0) + (P - P_0)D_P\overline{H}_0(P_0) + \epsilon^2\overline{H}_2(P_0) + \cdots,$$

and this expansion is carried up to order $N - 1$ in such a way that, formally,

$$\overline{H}_\epsilon(P) = \tilde{H}^N_\epsilon(P) + O(\epsilon^N + |P - P_0|^N).$$

Consider the change of coordinates

$$\begin{cases} p = P + D_x\tilde{u}^\epsilon_N(x, P), \\ X = x + D_P\tilde{u}^\epsilon_N(x, P). \end{cases}$$

Then, by (7) and (8), (4) is transformed into

$$\begin{cases} \dot{\mathbf{X}} = -D_P\overline{H}_\epsilon(\mathbf{P}) + O(\epsilon^N + |\mathbf{P} - P_0|^{N-1}), \\ \dot{\mathbf{P}} = O(\epsilon^N + |\mathbf{P} - P_0|^N). \end{cases}$$

3. Viscosity solutions, optimal control, and Mather measures. In general (5) does not admit smooth classical solutions. However, (5) has viscosity solutions which are known to be the appropriate notion of weak solution for Hamilton–Jacobi equations. In this section we review the necessary background, and in the rest of the paper we extend rigorously the classical perturbation procedure from the previous section to viscosity solutions.

For our purposes, a convenient definition of viscosity solution is the following: We say that a function u is a viscosity solution of (5), provided that it satisfies the fixed point identity

$$u(x) = \inf \int_0^t L(\mathbf{x}, \dot{\mathbf{x}}) + P\dot{\mathbf{x}} + \overline{H}(P)dt + u(\mathbf{x}(t)), \tag{11}$$

in which the infimum is taken over Lipschitz trajectories $\mathbf{x}(\cdot)$, with initial condition $\mathbf{x}(0) = x$, and $\overline{H}$ is the unique number for which (11) holds with $u(x)$ bounded. The Lagrangian L is the Legendre transform of the Hamiltonian

$$L(x, v) = \sup_p -vp - H(p, x).$$

This definition of viscosity solution is equivalent to the standard one [FS93], [BCD97], as long as the Hamiltonian and Lagrangian L are strictly convex. Note that viscosity solutions do not have to be smooth. However, they are semiconcave and therefore twice differentiable almost everywhere. Furthermore they are differentiable along the optimal trajectory $\mathbf{x}(t)$. In fact, the optimal trajectory $\mathbf{x}(t)$ in (11) and the momentum $\mathbf{p}(t) = P + D_x u(\mathbf{x}(t))$ are solutions of (4) for all $t > 0$.

First, let us quote an existence result [LPV88].

THEOREM 1 (Lions, Papanicolao, and Varadhan [LPV88]). *For each P there exists a number $\overline{H}(P)$ and a function $u(x, P)$, periodic in x, that solves* (5) *in the viscosity sense. Furthermore $\overline{H}(P)$ is convex in P, and $u(x, P)$ is Lipschitz in x.*

This theorem does not assert anything about uniqueness of the viscosity solution u. Indeed, such viscosity solutions are not unique even up to constants; see, for instance, [Con95]. However, as it was shown in [Gom02], under certain hypotheses one can prove uniqueness and even continuity of the viscosity solution u with respect to parameters. These hypotheses can be formulated in terms of the ergodic properties of certain measures—Mather measures (see Theorem 2)—that are invariant under the Hamiltonian dynamics.

The connection between classical mechanics and viscosity solutions is well known and was explored by several authors, for instance, [Fat97a], [Fat97b], [Fat98a], [Fat98b], [E99], [EG01], [EG02], and [Gom00b]. One of the most basic results is the following.

THEOREM 2 (A. Fathi, W. E). *Let u be a viscosity solution of* (5).

- *For each P there exists a set invariant under the dynamics* (4) *contained in the graph $(x, P + D_x u)$.*
- *There exists a probability measure $\mu(x, p)$ (Mather measure) invariant under* (4) *supported on this invariant set.*
- *This measure minimizes*

$$\int L(x, v) + Pv d\mu, \tag{12}$$

with $v = -D_p H(p, x)$, over all probability measures that are invariant under (4).

Conversely, any probability measure invariant under (4) *that minimizes* (12) *is supported on the graph $(x, P + D_x u)$ for any viscosity solution u of* (5).

One of the main advantages of the previous theorem is that one can translate properties of viscosity solutions into properties of Mather sets or measures and vice versa. Some properties of viscosity solutions are described in the following proposition.

PROPOSITION 1. *Suppose (x, p) is a point in the graph*

$$\mathcal{G} = \{(x, P + D_x u(x)) : \ u \text{ is differentiable at } x\}.$$

Then for all $t > 0$ the solution $(\mathbf{x}(t), \mathbf{p}(t))$ of (4) *with initial conditions (x, p) belongs to $\mathcal{G}$.*

Proof. The invariance of the graph for $t > 0$ is a consequence of the optimal control interpretation of viscosity solutions [FS93], and the reader may find a proof, for instance, in [Gom00b] or [Gom00a]. □

Finally, the following is an important identity [CIPP98].

PROPOSITION 2 (Contreras et al. [CIPP98]).

$$\overline{H}(P) = \inf_{\phi} \sup_{x} H(P + D_x \phi, x), \tag{13}$$

in which the infimum is taken over C^1 periodic functions ϕ.

This formula can be used to compute $\overline{H}(P)$ effectively [GO02a] and, in conjunction with properties of viscosity solutions, to detect nonintegrability of Hamiltonian systems [GO02b].

4. Estimates for the effective Hamiltonian. We start this section by proving that $\tilde{H}^N_\epsilon$ is an asymptotic expansion to $\overline{H}_\epsilon$. Then we will use convexity techniques to prove estimates for the derivatives of viscosity solutions that we will use in the subsequent sections.

PROPOSITION 3. *Suppose we can construct an approximate solution as in* (9). *Then*

$$\overline{H}_\epsilon(P) \le \tilde{H}^N_\epsilon(P) + O(\epsilon^N + |P - P_0|^N). \tag{14}$$

Remark. Note that the error term is a function of x, but, by periodicity, it can be estimated uniformly by $O(\epsilon^N + |P - P_0|^N)$.

Proof. The inf sup formula (13) implies

$$\overline{H}_\epsilon(P) \le \sup_x H_\epsilon(P + D_x\tilde{u}^\epsilon_N, x).$$

By expanding this expression in Taylor series and taking the supremum, we obtain the result. □

A converse inequality is also true.

PROPOSITION 4. *At any point at which $D_x u^\epsilon$ exists we have*

$$\begin{aligned}\overline{H}_\epsilon(P) \ge{}& H_\epsilon(P + D_x\tilde{u}^\epsilon_N, x) + D_p H_\epsilon(P + D_x\tilde{u}^\epsilon_N, x)(D_x\tilde{u}^\epsilon_N - D_x u^\epsilon)\\ &+ \frac{\gamma}{2}|D_x\tilde{u}^\epsilon_N - D_x u^\epsilon|^2\end{aligned} \tag{15}$$

for some positive constant $\gamma > 0$.

Proof. Since $H_\epsilon(p, x)$ is strictly convex, there exists a constant $\gamma > 0$ such that

$$\begin{aligned}\overline{H}_\epsilon(P) &\ge H_\epsilon(P + D_x u^\epsilon, x)\\ &\ge H_\epsilon(P + D_x\tilde{u}^\epsilon_N, x) + D_p H_\epsilon(P + D_x\tilde{u}^\epsilon_N, x)(D_x\tilde{u}^\epsilon_N - D_x u^\epsilon)\\ &\quad + \frac{\gamma}{2}|D_x\tilde{u}^\epsilon_N - D_x u^\epsilon|^2. \qquad \square\end{aligned}$$

The following corollary is going to be used in the next section.

COROLLARY 1. *If there exists an approximate solution as in Proposition* 3, *then there exists a point x_0 for which*

$$|D_x\tilde{u}^\epsilon(x_0) - D_x u^\epsilon(x_0)| \le C\epsilon^{N/2}.$$

Proof. Since u^ϵ is semiconcave, $\tilde{u}^\epsilon - u^\epsilon$ is semiconvex. Therefore at the maximum x_0, $D_x(\tilde{u}^\epsilon - u^\epsilon) = 0$; that is, the derivative exists and is zero. Then the two previous propositions yield the desired result. □

COROLLARY 2. *If there exists an approximate solution as in Proposition* 3, *then*

$$\overline{H}_\epsilon = \tilde{H}^N_\epsilon + O(\epsilon^N).$$

Proof. It suffices to combine Propositions 3 and 4 at the point x_0 given by the previous corollary. □

5. Uniform estimates. In this section we prove uniform estimates

$$\sup_x |u^\epsilon - \tilde{u}_N^\epsilon| = O(\epsilon^M)$$

under Diophantine conditions on the rotation vector ω_0 of the unperturbed problem. These results should be compared with the ones in [Gom02], which show that unique ergodicity of the Mather measure implies uniform continuity of the viscosity solution with respect to the parameters.

First we construct an approximate solution $\tilde{u}_N^\epsilon$ for a sufficiently large but finite N using the Lindstedt series expansion described in section 2. Then, by changing coordinates, we show that for long times there are uniform estimates along trajectories. Finally, by using Lipschitz estimates for u^ϵ and $\tilde{u}_N^\epsilon$ and the results by [Dum91] on ergodization times, we extend them to whole space.

THEOREM 3. *Suppose the rotation vector*

$$\omega_0 = D_P \overline{H}_0(P_0)$$

satisfies the Diophantine property (3). *Furthermore suppose that for every N there exists $P_\epsilon = P + O(\epsilon)$,*

$$D_P \tilde{H}_\epsilon^N(P_\epsilon) = \omega_0;$$

that is, the approximate rotation vector corresponding to P_ϵ is the original rotation vector ω_0. Let u^ϵ be a solution of

$$H_\epsilon(P_\epsilon + D_x u^\epsilon, x) = \overline{H}_\epsilon(P_\epsilon),$$

and let $\tilde{u}_N^\epsilon$ be the corresponding approximate solution using a Lindstedt series expansion up to order N.

Then for every M there exists $N(M)$ such that

$$\sup_x |u^\epsilon - \tilde{u}_N^\epsilon| = O(\epsilon^M).$$

Proof. Define P_ϵ by solving the equation

$$\omega_0 = D_P \tilde{H}_\epsilon^N(P_\epsilon),$$

that is,

$$\omega_0 = D_P \overline{H}_0(P_0) + \epsilon D_P \overline{H}_1(P_0) + (P_\epsilon - P_0) D_{PP}^2 \overline{H}_0(P_0) + \cdots,$$

with expansion taken up to order $N-1$. Under the nondegeneracy condition $\det D_{pp}^2 \overline{H}_0 \neq 0$ (which holds because $\overline{H}_0(P) = H_0(P)$ is strictly convex), the implicit function theorem yields a unique solution of the form

$$P_\epsilon = P_0 + \epsilon P_1 + \cdots$$

with $P_1 = -\left[D_{PP}^2 \overline{H}_0(P_0)\right]^{-1} D_P \overline{H}_1(P_0)$.

Define the new coordinates (P, X) by

$$\begin{cases} p = P + D_x \tilde{u}_N^\epsilon(x, P), \\ X = x + D_P \tilde{u}_N^\epsilon(x, P). \end{cases}$$

To simplify notation we denote $X = \phi_\epsilon(x)$.

Let x_0 be the point given by Corollary 1. Set

$$(\mathbf{x}(0), \mathbf{p}(0)) = (x_0, P_\epsilon + D_x u^\epsilon(x_0))$$

as the initial conditions for a trajectory $(\mathbf{x}(t), \mathbf{p}(t))$ of (4). Since $|D_x \tilde{u}^\epsilon(x_0) - D_x u^\epsilon(x_0)| \leq C\epsilon^{N/2}$, in the new coordinates we have

$$\mathbf{P}(0) = P_\epsilon + O(\epsilon^{N/2}).$$

Consider the Hamiltonian dynamics in the new coordinates (X, P), with initial condition $(\mathbf{X}(0), \mathbf{P}(0))$ (the value $\mathbf{X}(0)$ is not important):

$$\begin{cases} \dot{\mathbf{X}} = -D_P \tilde{H}_\epsilon(\mathbf{P}) + O(\epsilon^N + |\mathbf{P} - P_0|^{N-1}), \\ \dot{\mathbf{P}} = O(\epsilon^N + |\mathbf{P} - P_0|^N). \end{cases}$$

From this equation it follows that the momentum P in the new coordinates is conserved for long times.

Proposition 5.

$$\sup_{0 \leq t \leq \frac{1}{\epsilon^{N/2}}} |\mathbf{P}(t) - P_\epsilon| \leq O(\epsilon^{N/4}).$$

Proof. Note that

$$\begin{aligned} \frac{d}{dt}|\mathbf{P} - P_\epsilon|^2 &\leq C|\mathbf{P} - P_\epsilon|(\epsilon^N + |\mathbf{P} - P_\epsilon|^N) \\ &\leq C\epsilon^N |\mathbf{P} - P_\epsilon|^2 + C\epsilon^N \end{aligned}$$

as long as $|\mathbf{P} - P_\epsilon|^{N-1} \leq C\epsilon^N$. Note that for $N > 3$—and we are always assuming N large enough—we have $|\mathbf{P}(0) - P_\epsilon|^{N-1} \leq C\epsilon^N$. Thus the Gronwall inequality implies

$$|\mathbf{P}(T) - P_\epsilon|^2 \leq e^{C\epsilon^N T} \left(|\mathbf{P}(0) - P_\epsilon|^2 + C\epsilon^N T \right).$$

Therefore, up to $T = \frac{1}{\epsilon^{N/2}}$ we have

$$|\mathbf{P}(t) - P_\epsilon|^2 \leq C\epsilon^{N/2}$$

for ϵ sufficiently small. □

Observe that $\phi_\epsilon(x) = X$ is a diffeomorphism for small ϵ. Let

$$U(X) = u^\epsilon(\phi_\epsilon^{-1}(X)) - \tilde{u}_N^\epsilon(\phi_\epsilon^{-1}(X)),$$

in particular

$$U(\mathbf{X}(t)) = u^\epsilon(\mathbf{x}(t)) - \epsilon v_1(\mathbf{x}(t)) - \epsilon(P_\epsilon - P_0) D_P v_1(\mathbf{x}(t)) - \cdots.$$

Recall that $D_x u^\epsilon(\mathbf{x}(t)) = \mathbf{p}(t) - P_\epsilon$. Thus

$$\begin{aligned} \frac{d}{dt} U(\mathbf{X}(t)) &= (\mathbf{p}(t) - P_\epsilon) D_p H_\epsilon(\mathbf{p}(t), \mathbf{x}(t)) \\ &\quad - \epsilon D_x v_1(\mathbf{x}(t)) D_p H_\epsilon(\mathbf{p}(t), \mathbf{x}(t)) - \cdots \\ &= (\mathbf{p}(t) - P_\epsilon - \epsilon D_x v_1 - \cdots) D_p H_\epsilon \\ &= (\mathbf{P}(t) - P_\epsilon) D_p H_\epsilon(\mathbf{p}(t), \mathbf{x}(t)), \end{aligned}$$

since $\mathbf{p}(t) = \mathbf{P}(t) + \epsilon D_x v_1(\mathbf{x}(t)) + \cdots$. Therefore

$$\frac{d}{dt} U(\mathbf{X}(t)) = O(\epsilon^{N/4})$$

for $0 \leq t \leq \frac{1}{\epsilon^{N/2}}$.

We may also add a convenient constant to u^ϵ in such a way that $U(\mathbf{X}(0)) = 0$, and so we obtain

$$\sup_{0 \leq t \leq \frac{1}{\epsilon^{N/8}}} U(\mathbf{X}(t)) = O(\epsilon^{N/8})$$

along the trajectory.

Since, for small ϵ, ϕ_ϵ is a diffeomorphism, U is a Lipschitz function. The Diophantine property implies that the flow

$$\dot{\mathbf{X}} = D_P \overline{H}_0(P_0) + O(\epsilon^{N/2})$$

takes at most time $T = O\left(\frac{1}{\epsilon^{Ms}}\right)$ to get within distance ϵ^M of any point (see [BGW98], and also [Dum91], [DDG96]). Thus, if $M < \frac{N}{8s}$, we get for some $0 \leq t \leq \frac{1}{\epsilon^{N/8}}$ that

$$|X - \mathbf{X}(t)| \leq C\epsilon^M,$$

and so

$$|U(X)| \leq |U(X) - U(\mathbf{X}(t))| + |U(\mathbf{X}(t))| \leq C\epsilon^M.$$

Because ϕ_ϵ is a diffeomorphism, the same estimate carries over for the difference $u^\epsilon - \tilde{u}^\epsilon_N$. □

A final comment is that since $\tilde{u}^\epsilon_N - \tilde{u}^\epsilon_M = O(\epsilon^M)$, we also have

$$\sup_x |u^\epsilon - \tilde{u}^\epsilon_M| = O(\epsilon^M),$$

although we need the existence of $\tilde{u}^\epsilon_N$ to prove this estimate.

6. Applications: Stability of Mather sets and regularity. This last section is dedicated to proving estimates on the derivatives $D_x u^\epsilon - D_x \tilde{u}^\epsilon$. Such estimates rely on the uniform estimates from the previous section. Since the Mather sets are supported on the graph $(x, P + D_x u)$, estimates on the derivatives show the stability of Mather sets.

PROPOSITION 6. *Suppose $\omega_0 = D_P H_0(P_0)$ is Diophantine and (2) admits an approximate solution $\tilde{u}^\epsilon_N$. Then*

$$\begin{aligned} \frac{1}{T} \int_0^T \frac{\gamma}{2} |D_x u^\epsilon(\mathbf{x}(t)) - D_x \tilde{u}^\epsilon_N(\mathbf{x}(t))|^2 dt \\ \leq C\epsilon^N + \frac{2}{T} \sup_x |u^\epsilon - \tilde{u}^\epsilon_N|, \end{aligned}$$

in which the integral is taken along a trajectory $\mathbf{x}(\cdot)$ of

$$\dot{\mathbf{x}}(t) = -D_p H_\epsilon(P_\epsilon + D_x u^\epsilon(\mathbf{x}(t)), \mathbf{x}(t)).$$

Proof. The strict convexity of H_ϵ together with Corollary 2 implies

$$O(\epsilon^N) \geq D_p H_\epsilon(P_0 + D_x u^\epsilon(\mathbf{x}(t)), \mathbf{x}(t))(D_x \tilde{u}_N^\epsilon(\mathbf{x}(t)) - D_x u^\epsilon(\mathbf{x}(t))) + \frac{\gamma}{2}|D_x \tilde{u}_N^\epsilon(\mathbf{x}(t)) - D_x u^\epsilon(\mathbf{x}(t))|^2.$$

Integrating with respect to t and observing that

$$\begin{aligned}&\int_0^T D_p H_\epsilon(P_0 + D_x u^\epsilon(\mathbf{x}(t)), \mathbf{x}(t))(D_x \tilde{u}_N^\epsilon(\mathbf{x}(t)) - D_x u^\epsilon(\mathbf{x}(t)))dt \\ &= -\int_0^T \dot{\mathbf{x}}(t)(D_x \tilde{u}_N^\epsilon(\mathbf{x}(t)) - D_x\ u^\epsilon(\mathbf{x}(t)))dt \\ &= -u^\epsilon(\mathbf{x}(0)) + \tilde{u}_N^\epsilon(\mathbf{x}(0)) + u^\epsilon(\mathbf{x}(T)) - \tilde{u}_N^\epsilon(\mathbf{x}(T)),\end{aligned}$$

we obtain the result. □

This proposition is the key to proving the main result of this paper, which is discussed in the next theorem—pointwise estimates for first derivatives of viscosity solutions.

THEOREM 4. *Let $M > 0$. Suppose $\omega_0 = D_P H_0(P_0)$ is Diophantine such that the cell problem* (2) *admits an approximate solution $\tilde{u}_N^\epsilon$ for N sufficiently large such that Theorem* 3 *holds. Then*

$$\operatorname*{esssup}_x |D_x u^\epsilon - D_x \tilde{u}_N^\epsilon| \leq C\epsilon^{M/2}.$$

Proof. Since $\sup_x |\tilde{u}_N^\epsilon - u^\epsilon| = O(\epsilon^M)$ we have

$$\int_0^1 \frac{\gamma}{2}|D_x u^\epsilon(\mathbf{x}(t)) - D_x \tilde{u}_N^\epsilon(\mathbf{x}(t))|^2 dt \leq C\epsilon^M, \tag{16}$$

with $\mathbf{x}(t)$ as in the previous theorem and for any initial condition $\mathbf{x}(0) = x$.

Let G be the set of the points at which $D_x u^\epsilon$ exists and such that

$$|D_x u^\epsilon - D_x \tilde{u}_N^\epsilon| \leq C\epsilon^{M/2}$$

for some fixed constant C, and set

$$B = \{x \in G^c | u^\epsilon \text{ is differentiable at } x\}.$$

Since u^ϵ is Lipschitz, then $(B \cup G)^c$, which is the set of points of nondifferentiability of u^ϵ, is of zero Lebesgue measure.

Let x be a point for which $D_x u^\epsilon$ exists and $|D_x \tilde{u}_N^\epsilon - D_x u^\epsilon| > C\epsilon^{M/2}$. Define $p_x = P_\epsilon + D_x u^\epsilon(x)$. Let $(\mathbf{x}(t), \mathbf{p}(t))$ be the solution of (4) with initial conditions (x, p_x).

The estimate (16) implies that we may assume that there exists $0 < T < 1$ such that $\mathbf{x}(T) \in G$. Let $y = \mathbf{x}(T)$ and

$$\tilde{p}_y = P_\epsilon + D_x \tilde{u}_N^\epsilon(\mathbf{x}(T)), \quad p_y = P_\epsilon + D_x u^\epsilon(\mathbf{x}(T)).$$

Since $y \in G$ we have

$$|\tilde{p}_y - p_y| \leq C\epsilon^{M/2}.$$

Let $(\mathbf{x}(t), \tilde{\mathbf{p}}(t))$ be the solution of (4) with initial conditions $(y, \tilde{p}_y)$. Define $\tilde{x} = \mathbf{x}(-T)$, $p_{\tilde{x}} = \tilde{\mathbf{p}}(-T)$.

By standard ODE theory

$$|p_x - p_{\tilde{x}}| \leq C\epsilon^{M/2}, \qquad |x - \tilde{x}| \leq C\epsilon^{M/2}.$$

Note that

$$\begin{aligned}|D_x u^\epsilon(x) - D_x \tilde{u}_N^\epsilon(x)| \leq |P_\epsilon + D_x u^\epsilon(x) - p_{\tilde{x}}| + |p_{\tilde{x}} - P_\epsilon - D_x \tilde{u}_N^\epsilon(\tilde{x})| \\ + |D_x \tilde{u}_N^\epsilon(\tilde{x}) - D_x \tilde{u}_N^\epsilon(x)|.\end{aligned}$$

The first term is controlled by

$$|P_\epsilon + D_x u^\epsilon(x) - p_{\tilde{x}}| = |p_x - p_{\tilde{x}}| \leq C\epsilon^{M/2}.$$

The last term is controlled by the Lipschitz constant of $\tilde{u}_N^\epsilon$,

$$|D_x \tilde{u}_N^\epsilon(\tilde{x}) - D_x \tilde{u}_N^\epsilon(x)| \leq C|x - \tilde{x}| \leq C\epsilon^{M/2}.$$

Therefore it suffices to estimate $|p_{\tilde{x}} - P_\epsilon - D_x \tilde{u}_N^\epsilon(\tilde{x})|$. To that effect observe that from differentiating (10) with respect to x it follows that

$$D_p H(P_\epsilon + D_x \tilde{u}_N^\epsilon, x) D_{xx}^2 \tilde{u}_N^\epsilon + D_x H(P_\epsilon + D_x \tilde{u}_N^\epsilon, x) = O(\epsilon^N). \tag{17}$$

Then by combining

$$\begin{aligned}&\frac{d}{dt}\frac{|\tilde{\mathbf{p}}(t) - P_\epsilon - D_x \tilde{u}_N^\epsilon(\tilde{\mathbf{x}}(t)|^2}{2} \\ &\quad = (\tilde{\mathbf{p}}(t) - P_\epsilon - D_x \tilde{u}_N^\epsilon(\tilde{\mathbf{x}}(t)) \left[D_x H(\tilde{\mathbf{p}}(t), \tilde{\mathbf{x}}(t)) + D_{xx}^2 \tilde{u}_N^\epsilon(\tilde{\mathbf{x}}(t))\right] D_p H(\tilde{\mathbf{p}}(t), \tilde{\mathbf{x}}(t)))\end{aligned}$$

with (17), we obtain

$$\frac{d}{dt}\frac{|\tilde{\mathbf{p}}(t) - P_\epsilon - D_x \tilde{u}_N^\epsilon(\tilde{\mathbf{x}}(t)|^2}{2} \leq C|\tilde{\mathbf{p}}(t) - P_\epsilon - D_x \tilde{u}_N^\epsilon(\tilde{\mathbf{x}}(t))|^2 + O(\epsilon^N).$$

Then the Gronwall inequality yields

$$|p_{\tilde{x}} - P_\epsilon - D_x \tilde{u}_N^\epsilon(\tilde{x})|^2 \leq C\epsilon^N. \qquad \square$$

Acknowledgments. I would like to thank the referees, L. C. Evans, R. de la Llave, and C. Valls for many suggestions and comments on the original version of this paper.

REFERENCES

[AKN97] V. I. Arnold, V. V. Kozlov, and A. I. Neishtadt, *Mathematical Aspects of Classical and Celestial Mechanics*, Springer-Verlag, Berlin, 1997.

[BCD97] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Boston, MA, 1997.

[Bes] U. Bessi, *Smooth Approximation of Mather Sets*, preprint.

[BGGM98] F. Bonetto, G. Gallavotti, G. Gentile, and V. Mastropietro, *Lindstedt series, ultraviolet divergences and Moser's theorem*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 26 (1998), pp. 545–593.

[BGW98] J. Bourgain, F. Golse, and B. Wennberg, *On the distribution of free path lengths for the periodic Lorentz gas*, Comm. Math. Phys., 190 (1998), pp. 491–508.

[Con95] M. Concordel, *Periodic Homogenization of Hamilton–Jacobi Equations*, Ph.D. thesis, University of California, Berkeley, CA, 1995.

[CIPP98] G. Contreras, R. Iturriaga, G. P. Paternain, and M. Paternain, *Lagrangian graphs, minimizing measures and Mañé's critical values*, Geom. Funct. Anal., 8 (1998), pp. 788–809.

[Dum91] H. S. Dumas, *Ergodization rates for linear flow on the torus*, J. Dynam. Differential Equations, 3 (1991), pp. 593–610.

[DDG96] H. S. Dumas, L. Dumas, and F. Golse, *On the mean free path for a periodic array of spherical obstacles*, J. Statist. Phys., 82 (1996), pp. 1385–1407.

[E99] W. E, *Aubry-Mather theory and periodic solutions of the forced Burgers equation*, Comm. Pure Appl. Math., 52 (1999), pp. 811–828.

[EG01] L. C. Evans and D. Gomes, *Effective Hamiltonians and averaging for Hamiltonian dynamics.* I, Arch. Ration. Mech. Anal., 157 (2001), pp. 1–33.

[EG02] L. C. Evans and D. Gomes, *Effective Hamiltonians and averaging for Hamiltonian dynamics.* II, Arch. Ration. Mech. Anal., 161 (2002), pp. 271–305.

[Fat97a] A. Fathi, *Solutions KAM faibles conjuguées et barrières de Peierls*, C. R. Acad. Sci. Paris Sér. I Math., 325 (1997), pp. 649–652.

[Fat97b] A. Fathi, *Théorème KAM faible et théorie de Mather sur les systèmes lagrangiens*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 1043–1046.

[Fat98a] A. Fathi, *Orbite hétéroclines et ensemble de Peierls*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 1213–1216.

[Fat98b] A. Fathi, *Sur la convergence du semi-groupe de Lax-Oleinik*, C. R. Acad. Sci. Paris Sér. I Math., 327 (1998), pp. 267–270.

[FS93] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[FS86a] W. H. Fleming and P. E. Souganidis, *Asymptotic series and the method of vanishing viscosity*, Indiana Univ. Math. J., 35 (1986), pp. 425–447.

[FS86b] W. H. Fleming and P. E. Souganidis, *Erratum: "Asymptotic series and the method of vanishing viscosity,"* Indiana Univ. Math. J., 35 (1986), p. 925.

[Gom00a] D. Gomes, *Hamilton-Jacobi Equations, Viscosity Solutions and Asymptotics of Hamiltonian Systems*, Ph.D. Thesis, University of California, Berkeley, CA, 2000.

[Gom00b] D. Gomes, *Viscosity solutions of Hamilton-Jacobi equations, and asymptotics for Hamiltonian systems*, Calc. Var. Partial Differential Equations, 14 (2002), pp. 345–357.

[Gom02] D. Gomes, *Regularity theory for Hamilton-Jacobi equations*, J. Differential Equations, 187 (2003), pp. 359–374.

[GO02a] D. Gomes and A. Oberman, *Computing the Effective Hamiltonian—A Variational Approach to Homogenization*, preprint.

[GO02b] D. Gomes and A. Oberman, *Converse KAM Theory*, in preparation.

[LPV88] P. L. Lions, G. Papanicolao, and S. R. S. Varadhan, *Homogenization of Hamilton-Jacobi Equations*, manuscript.

[Mn92] R. Mañé, *On the minimizing measures of Lagrangian dynamical systems*, Nonlinearity, 5 (1992), pp. 623–638.

[Mn96] R. Mañé, *Generic properties and problems of minimizing measures of Lagrangian systems*, Nonlinearity, 9 (1996), pp. 273–310.

[Mat89a] J. N. Mather, *Minimal action measures for positive-definite Lagrangian systems*, in Proceedings of the Ninth International Congress on Mathematical Physics (Swansea, Wales, 1988), Hilger, Bristol, UK, 1989, pp. 466–468.

[Mat89b] J. N. Mather, *Minimal measures*, Comment. Math. Helv., 64 (1989), pp. 375–394.

[Mat91] J. N. Mather, *Action minimizing invariant measures for positive definite Lagrangian systems*, Math. Z., 207 (1991), pp. 169–207.

[MG95] A. Morbidelli and A. Giorgilli, *Superexponential stability of KAM tori*, J. Statist. Phys., 78 (1995), pp. 1607–1617.

SIAM J. MATH. ANAL.
Vol. 35, No. 1, pp. 148–159

UNIQUENESS IN IDENTIFICATION OF THE SUPPORT OF A SOURCE TERM IN AN ELLIPTIC EQUATION*

SUNGWHAN KIM† AND MASAHIRO YAMAMOTO†

Abstract. We consider an inverse problem of identifying the support D of a source term in an elliptic equation

$$-\Delta u(x) + q(x)\chi_D(x)u(x) = 0, \quad x \in \Omega, \qquad \text{and} \qquad u(x) = f(x), \quad x \in \partial\Omega.$$

Here q is a given positive function and χ_D is the characteristic function of a subdomain D such that $\overline{D} \subset \Omega$. In this paper, we prove the global uniqueness in this inverse problem within convex hulls of polygons D.

Key words. inverse problem, uniqueness, determination of supports

AMS subject classifications. 35R30, 35J25, 35B60

DOI. 10.1137/S0036141002412707

1. Introduction. We consider an inverse problem of recovering the shape and location of an unknown stationary heat source F. Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with smooth boundary and D a subdomain of Ω with Lipschitz boundary.

In this paper, we assume that the source F at $x = (x_1, x_2)$ is limited to D and proportional to the temperature u at x, that is, $F(x,t,u) = q(x)\chi_D(x)u(x,t)$. Here and henceforth, χ_D is the characteristic function of the subdomain $D \subset \Omega$, and $q \in C^2(\overline{\Omega})$, $q > 0$, on $\overline{\Omega}$.

If we apply a potential f to the boundary $\partial\Omega$ of Ω, then the resulting temperature u satisfies the Dirichlet problem

$$\begin{cases} -\Delta u + q\chi_D u = 0 & \text{in} \quad \Omega, \\ u = f & \text{on} \quad \partial\Omega. \end{cases} \tag{1.1}$$

It is well known that for a given domain D and $f \in H^{\frac{1}{2}}(\partial\Omega)$, there exists a unique solution $u \in H^1(\Omega)$ to (1.1). Thus we can define the Dirichlet-to-Neumann map $\Lambda_D : H^{\frac{1}{2}}(\partial\Omega) \to H^{-\frac{1}{2}}(\partial\Omega)$ by

$$\Lambda_D(f) := \frac{\partial u}{\partial \nu}|_{\partial\Omega}, \tag{1.2}$$

where ν is the unit outward normal vector to $\partial\Omega$.

Restricting D to a polygon such that $\overline{D} \subset \Omega$, we discuss an inverse problem of determining D by a single boundary measurement $(f, \Lambda_D(f))$.

There has been research related to our inverse problem, which is motivated by determination of transistor contact resistivity and contact window location in the equation $-\Delta u + \chi_D u = 0$ in Ω. See [3], [8], [12], [15]. In particular, a uniqueness result within a one-parameter monotone family from a one-point boundary measurement of the potential was obtained in [3]. Moreover [12] provides a global uniqueness result

*Received by the editors August 7, 2002; accepted for publication (in revised form) January 31, 2003; published electronically June 25, 2003.
http://www.siam.org/journals/sima/35-1/41270.html
†Graduate School of Mathematical Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro, Tokyo 153-8914, Japan (sungwhan@ms.u-tokyo.ac.jp, myama@ms.u-tokyo.ac.jp).

and a reconstruction scheme within the class of two- or three-dimensional balls from a single boundary measurement. In [15], the first author proved global uniqueness within some classes which can contain balls or ellipses. As for a general convex D, the uniqueness is still open (see, e.g., Problem 4.7.2 (p. 104) in [11]).

As for related inverse problems of determining piecewise continuous $\gamma = \gamma(x)$ in $\nabla \cdot (\gamma \nabla u) = 0$ in Ω, we can refer to [4], [5], [9], [13], [14], [17]. The case where $\gamma(x) = 1 + k\chi_D(x)$ and k is constant, in particular, has been studied by many researchers. The papers [5] and [17] dealt with the global uniqueness problem for polygons by one (or two) boundary measurement(s) under extra conditions. Our inverse problem is concerned with the determination of shapes of domains and is of a character similar to the classical inverse source problem or the inverse gravimetry where we are required to determine a domain D in $-\Delta u = q\chi_D$ by a single measurement of an exterior potential. As for the inverse source problem, we refer to the books [2], [10], [11] and the references therein. However, the methods for the inverse source problem are difficult to apply to the case where q depends on the whole components of x. Our method is applicable also to the inverse source problem.

The main purpose of this paper is to prove global uniqueness results within polygons under extra conditions. We always assume that the boundary of a polygon under consideration is a simply closed curve, and by a polygon we mean its interior. Moreover, throughout this paper, we assume

$$f \geq 0, \quad \not\equiv 0 \quad \text{on } \partial\Omega \quad \text{and} \quad q \in C^2(\overline{\Omega}), \quad q > 0, \quad \text{on } \overline{\Omega}. \tag{1.3}$$

We state our first main theorem. For $D \subset \mathbb{R}^2$, we denote the convex hull (i.e., the smallest convex set containing D) by $\mathrm{co}\,(D)$.

THEOREM 1.1. *If D_1 and D_2 are polygons such that $\overline{D_1}, \overline{D_2} \subset \Omega$ and $\Lambda_{D_1}(f) = \Lambda_{D_2}(f)$, then $co\,(D_1) = co\,(D_2)$.*

From Theorem 1.1, we can readily derive the following.

COROLLARY 1.2. *If D_1 and D_2 are convex polygons such that $\overline{D_1}, \overline{D_2} \subset \Omega$ and $\Lambda_{D_1}(f) = \Lambda_{D_2}(f)$, then $D_1 = D_2$.*

In Theorem 1.1, we cannot conclude that $D_1 = D_2$ without convexity. In the case of Figure 1, our argument does not work, and we do not know the uniqueness.

Next we show some uniqueness results for nonconvex polygons. First we show the uniqueness in a case where D_1 and D_2 have a common contact edge. For any domains D, E compactly contained in Ω, we denote the outer most boundary of $D \cup E$ by $\partial_{\mathrm{out}}(D \cup E)$, i.e.,

$$\partial_{\mathrm{out}}(D \cup E) = \{x \in \partial(D \cup E) |\, \text{there exists a continuous curve in } \Omega \setminus \overline{(D \cup E)} \text{ joining } x \text{ with some point of } \partial\Omega\}.$$

Here and henceforth, by a curve, we exclude the end points.

THEOREM 1.3. *Assume that D_1 and D_2 are polygons and that a line segment $\overline{A_0B_0} \subset \partial D_1 \cap \partial D_2$ lies on $\partial_{out}(D_1 \cup D_2)$. Then $\Lambda_{D_1}(f) = \Lambda_{D_2}(f)$ yields $D_1 = D_2$.*

Second we show the uniqueness in a case where all edges of D_1 and D_2 are parallel to two independent vectors.

THEOREM 1.4. *Assume that D_1 and D_2 are polygons such that there exist two independent vectors $\vec{a}$ and $\vec{b}$ such that all the edges of D_1 and D_2 are parallel to $\vec{a}$ or $\vec{b}$. Then $\Lambda_{D_1}(f) = \Lambda_{D_2}(f)$ yields $D_1 = D_2$.*

In particular, if polygons D_1 and D_2 are composed of rectangles in the forms of $\{(x_1, x_2) |\, a_1 < x_1 < b_1, a_2 < x_2 < b_2\}$, then Theorem 1.4 is applicable. Our

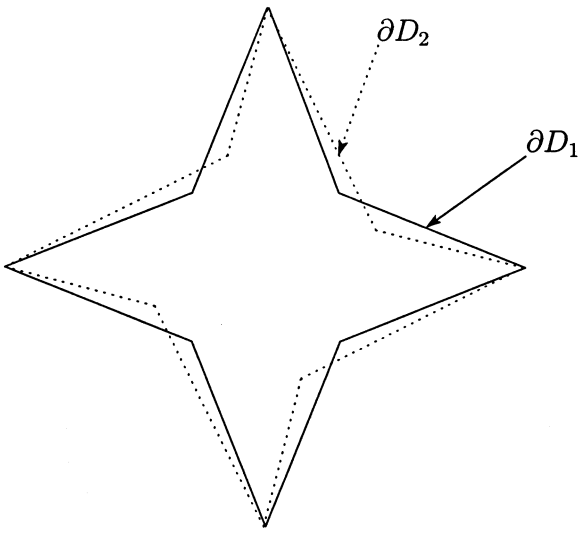


FIG. 1. *An example for $D_1 \neq D_2$ in the nonconvex case.*

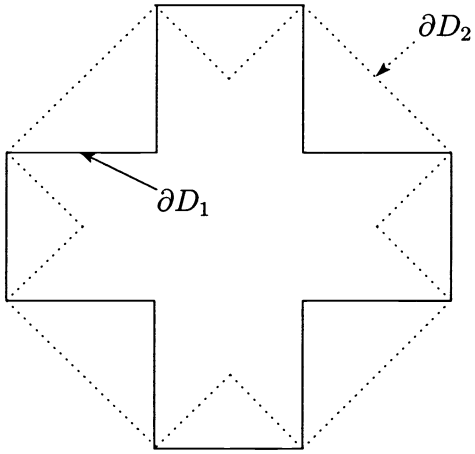


FIG. 2. *An example for $D_1 \neq D_2$ in the case where all the angles are right and all the edges are not parallel to the two fixed vectors.*

argument does not work even if all the vertex angles are right angles but all the edges are not parallel to one of the two fixed directions. See Figure 2. We think that the uniqueness results for nonconvex cases obtained so far are not comprehensive and should be improved.

Let u_j, $j = 1, 2$, be the solution to (1.1) corresponding to the domain D_j. We can prove (see, e.g., [7], [16]) that for any subdomain Ω' compactly contained in Ω, the solutions u_j, $j = 1, 2$, satisfy

$$u_j \in H^2(\Omega') \cap C^1(\Omega'). \tag{1.4}$$

Moreover, the maximum principle applied to u_j shows that

$$u_j > 0 \qquad \text{in} \quad \Omega, \quad j = 1, 2. \tag{1.5}$$

In the next section, we show a proposition on nonexistence of solutions to a Cauchy problem from which our main theorem is derived.

2. Nonexistence of an H^2-solution to a Cauchy problem for the Laplace equation. For the proof of the next nonexistence proposition, we need the following lemma about regularity of an H^1-solution to an elliptic equation. The proof of our lemma is essentially based on [6]. For completeness, we will give the proof.

LEMMA 2.1. *Let $\triangle P_1P_2P_3$ be the interior of a triangle which has three vertices $P_j \in \mathbb{R}^2$, $j = 1, 2, 3$. Assume that $f \in L^\mu(\triangle P_1P_2P_3)$ for some $\mu > 2$. If $v \in H^1(\triangle P_1P_2P_3)$ is the solution to a Dirichlet problem for the Laplace equation*

$$\begin{cases} \Delta v = f & \text{in} \quad \triangle P_1P_2P_3, \\ v = 0 & \text{on} \quad \overline{P_1P_2} \cup \overline{P_2P_3} \cup \overline{P_3P_1}, \end{cases} \tag{2.1}$$

then there exists a real number $p > 2$ such that

$$v \in W^{2,p}(\triangle P_1P_2P_3). \tag{2.2}$$

Proof. Let θ_1, θ_2, and θ_3 be the angles $\measuredangle P_3P_1P_2$, $\measuredangle P_1P_2P_3$, and $\measuredangle P_2P_3P_1$, respectively. Since $0 < \theta_j < \pi$ for any $j = 1, 2, 3$, we can take a real number $q_0 \in (1, 2)$ so that

$$\frac{2}{q_0} < \min\left\{\frac{\pi}{\theta_1}, \frac{\pi}{\theta_2}, \frac{\pi}{\theta_3}\right\}. \tag{2.3}$$

Let $p := \min\{\frac{q_0}{q_0 - 1}, \mu\}$. Clearly the number p is greater than 2. We claim that

$$v \in W^{2,p}(\triangle P_1P_2P_3). \tag{2.4}$$

Let $q := \frac{p}{p-1}$. Then by (2.3) we have

$$\frac{2}{q} = \frac{2(p-1)}{p} \le \frac{2}{q_0} < \min\left\{\frac{\pi}{\theta_1}, \frac{\pi}{\theta_2}, \frac{\pi}{\theta_3}\right\},$$

which implies that the number $\frac{2\theta_j}{q\pi}$ is not an integer for any $j = 1, 2, 3$. Since $p \le \mu$ and $f \in L^\mu(\triangle P_1P_2P_3)$, $f \in L^p(\triangle P_1P_2P_3)$. Therefore, it follows from Theorem 4.4.4.13 in [6] that there exist real numbers $c_{j,m}$ and a function w such that

$$w - \sum_{\substack{1 \le j \le 3 \\ -\frac{2}{q} < \lambda_{j,m} < 0 \\ \lambda_{j,m} \ne -1}} c_{j,m} S_{j,m} \in W^{2,p}(\triangle P_1P_2P_3)$$

and

$$\begin{cases} \Delta w = f & \text{in} \quad \triangle P_1P_2P_3, \\ w = 0 & \text{on} \quad \overline{P_1P_2} \cup \overline{P_2P_3} \cup \overline{P_3P_1}, \end{cases}$$

where m is a negative integer, $\lambda_{j,m} = \frac{m\pi}{\theta_j}$, and the functions $S_{j,m}$ are defined in equation $(4,4,3,7)$ in [6]. We note that $S_{j,m}$ does not necessarily belong to $W^{2,p}(\triangle P_1P_2P_3)$. The uniqueness of the Dirichlet problem yields

$$w = v.$$

Furthermore, our choice of constants p, q implies that there is no negative integer m such that

$$-\frac{2}{q} < \lambda_{j,m} = \frac{m\pi}{\theta_j} < 0.$$

Hence we can conclude that

$$v \in W^{2,p}(\triangle P_1P_2P_3). \qquad \square$$

Applying Lemma 2.1, we can show a proposition about the nonexistence of an H^2-solution to a Cauchy problem of the Laplace equation. This proposition plays the essential role in proving our theorems.

PROPOSITION 2.2. *Let $\triangle P_1P_2P_3$ be the interior of a triangle which has three vertices $P_j \in \mathbb{R}^2$, $j = 1, 2, 3$. Let $G \in W^{1,\infty}(\triangle P_1P_2P_3)$ be strictly positive in $\triangle P_1P_2P_3$. Then there exists no solution $y \in H^2(\triangle P_1P_2P_3)$ to*

$$\begin{cases} \Delta y = G & \text{in} \quad \triangle P_1P_2P_3, \\ y = |\nabla y| = 0 & \text{on} \quad \overline{P_1P_2} \cup \overline{P_1P_3}. \end{cases} \tag{2.5}$$

Proof. Let $\vec{a} = \overrightarrow{OP_2} - \overrightarrow{OP_1} := (a_1, a_2)$, $\vec{b} = \overrightarrow{OP_3} - \overrightarrow{OP_1} := (b_1, b_2)$, and $D = \triangle P_1P_2P_3$. We can take a rotation and a translation, if necessary, so that we may assume that the point P_1 is the origin O, $b_2 = 0$, and $a_2 > 0$. Let $\epsilon = \frac{1}{8}\text{dist}(O, \overline{P_2P_3})$, and for any real number $s > 0$ let us take a cut-off function χ_s satisfying $0 \le \chi_s \le 1$, $\chi_s \in C^\infty(\mathbb{R}^2)$, and

$$\chi_s(x) = \begin{cases} 1 & \text{if} \quad |x| < s\epsilon, \\ 0 & \text{if} \quad |x| > (s+1)\epsilon. \end{cases}$$

Suppose that $y \in H^2(D)$ satisfies (2.5). By $y \in H^2(D)$, the Sobolev imbedding theorem yields that for any $\ell \ge 2$

$$y\chi_6 \in W^{1,\ell}(D). \tag{2.6}$$

Since $y \in H^2(D)$ and $y = |\nabla y| = 0$ on $\overline{OP_2} \cup \overline{OP_3}$, the function $y\chi_4$ belongs to $H^1(D)$ and is the solution to

$$\begin{cases} \Delta(y\chi_4) = G\chi_4 + 2\nabla y \cdot \nabla\chi_4 + y(\Delta\chi_4) & \text{in} \quad D, \\ y\chi_4 = 0 & \text{on} \quad \partial D. \end{cases}$$

Relation (2.6) implies that $G\chi_4 + 2\nabla y \cdot \nabla\chi_4 + y(\Delta\chi_4) \in L^\ell(D)$ for any $\ell > 2$. It follows from Lemma 2.1 that there exists a real number $\mu > 2$ such that

$$y\chi_4 \in W^{2,\mu}(D), \tag{2.7}$$

and hence

$$y \in W^{2,\mu}(D \cap B(O, 4\epsilon)), \tag{2.8}$$

where $B(O, 4\epsilon) = \{z \in \mathbb{R}^2 \mid |z| < 4\epsilon\}$. Next from (2.8) we can see that $\partial_{x_j}(y\chi_2)$, $j = 1, 2$, belongs to $H^1(D)$ and is the solution to

$$\begin{cases} \Delta(\partial_{x_j}(y\chi_2)) = \partial_{x_j}[G\chi_2 + 2\nabla y \cdot \nabla\chi_2 + y(\Delta\chi_2)] & \text{in} \quad D, \\ \partial_{x_j}(y\chi_2) = 0 & \text{on} \quad \partial D. \end{cases}$$

Relation (2.8) implies that

$$\partial_{x_j}[G\chi_2 + 2\nabla y \cdot \nabla\chi_2 + y(\Delta\chi_2)] \in L^{\mu}(D).$$

Since $\mu > 2$, by Lemma 2.1, there exists a real number $p > 2$ such that

$$\partial_{x_j}(y\chi_2) \in W^{2,p}(D). \tag{2.9}$$

Therefore, we have

$$y \in W^{3,p}(D \cap B(O, 2\epsilon)). \tag{2.10}$$

The imbedding theorem (see, e.g., [1]) implies that

$$y \in C^2(\overline{D \cap B(O, \epsilon)}). \tag{2.11}$$

Then $y(b_1 t, 0) = \partial_{x_2} y(b_1 t, 0) = 0$ and $y(a_1 t, a_2 t) = 0$ for $0 \le t \le \delta$, where $\delta > 0$ is sufficiently small. Therefore,

$$(\partial_{x_1}\partial_{x_2} y)(b_1 t, 0) = (\partial_{x_1}^2 y)(b_1 t, 0) = 0$$

and

$$0 = \frac{d^2 y(a_1 t, a_2 t)}{dt^2} = a_1^2(\partial_{x_1}^2 y)(a_1 t, a_2 t) + 2a_1 a_2(\partial_{x_1}\partial_{x_2} y)(a_1 t, a_2 t) + a_2^2(\partial_{x_2}^2 y)(a_1 t, a_2 t)$$

for $0 \le t \le \delta$. Hence, by $y \in C^2(\overline{D \cap B(O, \epsilon)})$, we have

$$\partial_{x_1}^2 y(0, 0) = \partial_{x_2}^2 y(0, 0) = \partial_{x_1}\partial_{x_2} y(0, 0) = 0,$$

so that $\Delta y(0, 0) = G(0, 0) = 0$, which contradicts that $G > 0$ in D. □

REMARK 2.3. *In Proposition* 2.2, *it is essential that* $\overline{P_1P_2}$ *and* $\overline{P_1P_3}$ *intersect at* P_1 *transversally. In fact, if a curve joining* P_1, P_2 *and a curve joining* P_1, P_3 *intersect smoothly at* P_1, *there may exist a solution* $y \in H^2(D)$ *for some positive* $G \in H^1(D)$ *with* $\partial_{x_2} G \in L^{\infty}(D)$.

Example for existence for a smooth curve passing P_1, P_2, P_3. Let

$$D = \left\{(x_1, x_2) \middle| \, 0 \le x_1 < \frac{1}{2},\ 0 < x_2 < -\frac{1}{8}\left(x_1 - \frac{1}{2}\right)\right\}$$
$$\cup \left\{(x_1, x_2) \middle| \, -\frac{1}{2} < x_1 < 0,\ x_1^3 < x_2 < -\frac{1}{8}\left(x_1 - \frac{1}{2}\right)\right\}$$

and

$$y(x_1,x_2)=\begin{cases} x_2^2, & x_1\geq 0,\\ (x_1^3-x_2)^2, & x_1<0,\end{cases}$$

$$G(x_1,x_2)=\begin{cases} 2, & x_1\geq 0,\\ 2-12x_1x_2+30x_1^4, & x_1<0.\end{cases}$$

We regard $\{(x_1,0)|0\leq x_1<\frac{1}{2}\}\cup\{(x_1,x_1^3)|-\frac{1}{2}<x_1<0\}$ as the curve $P_2P_1P_3$. Then the two parts of the curve $P_2P_1P_3$ connect smoothly at $P_1\equiv O$. Moreover, we can directly verify that $y\in C^2(\overline{D})$, $G\in H^1(D)$ and $\partial_{x_2}G\in L^\infty(D)$, $y=|\nabla y|=0$ on the curve $P_2P_1P_3$ and $\Delta y=G>0$ in D.

3. Proof of Theorem 1.1. Let us define $y:=u_1-u_2$ in Ω. Then by (1.1) and (1.5), the function y satisfies

$$\Delta y=0 \quad \text{in} \quad \Omega\setminus(\overline{D_1\cup D_2}), \tag{3.1}$$

$$\Delta y=qu_1>0 \quad \text{in} \quad D_1\setminus\overline{D_2}, \tag{3.2}$$

$$\Delta y=-qu_2<0 \quad \text{in} \quad D_2\setminus\overline{D_1}, \tag{3.3}$$

$$\Delta y=qy \quad \text{in} \quad D_1\cap D_2, \tag{3.4}$$

$$y=|\nabla y|=0 \quad \text{on} \quad \partial\Omega. \tag{3.5}$$

Henceforth F is the component of $\Omega\setminus(\overline{D_1\cup D_2})$ which is connected with $\partial\Omega$. Since y is harmonic in $\Omega\setminus(\overline{D_1\cup D_2})$ and $y=\frac{\partial y}{\partial\nu}=0$ on $\partial\Omega$, the unique continuation (see, e.g., [10]) implies that

$$y\equiv 0 \quad \text{on} \quad \overline{F}. \tag{3.6}$$

Then we can see the following two facts:

If $D,E\subset\Omega$ are convex polygons and $D\neq E$, then there exists

a vertex O of D such that $O\in\Omega\setminus\overline{E}$

or a vertex O of E such that $O\in\Omega\setminus\overline{D}$. (3.7)

(3.8) If $D,E\subset\Omega$ are convex polygons, then $\Omega\setminus\overline{(D\cup E)}$ is connected.

Now we will complete the proof of Theorem 1.1. Assume contrarily that $\operatorname{co}(D_1)\neq\operatorname{co}(D_2)$. Then, by (3.7), there exists a vertex O of $\operatorname{co}(D_1)$ such that $O\in\Omega\setminus\overline{\operatorname{co}(D_2)}$ or a vertex O of $\operatorname{co}(D_2)$ such that $O\in\Omega\setminus\overline{\operatorname{co}(D_1)}$. Without loss of generality, we may assume the former case. Then, since $O\in\Omega\setminus\overline{\operatorname{co}(D_2)}$, we can take a sufficiently small triangle $\triangle OAB$ such that

$$\overline{OA}\cup\overline{OB}\subset\partial(\operatorname{co}(D_1)) \qquad \text{and} \qquad \triangle OAB\subset\operatorname{co}(D_1)\setminus\overline{\operatorname{co}(D_2)}.$$

By (3.8), we have

$$\overline{OA}\cup\overline{OB}\subset\Omega\setminus(\operatorname{co}(D_1)\cup\operatorname{co}(D_2))\subset\overline{F}. \tag{3.9}$$

Any vertex of $\mathrm{co}\,(D_1)$ is a convex vertex of D_1; that is, in a neighborhood of that vertex, D_1 is convex. Therefore O is a convex vertex of D_1. By $\mathrm{co}\,(D_1) \supset D_1$, we can take $\triangle OA'B'$ such that

$$\overline{OA'} \cup \overline{OB'} \subset \partial D_1 \qquad \text{and} \qquad \triangle OA'B' \subset \triangle OAB.$$

Hence it follows from $\triangle OAB \subset \mathrm{co}\,(D_1) \setminus \overline{\mathrm{co}\,(D_2)}$ that $\triangle OA'B' \subset \mathrm{co}\,(D_1) \setminus \overline{\mathrm{co}\,(D_2)}$. Moreover, by (3.9), we see that $\overline{OA'} \cup \overline{OB'}$ is included in $\overline{F}$. Therefore, by (3.2) and (3.6), we have $\Delta y = qu_1 > 0$ in $\triangle OA'B'$ and $y = |\nabla y| = 0$ on $\overline{OA'} \cup \overline{OB'}$. Again by (1.4), we see that $qu_1 \in H^1(\triangle OA'B')$ and $|\nabla(qu_1)|$ is bounded in $\overline{\triangle OA'B'}$, and so we apply Proposition 2.2, which yields a contradiction. Hence $\mathrm{co}\,(D_1) = \mathrm{co}\,(D_2)$ follows. Thus the proof of Theorem 1.1 is complete.

4. Proof of Theorem 1.3. Let E be the connected component of $D_1 \cap D_2$ such that $\overline{A_0B_0} \subset \partial E$. Since $\overline{A_0B_0} \subset \partial_{\mathrm{out}}(D_1 \cup D_2)$ and $\Delta y - qy = 0$ in E, the unique continuation implies that

$$y = 0 \quad \text{in } E. \tag{4.1}$$

We represent the boundary ∂D_j, $j = 1, 2$, by a continuous curve $\alpha_j : [0, 1] \to \partial D_j$ such that α_j is injective in $[0, 1)$, $\alpha_j(0) = A_0$, $\alpha_j(\frac{1}{2}) = B_0$, and $\alpha_j(1) = \alpha_j(0)$. Exchanging A_0 with B_0 if necessary, we may assume that the curves α_j are oriented in the positive direction; that is, the outward normal vector to ∂D_j and the oriented tangential vector of ∂D_j form a right-handed system at any point of ∂D_j.

Let

$$a = \inf\{t \in [0, 1] |\, \alpha_1(t) \neq \alpha_2(t)\}.$$

Then we note that $\alpha_1(t) = \alpha_2(t)$ if $0 \leq t \leq a$.

We will prove the theorem by reduction to absurdity. That is, assume that $D_1 \neq D_2$. Then, by $\alpha_1(1/2) = \alpha_2(1/2)$ and $\alpha_1(1) = \alpha_2(1)$, we can take a number $\frac{1}{2} \leq a < b \leq 1$ such that $\alpha_1(t) \neq \alpha_2(t)$ for $t \in (a, b)$ and $\alpha_1(b) = \alpha_2(b)$.

Since $\alpha_1(t) = \alpha_2(t)$ for $0 \leq t \leq a$ and $\alpha_1(t) \neq \alpha_2(t)$ for $t \in (a, b)$, the point $\alpha_1(a)$ is a vertex of D_1 or a vertex of D_2. Therefore we see that $\alpha_1(a, b)$ is outside $\overline{D_2}$ or $\alpha_2(a, b)$ is outside $\overline{D_1}$. Therefore either $\alpha_1[a, b]$ or $\alpha_2[a, b]$ is on $\partial_{\mathrm{out}}(D_1 \cup D_2)$.

In fact, let $\alpha_1(a, b)$ be outside $\overline{D_2}$. For any $x \in \alpha_1[a, b]$, there exists a continuous curve γ_1 connecting x and some $y \in \alpha_1[0, \frac{1}{2}]$ such that $\gamma_1 \setminus \{x, y\} \subset \Omega \setminus \overline{(D_1 \cup D_2)}$. Since $\alpha_1[0, \frac{1}{2}] \subset \partial_{\mathrm{out}}(D_1 \cup D_2)$, we can take a continuous curve γ_2 connecting y and some $x_0 \in \partial\Omega$ such that $\gamma_2 \setminus \{y\} \subset \Omega \setminus \overline{(D_1 \cup D_2)}$. Hence we can choose a continuous curve γ such that γ is sufficiently close to $\gamma_1 \cup \gamma_2$, $\gamma \subset \Omega \setminus \overline{(D_1 \cup D_2)}$, and γ connects x and x_0. Thus $\alpha_1[a, b] \subset \partial_{\mathrm{out}}(D_1 \cup D_2)$.

Without loss of generality, we may assume that

$$\alpha_1[a, b] \text{ is contained in } \partial_{\mathrm{out}}(D_1 \cup D_2). \tag{4.2}$$

Let $a < t_1^j < \cdots < t_{k_j}^j < b$, $j = 1, 2$, be a partition of $[a, b]$ such that $\alpha_j(t_1^j), \ldots, \alpha_j(t_{k_j}^j)$ are all the vertices of D_j on $\alpha_j(a, b)$.

We will claim that

$$\alpha_1(a) \text{ is a vertex of both } D_1 \text{ and } D_2. \tag{4.3}$$

In fact, since $\alpha_1(t) = \alpha_2(t)$ for $t \in [0, a]$ and $\alpha_1(t) \neq \alpha_2(t)$ for $t \in (a, b)$, the point $\alpha_1(a)$ cannot be simultaneously on an edge of D_1 and on an edge of D_2. Here and henceforth, by an edge, we mean that it does not contain any vertices.

Moreover, if $\alpha_1(a)$ is on an edge of one domain and is a vertex of the other, then, in terms of (4.2), we can take a triangle $\triangle\alpha_2(t_1^2)\alpha_1(a)\alpha_1(t_1^1)$, so that

$$\begin{cases} \overline{\alpha_1(a)\alpha_2(t_1^2)} \subset \partial E, \quad \overline{\alpha_1(a)\alpha_1(t_1^1)} \subset \partial_{\text{out}}(D_1 \cup D_2), \\ \text{and the interior of this triangle is contained in} \quad D_1 \setminus \overline{D_2}. \end{cases} \tag{4.4}$$

By (4.1) and (3.6), we apply Proposition 2.2 to be led to a contradiction. Thus we have proved (4.3).

We choose small $\varepsilon > 0$, so that $\alpha_1(t) = \alpha_2(t)$ is on an edge of D_j, $j = 1, 2$, for $t \in [a - \varepsilon, a]$. Furthermore we can take a suitable rotation, if necessary, so that $\alpha_1(t)$ is on the x_1-axis for $t \in [a - \varepsilon, a]$ and the x_1-component of $\alpha_1(a - \varepsilon)$ is smaller than the one of $\alpha_1(a)$. Then, by the orientation of α_1 and α_2, the domains D_1 and D_2 are located in the upper half plane $\mathbb{R}^2_+ := \{(x_1, x_2) | x_2 > 0\}$ locally near the edge $\overline{\alpha_1(a - \varepsilon)\alpha_1(a)}$.

Furthermore

$$\text{the edge } \overline{\alpha_1(a)\alpha_1(t_1^1)} \text{ lies in the lower half plane } \mathbb{R}^2_-$$

$$:= \{(x_1, x_2) | x_2 < 0\} \text{ and the edge } \overline{\alpha_1(a)\alpha_2(t_1^2)} \text{ in } \mathbb{R}^2_+. \tag{4.5}$$

In fact, assume contrarily. Then, by (4.3), we alternatively have two cases:

(i) $\overline{\alpha_1(a)\alpha_1(t_1^1)} \subset \mathbb{R}^2_+$, $\overline{\alpha_1(a)\alpha_2(t_1^2)} \subset \mathbb{R}^2_-$.

(ii) $\overline{\alpha_1(a)\alpha_1(t_1^1)} \cup \overline{\alpha_1(a)\alpha_2(t_1^2)} \subset \mathbb{R}^2_+$ or $\mathbb{R}^2_-$.

Case (i) is impossible, because the domains D_1 and D_2 are located in $\mathbb{R}^2_+$ locally near $\overline{\alpha_1(a - \varepsilon)\alpha_1(a)}$, and so, if (i) occurs, then $\overline{\alpha_1(a)\alpha_1(t_1^1)} \subset \partial_{\text{out}}(D_1 \cup D_2)$ does not hold. This contradicts (4.2). Case (ii) is impossible also. In fact, assume that case (ii) occurs. Then, by (4.2), we can take a triangle $\triangle\alpha_2(t_1^2)\alpha_1(a)\alpha_1(t_1^1)$ satisfying (4.4). This is again a contradiction of Proposition 2.2. Hence we have proved (4.5).

By (4.2), we have $\alpha_2[a, b] \subset \partial E$, so that Proposition 2.2 implies that

$$\alpha_1(t_i^1) \notin \text{CV}(D_1), \quad i = 1, \ldots, k_1, \tag{4.6}$$

and

$$\alpha_2(t_i^2) \in \text{CV}(E), \quad i = 1, \ldots, k_2. \tag{4.7}$$

Here $\text{CV}(D)$ denotes the set of all convex vertices of a polygon D.

We will prove (4.6) and (4.7). In fact, otherwise, there is a vertex $\alpha_1(t_{i_0}^1) \in CV(D_1)$ or $\alpha_2(t_{j_0}^2) \notin CV(E)$. First let $\alpha_1(t_{i_0}^1) \in CV(D_1)$ for some i_0. Then, by (4.2), we can take a triangle $\triangle P_1\alpha_1(t_{i_0}^1)Q_1 \subset D_1 \setminus \overline{D_2}$ such that $y = |\nabla y| = 0$ on the parts $\overline{P_1\alpha_1(t_{i_0}^1)}$ and $\overline{\alpha_1(t_{i_0}^1)Q_1}$ of the edges of D_1. This is a contradiction by Proposition 2.2. Therefore (4.6) must hold. Second let $\alpha_2(t_{j_0}^2) \notin CV(E)$ for some j_0. By (4.1), $y = |\nabla y| = 0$ on the parts $\overline{P_2\alpha_2(t_{j_0}^2)}$ and $\overline{\alpha_2(t_{j_0}^2)Q_2}$ of the edges of D_2. Moreover, by (4.2), we see that $\triangle P_2\alpha_2(t_{j_0}^2)Q_2 \subset D_1 \setminus \overline{D_2}$. This is a contradiction again by Proposition 2.2. Thus the proof of (4.7) is complete.

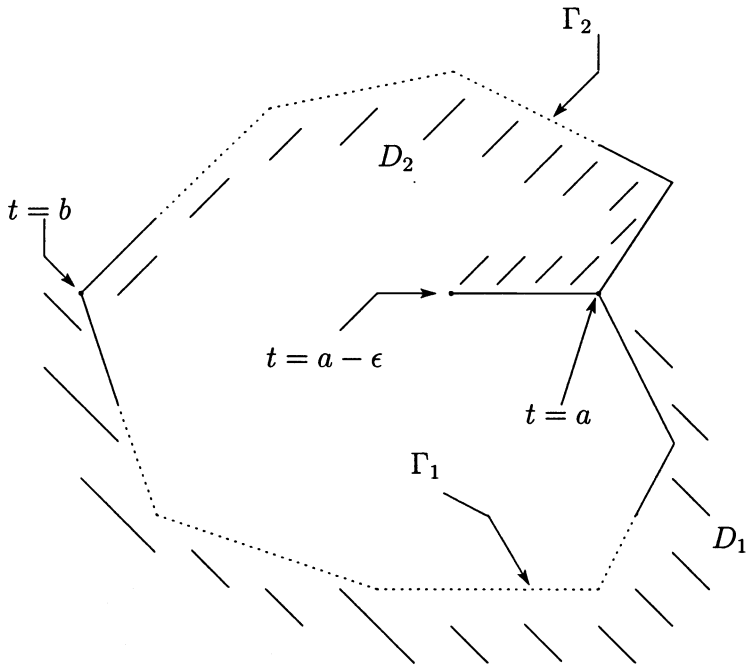


FIG. 3. *The figure for the proof of Theorem* 1.3.

Let us trace the curves $\Gamma_1 := \alpha_1[a-\varepsilon, b]$ and $\Gamma_2 := \alpha_2[a-\varepsilon, b]$. Both curves coincide from $t = a-\varepsilon$ to $t = a$. By (4.6) and (4.7), the former is oriented clockwise, while the latter is oriented counterclockwise. By $\alpha_1(b) = \alpha_2(b)$, the curves $(-\Gamma_1) \cup \Gamma_2 \setminus \alpha_1[a-\varepsilon, a]$ is a closed curve and surrounds a polygon $\widetilde{D}$. Here we regard $-\Gamma_1$ as a curve oriented from $\alpha_1(b)$ to $\alpha_1(a-\varepsilon)$. Moreover, the intersection of $\widetilde{D}$ and some neighborhood of Γ_2 is in D_2, while the intersection of $\Omega \setminus \overline{\widetilde{D}}$ and some neighborhood of $-\Gamma_1$ is in D_1 (Figure 3).

Therefore Γ_1 cannot be connected to $\partial\Omega$ by any continuous curve in $\partial_{\text{out}}(D_1 \cup D_2)$. In fact, for any $x \in \partial\Omega$ and $\widetilde{x} \in \Gamma_1$, let γ be an arbitrary continuous curve connecting x and $\widetilde{x}$. Then γ must intersect Γ_1 or Γ_2 transversally. If γ intersects Γ_1 transversally, then γ must pass in D_1. If γ intersects Γ_2 transversally, then γ must pass in D_2. Therefore $\gamma \not\subset \Omega \setminus \overline{(D_1 \cup D_2)}$. This contradicts (4.2). Thus, by reduction to absurdity, the proof of Theorem 1.3 is complete.

5. Proof of Theorem 1.4. Without loss of generality, we may assume that $\vec{b} = (1, 0)$. Since $a_2 \neq 0$ by the linear independence of $\vec{a}$ and $\vec{b}$, we can choose $a_2 = 1$. Let us set $\vec{\mu} = (1, -a_2)$. We set $x = (x_1, x_2) \in \mathbb{R}^2$, and

$$t_0 = \min\{x \cdot \vec{\mu} | x \in \overline{D_1}\} \quad \text{and} \quad s_0 = \min\{x \cdot \vec{\mu} | x \in \overline{D_2}\}. \tag{5.1}$$

Here and henceforth, $x \cdot \vec{\mu}$ denotes the scalar product of $x, \vec{\mu} \in \mathbb{R}^2$. Then

$$t_0 = s_0. \tag{5.2}$$

In fact, otherwise, we may assume that $t_0 < s_0$. Then

$$\{x \in D_1 | t_0 < x \cdot \vec{\mu} < s_0\} \subset \Omega \setminus \overline{D_2},$$

and there exists a vertex O of D_1 with $O \cdot \vec{\mu} = t_0$. Therefore, we can take a small triangle such that

$$\overline{OA} \cup \overline{OB} \subset \partial D_1 \quad \text{and} \quad \triangle OAB \subset \{x \in D_1 | t_0 < x \cdot \vec{\mu} < s_0\}.$$

We recall that F is the connected component of $\Omega \setminus \overline{(D_1 \cup D_2)}$ with $\partial\Omega$. Then we see that $\overline{OA} \cup \overline{OB} \subset \overline{F}$. Hence, by (3.6), we have $y = |\nabla y| = 0$ on $\overline{OA} \cup \overline{OB}$. In terms of (3.2), we apply Proposition 2.2, so that nonexistence of y is shown, which is a contradiction. Thus (5.2) has been proved.

Next for $j = 1, 2$, let $q_j = \sup\{x_2 | x = (x_1, x_2) \in \partial D_j \text{ and } x \cdot \vec{\mu} = t_0\}$ and let $P_j = (p_j, q_j)$ be the intersection point of $x_2 = q_j$ and $x \cdot \vec{\mu} = t_0$. If $q_1 \neq q_2$, then we may assume that $q_1 > q_2$. Then we can take a small triangle $\triangle P_1 Q R$ such that

$$\overline{P_1 Q} \cup \overline{P_1 R} \subset \partial D_1 \quad \text{and} \quad \triangle P_1 Q R \subset D_1 \setminus \overline{D_2}.$$

Then $\overline{P_1 Q} \cup \overline{P_1 R} \subset \overline{F}$, by (3.2) and (3.6), we apply Proposition 2.2, so that nonexistence of y is shown, which is a contradiction. Therefore,

$$q_1 = q_2. \tag{5.3}$$

Relations (5.2) and (5.3) imply that $\partial D_1 \cap \partial D_2 \cap \{x \mid x \cdot \vec{\mu} = t_0\}$ must contain a common line segment in $\partial_{\text{out}}(D_1 \cup D_2)$. By Theorem 1.3, we can conclude that $D_1 = D_2$.

Acknowledgments. The authors thank Professor Victor Isakov (Wichita State University) for helpful comments. Moreover, they are grateful to the referees for their invaluable comments. In particular, they highly appreciate the proposal by one referee for simplification of the proof of Proposition 2.2. In fact, the authors can simplify the proof, although their proof is slightly different from the proposal for the sake of more safety.

REFERENCES

[1] A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] G. Anger, *Inverse Problems in Differential Equations*, Plenum, New York, 1990.
[3] W. Fang and E. Cumberbatch, *Inverse problems for metal oxide semiconductor field-effect transistor contact resistivity*, SIAM J. Appl. Math., 52 (1992), pp. 699–709.
[4] A. Friedman, *Detection of mines by electric measurements*, SIAM J. Appl. Math., 47 (1987), pp. 201–212.
[5] A. Friedman and V. Isakov, *On the uniqueness in the inverse conductivity problem with one measurement*, Indiana Univ. Math. J., 38 (1989), pp. 563–579.
[6] P. Grisvard, *Elliptic Problem in Nonsmooth Domains*, Pitman, Boston, MA, 1985.
[7] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, Heidelberg, 1977.
[8] F. Hettlich and W. Rundell, *Recovery of the support of a source term in an elliptic differential equation*, Inverse Problems, 13 (1997), pp. 959–976.
[9] V. Isakov, *On uniqueness of recovery of a discontinuous conductivity coefficient*, Comm. Pure Appl. Math., 41 (1988), pp. 865–877.
[10] V. Isakov, *Inverse Source Problems*, AMS, Providence, RI, 1990.
[11] V. Isakov, *Inverse Problems for Partial Differential Equations*, Springer-Verlag, Berlin, 1998.
[12] H. Kang, K. Kwon, and K. Yun, *Recovery of an inhomogeneity in an elliptic equation*, Inverse Problems, 17 (2001), pp. 25–44.
[13] H. Kang and J.K. Seo, *The layer potential technique for the inverse conductivity problem*, Inverse Problems, 12 (1996), pp. 267–278.
[14] H. Kang and J.K. Seo, *Inverse conductivity problem with one measurement: Uniqueness of balls in* $\mathbb{R}^3$, SIAM J. Appl. Math., 59 (1999), pp. 1533–1539.

[15] S. Kim, *Unique determination of inhomogeneity in an elliptic equation*, Inverse Problems, 18 (2002), pp. 1325–1332.
[16] O.A. Ladyzhenskaya and N.N. Ural'tseva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.
[17] J.K. Seo, *On the uniqueness in the inverse conductivity problem*, J. Fourier Anal. Appl., 2 (1996), pp. 227–235.

SIAM J. MATH. ANAL.
Vol. 35, No. 1, pp. 160–171

TRAVELING WAVES IN A SUSPENSION BRIDGE SYSTEM*

ZHONGHAI DING†

Abstract. In this paper, we study traveling waves in a suspension bridge system governed by the coupled nonlinear wave and beam equations describing oscillations in the supporting cable and roadbed. By applying the variational method, it is proved that the suspension bridge system has at least one nontrivial traveling wave.

Key words. suspension bridge system, traveling wave, mountain pass lemma

AMS subject classifications. 35Q72, 58E05, 73K12, 73V25

DOI. 10.1137/S0036141002412690

1. Introduction.

The suspension bridge is a common type of civil engineering structure. It is well known that suspension bridges may display certain oscillations under external aerodynamic forces. Under the action of a strong wind, for example, a narrow and very flexible suspension bridge can undergo dangerous oscillations [2]. Therefore, it is imperative to investigate dynamic oscillations in suspension bridges, especially the destructive large-amplitude oscillations, and to develop design techniques to prevent such destructive oscillations. In the last decade, Lazer and McKenna [13], [14] proposed new mathematical models describing oscillations in suspension bridges, which are based upon the observation of the fundamental nonlinearity in suspension bridges that the stays connecting the supporting cables and the roadbed resist expansion but do not resist compression. These new models are described by systems of coupled nonlinear partial differential equations.

The new study of suspension bridges initiated by Lazer and McKenna has obtained many important and interesting results. Multiple large-amplitude periodic oscillations have been found theoretically and numerically in the single Lazer–McKenna suspension bridge equation (see [3], [4], [10], [11], [12], [13], [14], [16], [17], and the references therein). Recently, we studied the Lazer–McKenna suspension bridge system governed by coupled nonlinear beam and wave equations and obtained multiple periodic nonlinear oscillations [6], [7], [8], [9] by applying the variational critical point theory.

In this paper, we are interested in traveling waves in suspension bridge systems. Consider a simplified suspension bridge configuration: the roadbed of length L is modeled by a horizontal vibrating beam with both ends being simply supported; the supporting cable of length L is modeled by a horizontal vibrating string with both ends being fixed; and the vertical stays connecting the roadbed to the supporting cable are modeled by one-sided springs which resist expansion but do not resist compression. Let $u(x,t)$ and $w(x,t)$ denote the downward deflections of the cable and the roadbed, respectively. The following suspension bridge model was proposed by Lazer

*Received by the editors August 6, 2002; accepted for publication (in revised form) January 15, 2003; published electronically June 25, 2003. Part of this work was completed during the author's sabbatical leave (Fall, 2002).
http://www.siam.org/journals/sima/35-1/41269.html
†Department of Mathematical Sciences, University of Nevada, Las Vegas, NV 89154-4020 (dingz@unlv.edu).

and McKenna [13]:

$$\text{(1.1)}\qquad \begin{cases} m_c u_{tt} - Q u_{xx} - K(w-u)^+ = m_c g + f_1(x,t), \quad 0 < x < L, \\ m_b w_{tt} + EI w_{xxxx} + K(w-u)^+ = m_b g + f_2(x,t), \quad 0 < x < L, \\ u(0,t) = u(L,t) = 0, \\ w(0,t) = w(L,t) = 0, \quad w_{xx}(0,t) = w_{xx}(L,t) = 0, \end{cases}$$

where $(w-u)^+ = \max\{w-u,\ 0\}$; m_c and m_b are the mass densities of the cable and the roadbed, respectively; Q is the coefficient of cable tensile strength; EI is the roadbed flexural rigidity; K is Hooke's constant of the stays; and f_1 and f_2 represent the external aerodynamic forces. If L is sufficiently large and $f_1 = f_2 = 0$, then (1.1) can be replaced approximately by the following model:

$$\text{(1.2)}\qquad \begin{cases} m_c u_{tt} - Q u_{xx} - K(w-u)^+ = m_c g, \quad x \in R,\ t \in R, \\ m_b w_{tt} + EI w_{xxxx} + K(w-u)^+ = m_b g, \quad x \in R,\ t \in R. \end{cases}$$

By letting $u = 0$, the second equation in (1.2) is the single Lazer–McKenna suspension bridge equation. The traveling waves in the single Lazer–McKenna suspension bridge equation were studied by McKenna and Walter [17] and Chen and McKenna [4] numerically and analytically. However, there has been little discussion of traveling waves in suspension bridge systems such as system (1.2). By applying the variational method and the mountain pass lemma, we show in this paper that (1.2) admits at least one nontrivial traveling wave.

The organization of this paper is as follows. In section 2, we formulate an equivalent system of (1.2). In section 3, we formulate the corresponding variational problem and follow the same idea presented in [4] by Chen and McKenna to prove that the system (1.2) admits at least one nontrivial traveling wave.

2. An equivalent system of (1.2). In order to investigate traveling waves in the suspension bridge system (1.2), we assume $u(x,t) = u_e(x) + y(x-ct)$ and $w(x,t) = w_e(x) + z(x-ct)$, where c is the wave speed and (u_e, w_e) are given by

$$\begin{cases} u_e(x) = -\dfrac{(m_c+m_b)g}{2Q} x^2, \\ w_e(x) = -\dfrac{(m_c+m_b)g}{2Q} x^2 + \dfrac{m_b g}{K}, \end{cases}$$

which satisfy the steady state equations of (1.2),

$$\begin{cases} -Q u_{xx} - K(w-u)^+ = m_c g, & x \in R, \\ EI w_{xxxx} + K(w-u)^+ = m_b g, & x \in R. \end{cases}$$

Thus $(y(s), z(s))$, where $s = x - ct$, satisfies

$$\text{(2.1)}\qquad \begin{cases} (c^2 m_c - Q) y'' - K\left[\left(z - y + \dfrac{m_b g}{K}\right)^+ - \dfrac{m_b g}{K}\right] = 0, & s \in R, \\ \\ c^2 m_b z'' + EI z^{(4)} + K\left[\left(z - y + \dfrac{m_b g}{K}\right)^+ - \dfrac{m_b g}{K}\right] = 0, & s \in R. \end{cases}$$

Let $H^r(R)$ denote the usual Sobolev space on R of order r. We are interested in those solutions (y,z) of (2.1) such that $(y,z) \in H^2(R) \times H^4(R)$.

Assume

$$0 < c^2 < \frac{Q}{m_c + m_b}. \tag{2.2}$$

By adding the two equations in (2.1) together and by integrating the resulting equation twice, we obtain

$$y = \frac{1}{Q - c^2 m_c}\left[EIz'' + c^2 m_b z\right] \stackrel{def}{=} L_1 z. \tag{2.3}$$

Thus L_1 is a bounded linear operator from $H^{r+2}(R)$ to $H^r(R)$. By substituting (2.3) into the second equation of (2.1), we obtain

$$c^2 m_b z'' + EIz^{(4)} + K\left[\left(\frac{1}{Q-c^2 m_c}\left(-EIz'' + (Q - c^2(m_c + m_b))z\right) + \frac{m_b g}{K}\right)^+ - \frac{m_b g}{K}\right] = 0. \tag{2.4}$$

Define operators L_2 and L_3 by

$$L_2 z = \frac{1}{Q - c^2 m_c}\left[-EIz'' + (Q - c^2(m_c + m_b))z\right],$$

$$L_3 z = EIz^{(4)} + c^2 m_b z''.$$

Thus L_2 is a bounded linear operator from $H^{r+2}(R)$ to $H^r(R)$, and L_3 is a bounded linear operator from $H^{r+4}(R)$ to $H^r(R)$. Then (2.4) can be written as

$$L_3 z + K\left[\left(L_2 z + \frac{m_b g}{K}\right)^+ - \frac{m_b g}{K}\right] = 0. \tag{2.5}$$

Under condition (2.2), one can verify easily that L_2 is invertible from $H^r(R)$ to $H^{r+2}(R)$. Let $\widetilde{z} = L_2 z$; then (2.5) can be written as

$$L_3 L_2^{-1}\widetilde{z} + K\left[\left(\widetilde{z} + \frac{m_b g}{K}\right)^+ - \frac{m_b g}{K}\right] = 0. \tag{2.6}$$

Define

$$\mathcal{A}\widetilde{z} = L_3 L_2^{-1}\widetilde{z}.$$

Then $\mathcal{A}$ is a bounded linear operator from $H^{r+2}(R)$ to $H^r(R)$. Let $v = \frac{K}{m_b g}\widetilde{z}$. Then (2.6) becomes

$$\mathcal{A}v + K[(v+1)^+ - 1] = 0. \tag{2.7}$$

Therefore we have the following theorem.

THEOREM 2.1. *Let condition* (2.2) *be satisfied. If* (2.7) *admits a nontrivial solution* $v \in H^2(R)$*, then the suspension bridge system* (2.1) *admits a nontrivial solution* $(y,\ z) \in H^2(R) \times H^4(R)$*. The relations between* (y, z) *and* v *are given by*

$$y = \frac{m_b g}{K} L_1 L_2^{-1} v, \qquad z = \frac{m_b g}{K} L_2^{-1} v.$$

Thus studying the existence of traveling waves in (1.2) becomes a matter of proving the existence of nontrivial solutions of (2.7).

3. Nonlinear traveling waves. To study nontrivial solutions of (2.7), we define a functional $I(v) : H^1(R) \longrightarrow R$ by

$$I(v) = \frac{1}{2}\int_R \mathcal{A}v \cdot v ds + \frac{K}{2}\int_R [((v+1)^+)^2 - 1] ds - K\int_R v ds. \tag{3.1}$$

$I(v)$ can also be rewritten as

$$I(v) = \frac{1}{2}\int_R \mathcal{A}v \cdot v ds + \frac{K}{2}\int_{v>-1} v^2 ds - K\int_{v\le -1}\left(v + \frac{1}{2}\right) ds.$$

It is easy to show the following lemma.

LEMMA 3.1. *Assume condition* (2.2) *is satisfied. Then* $I(v)$ *is continuously Fréchet differentiable with*

$$I'(v)\varphi = \int_R \mathcal{A}v \cdot \varphi ds + K\int_{v>-1} v\varphi ds - K\int_{v\le -1}\varphi ds$$

for any $v,\ \varphi \in H^1(R)$. *Consequently, the solutions of* (2.7) *in* $H^1(R)$ *correspond to critical points of* $I(v)$ *in* $H^1(R)$.

For any $v \in H^1(R)$, let $\|v\|_{H^1}$ denote the usual H^1-norm given by

$$\|v\|_{H^1}^2 = \int_R \left(|v'|^2 + |v|^2\right) ds.$$

For any $v \in H^1(R)$, define

$$\|v\|_*^2 = \int_R \mathcal{A}v \cdot v ds + K\int_R |v|^2 ds. \tag{3.2}$$

Assume throughout this paper that

$$\begin{cases} \dfrac{m_b}{m_c} < \dfrac{\sqrt{KEI}}{Q - \sqrt{KEI}} & \text{if } \ 0 < K < \dfrac{Q^2}{EI}; \\ \text{no condition on } m_b \text{ and } m_c & \text{if } \ K \ge \dfrac{Q^2}{EI}. \end{cases} \tag{3.3}$$

Under condition (3.3), one can verify easily

$$\max\left\{0,\ \frac{Q - \sqrt{KEI}}{m_c}\right\} < \frac{Q}{m_c + m_b}.$$

Assume throughout this paper that

$$\max\left\{0,\ \frac{Q - \sqrt{KEI}}{m_c}\right\} < c^2 < \frac{Q}{m_c + m_b}, \tag{3.4}$$

which is exactly condition (2.2) if $K \ge Q^2/EI$ and is stronger than condition (2.2) if $0 < K < Q^2/EI$. We have the following useful lemma.

LEMMA 3.2. *Assume conditions* (3.3) *and* (3.4) *are satisfied. Then*

$$C_1\|v\|_{H^1}^2 \le \|v\|_*^2 \le C_2\|v\|_{H^1}^2 \qquad \forall v \in H^1(R),$$

where

$$C_1 = \min\left\{Q - c^2 m_c,\ K - \frac{(Q - c^2 m_c)^2}{EI}\right\} > 0 \quad \textit{and} \quad C_2 = \max\left\{K,\ Q - c^2 m_c\right\} > 0.$$

Proof. For any $v \in H^1(R)$, denote by $\widehat{v}$ the Fourier transform of v, which is defined by

$$\widehat{v}(\xi) = \mathcal{F}(v)(\xi) = \frac{1}{\sqrt{2\pi}} \int_R e^{-i\xi s} v(s) ds.$$

By Plancherel's theorem, we have

$$\|v\|_{H^1}^2 = \int_R \left(1 + \xi^2\right) |\widehat{v}(\xi)|^2 d\xi.$$

Note that

$$\begin{aligned} \|v\|_*^2 &= \int_R \mathcal{A}v \cdot v ds + K \int_R v^2 ds \\ &= \int_R \widehat{v}(\xi) \overline{\widehat{\mathcal{A}v}(\xi)} d\xi + K \int_R |\widehat{v}(\xi)|^2 d\xi \\ &= \int_R (H(\xi) + K) |\widehat{v}(\xi)|^2 d\xi, \end{aligned}$$

where

$$H(\xi) = \frac{(Q - c^2 m_c)(EI\xi^4 - c^2 m_b \xi^2)}{EI\xi^2 + Q - c^2(m_c + m_b)}. \tag{3.5}$$

Under condition (3.4), we have

$$H(\xi) + K \le (Q - c^2 m_c)\xi^2 + K \le C_2(1 + \xi^2),$$

where $C_2 = \max\left\{K,\ Q - c^2 m_c\right\} > 0$. Note that

$$\begin{aligned} H(\xi) + K &= (Q - c^2 m_c)\xi^2 - (Q - c^2 m_c)^2 \frac{\xi^2}{EI\xi^2 + Q - c^2(m_c + m_b)} + K \\ &\ge (Q - c^2 m_c)\xi^2 - \frac{(Q - c^2 m_c)^2}{EI} + K. \end{aligned}$$

Under conditions (3.3) and (3.4), we have $K - \dfrac{(Q - c^2 m_c)^2}{EI} > 0$. Let

$$C_1 = \min\left\{Q - c^2 m_c,\ K - \frac{(Q - c^2 m_c)^2}{EI}\right\} > 0;$$

then

$$H(\xi) + K \ge C_1(\xi^2 + 1).$$

Thus upon applying Plancherel's theorem, we have

$$C_1 \|v\|_{H^1}^2 \le \|v\|_*^2 \le C_2 \|v\|_{H^1}^2 \qquad \forall v \in H^1(R). \qquad \square$$

LEMMA 3.3. *For any* $v \in H^1(R)$,

$$\|v\|_\infty \le \sqrt{2}\|v\|_{H^1},$$

where $\|v\|_\infty = \sup_{s\in R}|v(s)|$.

The proof of this lemma can be found in [4, Lemma 2.3].

To prove the existence of nontrivial critical points of $I(v)$ in $H^1(R)$, we use the mountain pass lemma and the concentration-compactness principle of P. L. Lions [18].

THEOREM 3.4 (mountain pass lemma [15]). *Let* X *be a real Banach space,* $B_\rho = \{v \in X \,|\, \|v\|_X < \rho\}$, *and* $\partial B_\rho = \{v \in E \,|\, \|v\|_X = \rho\}$. *Assume that* $J \in C^1(X, R)$ *satisfies the following conditions:*

(a) *for some* $v_0 \in X$, *there are constants* ρ, $\alpha > 0$ *such that* $J|_{v_0+\partial B_\rho} \ge J(v_0)+\alpha$;

(b) *there is an* $e \in X\backslash\overline{v_0 + B_\rho}$ *such that* $J(e) \le J(v_0)$.

Then there exists a sequence $\{v_n\}$ *in* X *such that* $J(v_n) \to \beta$ *and* $J'(v_n) \to 0$ *as* $n \to \infty$, *where* β *can be characterized by*

$$\beta = \inf_{\gamma\in\Gamma}\max_{v\in\gamma([0,1])} J(v),$$

where $\Gamma = \{\gamma \in C([0,1], X) \mid \gamma(0) = v_0,\ \gamma(1) = e\}$.

We are going to show that the functional $I(v)$ given by (3.1) satisfies the ingredients in the mountain pass lemma. Note that $I(0) = 0$ and $I \in C^1(H^1(R), R)$ by Lemma 3.1.

LEMMA 3.5. *Assume conditions* (3.3) *and* (3.4) *are satisfied. Then there exist constants* ρ, $\alpha > 0$ *such that*

$$I(v) \ge \alpha \qquad \forall v \in \partial B_\rho,$$

where $B_\rho = \{v \in H^1(R) \,|\, \|v\|_{H^1} < \rho\}$.

Proof. By Lemma 3.3, $\|v\|_\infty \le \sqrt{2}\|v\|_{H^1}$ for any $v \in H^1(R)$.

Let $\rho = \frac{\sqrt{2}}{2}$. Thus for any $v \in B_\rho$, $\|v\|_\infty < 1$; hence $v(s)+1 > 0$ on R. Therefore, by Lemma 3.2,

$$\begin{aligned} I(v) &= \frac{1}{2}\int_R \mathcal{A}v\cdot v ds + \frac{K}{2}\int_R [(v+1)^2 - 1]ds - K\int_R v ds \\ &= \frac{1}{2}\left[\int_R \mathcal{A}v\cdot v ds + K\int_R v^2 ds\right] = \frac{1}{2}\|v\|_*^2 \ge \frac{C_1}{2}\|v\|_{H^1}^2. \end{aligned}$$

Thus 0 is a local minimum of $I(v)$ in $H^1(R)$. Let $\alpha = \frac{\sqrt{2}C_1}{4}$. Then

$$I(v) \ge \alpha \qquad \forall v \in \partial B_\rho. \qquad \square$$

LEMMA 3.6. *Assume conditions* (3.3) *and* (3.4) *are satisfied. Then there exists an* $e \in H^1(R)$ *such that* $I(e) \le I(0) = 0$.

Proof. Let $\varphi \in H^1(R)$ be given such that $\varphi \le 0$ on R and

$$\overline{\text{supp}(\varphi)} = \overline{\{s \in R \mid \varphi(s) \ne 0\}}$$

is compact. Note that

$$\int_R \xi^2|\widehat{\varphi}(\xi)|^2 d\xi \le \int_R (1+\xi^2)|\widehat{\varphi}(\xi)|^2 d\xi = \|\varphi\|_{H^1}^2.$$

Let ε be a given very small positive number; then there exists an $M > 0$ such that

$$\int_{|\xi|\geq M} \xi^2 |\widehat{\varphi}(\xi)|^2 d\xi \leq \varepsilon$$

and

$$\int_R \xi^2 |\widehat{\varphi}(\xi)|^2 d\xi - \varepsilon \leq \int_{|\xi|<M} \xi^2 |\widehat{\varphi}(\xi)|^2 d\xi \leq \int_R \xi^2 |\widehat{\varphi}(\xi)|^2 d\xi + \varepsilon.$$

Let $\psi_\lambda(s) = \varphi(\lambda s)$ for some $\lambda > 0$.

Then

$$\int_R \mathcal{A}\psi_\lambda \cdot \psi_\lambda ds = \int_R \widehat{\psi}_\lambda(\xi) \overline{\widehat{\mathcal{A}\psi}_\lambda(\xi)} d\xi = \frac{1}{\lambda} \int_R H(\lambda\xi) |\widehat{\varphi}(\xi)|^2 d\xi,$$

where $H(\xi)$ is defined by (3.5). Thus

$$\begin{aligned}
&\int_R \mathcal{A}\psi_\lambda \cdot \psi_\lambda ds \\
&= \lambda(Q - c^2 m_c) \int_R \left(1 - \frac{Q - c^2 m_c}{EI\lambda^2\xi^2 + Q - c^2(m_c + m_b)}\right) \xi^2 |\widehat{\varphi}(\xi)|^2 d\xi \\
&\leq \lambda(Q - c^2 m_c) \left[\varepsilon + \int_{|\xi|<M} \left(1 - \frac{Q - c^2 m_c}{EI\lambda^2\xi^2 + Q - c^2(m_c + m_b)}\right) \xi^2 |\widehat{\varphi}(\xi)|^2 d\xi\right] \\
&\leq \lambda(Q - c^2 m_c) \left[\varepsilon + \int_{|\xi|<M} \left(1 - \frac{Q - c^2 m_c}{EI\lambda^2 M^2 + Q - c^2(m_c + m_b)}\right) \xi^2 |\widehat{\varphi}(\xi)|^2 d\xi\right] \\
&= \lambda(Q - c^2 m_c) \left[\varepsilon - \frac{c^2 m_b - EI\lambda^2 M^2}{EI\lambda^2 M^2 + Q - c^2(m_c + m_b)} \int_{|\xi|<M} \xi^2 |\widehat{\varphi}(\xi)|^2 d\xi\right].
\end{aligned}$$

Under conditions (3.3) and (3.4), we can choose λ to be a very small positive number such that

$$\int_R \mathcal{A}\psi_\lambda \cdot \psi_\lambda ds \leq \lambda(Q - c^2 m_c) \left[\varepsilon - \frac{c^2 m_c - EI\lambda^2 M^2}{EI\lambda^2 M^2 + Q - c^2(m_c + m_b)} \left(\int_R \xi^2 |\widehat{\varphi}(\xi)|^2 d\xi - \varepsilon\right)\right] < 0.$$

Let $\psi_0(s) = \psi_{\lambda_0}(s)$ for some $\lambda_0 > 0$ such that

$$\int_R \mathcal{A}\psi_{\lambda_0} \cdot \psi_{\lambda_0} ds = -\delta < 0. \tag{3.6}$$

Then, for any $B > 0$,

$$\begin{aligned}
\frac{1}{B} I(B\psi_0) = \frac{B}{2} \int_R \mathcal{A}\psi_0 \cdot \psi_0 ds + \frac{KB}{2} \int_{0>\psi_0>-1/B} \psi_0^2 ds \\
- K \int_{\psi_0 \leq -1/B} \left(\psi_0 + \frac{1}{2B}\right) ds.
\end{aligned} \tag{3.7}$$

Since φ has a compact support, $\psi_0(s) = \varphi(\lambda_0 s)$ also has a compact support. Let $\mu = |\operatorname{supp}(\psi_0)| < \infty$. Then we have

$$0 \leq \lim_{B\to\infty} \frac{KB}{2} \int_{0>\psi_0>-1/B} \psi_0^2 ds \leq \lim_{B\to\infty} \frac{K\mu}{2B} = 0 \tag{3.8}$$

and

$$0 \le \lim_{B\to\infty} \int_{\psi_0 \le -1/B} \frac{1}{2B} ds \le \lim_{B\to\infty} \frac{\mu}{2B} = 0. \tag{3.9}$$

By Lemma 3.3, we have

$$0 \le \lim_{B\to\infty} \int_{\psi_0 \le -1/B} |\psi_0| ds \le \|\psi_0\|_\infty \mu \le \sqrt{2}\|\psi_0\|_{H^1}\mu < \infty. \tag{3.10}$$

Thus it follows from (3.6)–(3.10) that

$$\lim_{B\to\infty} \frac{1}{B} I(B\psi_0) \le -\lim_{B\to\infty} \frac{B\delta}{2} + \sqrt{2}\|\psi_0\|_{H^1}\mu = -\infty.$$

Therefore, there exists a $B_0 > 0$ such that $I(B_0\psi_0) < 0$. Then we can choose $e = B_0\psi_0 \in H^1(R)$, which satisfies $I(e) < 0$. □

By Lemmas 3.1, 3.5, and 3.6, $I(v)$ satisfies conditions in the mountain pass lemma. Thus, by applying the mountain pass lemma, we have the following lemma.

LEMMA 3.7. *Assume conditions* (3.3) *and* (3.4) *are satisfied. Then there exists a sequence* $\{v_n\}$ *in* $H^1(R)$ *such that* $I(v_n) \to \beta$ *and* $I'(v_n) \to 0$ *as* $n \to \infty$, *where* β *can be characterized by*

$$\beta = \inf_{\gamma\in\Gamma} \max_{v\in\gamma([0,1])} I(v),$$

where $\Gamma = \{\gamma \in C([0,1], H^1(R)) \mid \gamma(0) = 0,\ \gamma(1) = e\}$.

Note that $0 < \alpha < \beta < \infty$ and, if $v \in H^1(R)$ is a solution of (2.7), then any translation of $v(s)$ is also a solution of (2.7). Thus the sequence $\{v_n\}$ given in Lemma 3.7 does not necessarily have a convergent subsequence. By using the concentration-compactness principle of P. L. Lions [18], we show that $I(v)$ has a nontrivial critical point in $H^1(R)$.

LEMMA 3.8. *Assume conditions* (3.3) *and* (3.4) *are satisfied. Then the sequence* $\{v_n\}$ *defined in Lemma* 3.7 *is bounded in* $H^1(R)$.

Proof. Suppose $\{v_n\}$ is unbounded in $H^1(R)$. Then $\{v_n\}$ has a subsequence, denoted also by $\{v_n\}$, such that $\|v_n\|_* \to \infty$ as $n \to \infty$. Since $I(v_n) \to \beta < \infty$ and $I'(v_n) \to 0$, we have $\frac{I(v_n)}{\|v\|_*} \to 0$ and $\frac{I'(v_n)v_n}{\|v_n\|_*} \to 0$. By Lemmas 3.2 and 3.3, we have $\|v_n\|_\infty \le \sqrt{2/C_1}\|v_n\|_*$. By Lemma 3.1, we have

$$2I(v_n) = \|v_n\|_*^2 + K\int_{v_n\le -1} (-v_n^2 - 2v_n - 1)ds$$

and

$$I'(v_n)v_n = \|v_n\|_*^2 + K\int_{v_n\le -1} (-v_n^2 - v_n)ds.$$

Thus

$$\begin{aligned}
0 &= \lim_{n\to\infty} \frac{I'(v_n)v_n}{\|v_n\|_*^2} \\
&= 1 + K \lim_{n\to\infty} \int_{v_n\le -1} \frac{-v_n^2 - v_n}{\|v_n\|_*^2} ds \\
&= 1 - K \lim_{n\to\infty} \int_{v_n\le -1} \frac{|v_n|(|v_n|-1)}{\|v_n\|_*^2} ds \\
&\ge 1 - K \lim_{n\to\infty} \frac{\|v_n\|_\infty}{\|v_n\|_*} \int_{v_n\le -1} \frac{|v_n|-1}{\|v_n\|_*} ds \\
&\ge 1 - K \frac{\sqrt{2}}{\sqrt{C_1}} \lim_{n\to\infty} \int_{v_n\le -1} \frac{|v_n|-1}{\|v_n\|_*} ds \\
&= 1 - \frac{\sqrt{2}}{\sqrt{C_1}} \lim_{n\to\infty} \frac{2I(v_n) - I'(v_n)v_n}{\|v_n\|_*} \\
&= 1,
\end{aligned}$$

which is a contradiction. Therefore $\{v_n\}$ is bounded. □

LEMMA 3.9. *Assume conditions* (3.3) *and* (3.4) *are satisfied. Then there is a* $\tilde{v} \in H^1(R)$, $\tilde{v} \neq 0$, *such that* $I'(\tilde{v}) = 0$.

Proof. By Lemma 3.8, $\{v_n\}$ defined in Lemma 3.7 is bounded. Let $M > 0$ such that $\|v_n\|_* \le M$ for any n. Note that

$$2\beta = \lim_{n\to\infty} (2I(v_n) - I'(v_n)v_n) = \lim_{n\to\infty} \int_{v_n\le -1} (-v_n - 1) ds$$

and

$$\int_{v_n\le -1} (-v_n - 1) ds \le \int_{v_n\le -1} |v_n|^3 ds \le \int_R |v_n|^3 ds \le \|v_n\|_\infty \|v_n\|_{H^1}^2.$$

Thus there exists $n_0 > 0$ such that, for $n \ge n_0$, we have

$$0 < \beta < \|v_n\|_\infty \|v_n\|_{H^1}^2 \le \frac{1}{C_1} \|v_n\|_\infty \|v_n\|_*^2 \le \frac{M^2}{C_1} \|v_n\|_\infty.$$

Thus

$$\|v_n\|_\infty = \sup_{s\in R} |v_n(s)| \ge \frac{C_1\beta}{M^2} = C_0 > 0 \qquad \forall n \ge n_0.$$

Let $n \ge n_0$. For each $v_n \in H^1(R)$, let $s_n \in R$ such that $v_n(s_n) = \|v_n\|_\infty$. Let $\tilde{v}_n(s) = v_n(s_n + s)$. Then $\{\tilde{v}_n\} \subset H^1(R)$ is bounded and $\{\tilde{v}_n\}$ has a weakly convergent subsequence which converges to $\tilde{v} \in H^1(R)$ [18]. Since $\tilde{v}_n(0) \ge C_0$, we have $\tilde{v}(0) \ge C_0$. Thus $\tilde{v} \neq 0$. Note that $I(\tilde{v}_n) \to \beta$ and $I'(\tilde{v}_n) \to 0$ as $n \to \infty$. We will show $I'(\tilde{v}) = 0$.

For any $\varphi \in H^1(R)$, by Lemma 3.2 and the definition of weak convergence, we have

$$\lim_{n\to\infty} \int_R (\mathcal{A}\tilde{v}_n \cdot \varphi + K\tilde{v}_n\varphi)\, ds = \int_R (\mathcal{A}\tilde{v} \cdot \varphi + K\tilde{v}\varphi)\, ds. \tag{3.11}$$

For any given $\varepsilon > 0$, we can choose a compact set $K \subset R$ such that

$$\|\varphi\|_{L^2(R\setminus K)} < \frac{\varepsilon}{2N_1},$$

where N_1 is a constant such that $\|\tilde{v}_n\|_2 + \|\tilde{v}\|_2 \le N$ for any n. Then

$$\begin{aligned}
&\left| \int_{R\setminus K} \left[(\tilde{v}_n + 1)^- - (\tilde{v} + 1)^- \right] \varphi ds \right| \\
&\le \int_{R\setminus K} \left| (\tilde{v}_n + 1)^- - (\tilde{v} + 1)^- \right| |\varphi| ds \\
&\le \int_{R\setminus K} |\tilde{v}_n - \tilde{v}| \, |\varphi| ds \\
&\le \|\tilde{v}_n - \tilde{v}\|_{L^2(R\setminus K)} \|\varphi\|_{L^2(R\setminus K)} \\
&\le \frac{\varepsilon}{2}.
\end{aligned}$$

By the Sobolev imbedding theorem [1], the weak convergence of $\{\tilde{v}_n\}$ in $H^1(R)$ implies the uniform convergence of $\{\tilde{v}_n\}$ in $C(K)$. Thus there exists an $n_1 > 0$ such that for any $n \ge n_1$,

$$\sup_{s\in K} |\tilde{v}_n(s) - \tilde{v}(s)| < \frac{\varepsilon}{2N_2},$$

where N_2 is a constant such that $\displaystyle\int_K |\varphi| ds \le N_2$. Thus

$$\begin{aligned}
&\left| \int_K \left[(\tilde{v}_n + 1)^- - (\tilde{v} + 1)^- \right] \varphi ds \right| \\
&\le \int_K \left| (\tilde{v}_n + 1)^- - (\tilde{v} + 1)^- \right| |\varphi| ds \\
&\le \int_K |\tilde{v}_n - \tilde{v}| \, |\varphi| ds \\
&\le \frac{\varepsilon}{2}.
\end{aligned}$$

Therefore for any $\varphi \in H^1(R)$

$$\lim_{n\to\infty} \int_R (\tilde{v}_n + 1)^- \varphi ds = \int_R (\tilde{v} + 1)^- \varphi ds. \tag{3.12}$$

By Lemma 3.1, (3.11), and (3.12), we have

$$0 = \lim_{n\to\infty} I'(\tilde{v}_n)\varphi = I'(\tilde{v})\varphi \qquad \forall \varphi \in H^1(R).$$

Thus $\tilde{v}$ is a nontrivial critical point of $I(v)$ on $H^1(R)$. □

By Lemmas 3.1 and 3.9, we have the following theorem.

THEOREM 3.10. *Let conditions* (3.3) *and* (3.4) *be satisfied. Then* (2.7) *admits at least one nontrivial solution in* $H^1(R)$.

By Theorems 2.1 and 3.10, and by using the bootstrapping technique, we obtain the main theorem in this paper.

THEOREM 3.11. *Let conditions* (3.3) *and* (3.4) *be satisfied. Then* (2.1) *admits at least one nontrivial solution* $(y, z) \in H^2(R) \times H^4(R)$. *Consequently, the suspension bridge system* (1.2) *admits at least one nontrivial traveling wave.*

It should be pointed out that Lemmas 3.5–3.7 may shed light on computing solutions of (2.7) by using the numerical mountain pass algorithm [5]. This algorithm has been applied successfully to the single Lazer–McKenna suspension bridge equation (the second equation of (1.2) with $u = 0$) to obtain traveling waves numerically [4]. Due to the facts that $\mathcal{A}$ is a pseudodifferential operator and that the domain is the real line R, the implementation of the numerical mountain pass algorithm to $I(v)$ defined by (3.1) is not trivial. Since the finite difference method or the finite element method cannot be used directly to handle the pseudodifferential operator $\mathcal{A}$, some different approximation methods would be needed to overcome this difficulty. By devising a wavelet-type approximation method, we are able to implement the numerical mountain pass algorithm to $I(v)$ and obtain numerically some traveling waves in the suspension bridge system (1.2). The detailed description and numerical examples will be reported in a separate paper.

Acknowledgment. The author thanks the University of Nevada Las Vegas for its support.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1978.
[2] O. H. AMANN, T. VON KARMAN, AND G. B. WOODRUFF, *The Failure of the Tacoma Narrows Bridge*, Federal Works Agency, Washington, D.C., 1941.
[3] Q. H. CHOI, T. JUNG, AND P. J. MCKENNA, *The study of a nonlinear suspension bridge equation by a variational reduction method*, Appl. Anal., 50 (1993), pp. 73–92.
[4] Y. CHEN AND P. J. MCKENNA, *Traveling waves in a nonlinearly suspended beam: Theoretical results and numerical observations*, J. Differential Equations, 136 (1997), pp. 325–355.
[5] Y. S. CHOI AND P. J. MCKENNA, *A mountain pass method for the numerical solutions of semilinear elliptic problems*, Nonlinear Anal., 20 (1993), pp. 417–437.
[6] Z. DING, *Nonlinear periodic oscillations in suspension bridges*, in Control of Nonlinear Distributed Systems, Lecture Notes in Pure and Appl. Math. 218, Marcel Dekker, New York, 2000, pp. 69–84.
[7] Z. DING, *Nonlinear periodic oscillations in a suspension bridge system under periodic external aerodynamic forces*, Nonlinear Anal., 49 (2002), pp. 1079–1097.
[8] Z. DING, *On nonlinear oscillations in a suspension bridge system*, Trans. Amer. Math. Soc., 354 (2002), pp. 265–274.
[9] Z. DING, *Multiple periodic oscillations in a nonlinear suspension bridge system*, J. Math. Anal. Appl., 269 (2002), pp. 726–746.
[10] J. GLOVER, A. C. LAZER, AND P. J. MCKENNA, *Existence and stability of large scale nonlinear oscillations in suspension bridges*, Z. Angew. Math. Phys., 40 (1989), pp. 172–200.
[11] L. D. HUMPHREYS, *Numerical mountain pass solutions of a suspension bridge equation*, Nonlinear Anal., 28 (1997), pp. 1811–1826.
[12] L. D. HUMPHREYS AND P. J. MCKENNA, *Multiple periodic solutions for a nonlinear suspension bridge equation*, IMA J. Appl. Math., 63 (1999), pp. 37–49.
[13] A. C. LAZER AND P. J. MCKENNA, *Large-amplitude periodic oscillations in suspension bridges: Some new connections with nonlinear analysis*, SIAM Rev., 32 (1990), pp. 537–578.
[14] A. C. LAZER AND P. J. MCKENNA, *Large scale oscillation behavior in loaded asymmetric systems*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 4 (1987), pp. 243–274.

[15] J. Mawhin and M. Willem, *Critical Point Theory and Hamiltonian Systems*, Springer-Verlag, New York, 1989.
[16] P. J. McKenna and W. Walter, *Nonlinear oscillations in a suspension bridge*, Arch. Ration. Mech. Anal., 98 (1987), pp. 167–177.
[17] P. J. McKenna and W. Walter, *Traveling waves in a suspension bridge*, SIAM J. Appl. Math., 50 (1990), pp. 703–715.
[18] M. Struwe, *Variational Methods*, 2nd ed., Springer-Verlag, New York, 1996.

SIAM J. MATH. ANAL.
Vol. 35, No. 1, pp. 172–182

THE CALDERON PROBLEM FOR TWO-DIMENSIONAL MANIFOLDS BY THE BC-METHOD*

M. I. BELISHEV†

Dedicated to the jubilee of O. A. Ladyzhenskaya

Abstract. As was shown by Lassas and Uhlmann [*Ann. Sci. École Norm. Sup.* (4), 34 (2001), pp. 771–787], the smooth two-dimensional compact orientable Riemann manifold with the boundary is uniquely determined by its Dirichlet-to-Neumann map (DN-map) up to conformal equivalence. We give a new proof of this fact based on relations between the Calderon problem and function algebras: the manifold is identified with the spectrum of the algebra of holomorphic functions determined by the DN-map up to isometry; as such, the manifold is recovered from the DN-map by the use of the Gelfand transform. A simple formula linking the DN-map to the Euler characteristic of the manifold is derived.

Key words. Riemann manifolds, Dirichlet-to-Neumann map, reconstruction problem, function algebras, BC-method

AMS subject classifications. 35R30, 35J25, 46J20, 58J32

DOI. 10.1137/S0036141002413919

Introduction.

0.1. General statement. Let (Ω, g) be a smooth[1] n-dimensional Riemann manifold with the boundary Γ; let g be the metric tensor; and let Δ_g be the Beltrami–Laplace operator. Consider the Dirichlet problem

$$\Delta_g u = 0 \quad \text{in } \operatorname{int}\Omega, \tag{0.1}$$

$$u = f \quad \text{on } \Gamma; \tag{0.2}$$

let $u = u^f(x)$ be the solution for a smooth f. With the problem (0.1), (0.2) one associates the Dirichlet-to-Neumann map (DN-map) $\Lambda_g : f \to \frac{\partial u^f}{\partial \nu}|_\Gamma$ (ν is the outward normal).

Assume that the boundary Γ is given and the operator $\Lambda_g : C^\infty(\Gamma) \mapsto C^\infty(\Gamma)$ is known; the *Calderon problem* is to recover (Ω, g). In another formulation, one needs to construct a manifold with a prescribed DN-map.

0.2. The case of $n = 2$. The results. The well-known peculiarity of the 2-dimensional case is the following. If g' and g'' are two metrics on Ω such that $g'' = \rho g'$ with a positive function ρ (i.e., g'' is a conformal deformation of g'), then the DN-maps of the manifolds (Ω, g') and (Ω, g'') are connected through the relation $\Lambda_{g''} = \frac{1}{\sqrt{\rho}}\Lambda_{g'}$, whereas the additional condition $\rho|_\Gamma = 1$ implies $\Lambda_{g''} = \Lambda_{g'}$. This motivates the following definition.

*Received by the editors August 29, 2002; accepted for publication (in revised form) January 15, 2003; published electronically June 25, 2003. This work was supported by RFBR grants 02-01-00260 and 00-15-96017.

http://www.siam.org/journals/sima/35-1/41391.html

†Saint-Petersburg Department of the Steklov Mathematical Institute (POMI), 27 Fontanka, St. Petersburg 191011, Russia (belishev@pdmi.ras.ru).

[1]Everywhere in the paper "smooth" means C^∞-smooth.

Let (Ω', g') and (Ω'', g'') be two smooth manifolds with the common boundary $\partial\Omega' = \partial\Omega'' = \Gamma$; we call them *conformally equivalent* if there exist a diffeomorphism $\beta : \Omega' \mapsto \Omega'', \quad \beta(\Omega') = \Omega'', \quad \beta|_\Gamma = \mathrm{id}$ and a positive function $\rho \in C^\infty(\Omega'), \quad \rho|_\Gamma = 1$ such that β is an isometry of $(\Omega', \rho g')$ onto (Ω'', g'').

The conformal equivalence of manifolds implies the coincidence of their DN-maps. The remarkable fact is that the converse is also true.

THEOREM 1. *Two 2-dimensional compact orientable manifolds with the common boundary are conformally equivalent iff their DN-maps coincide.*

In other words, in dimension 2 the class of conformally equivalent manifolds is determined by its DN-map. This fact was first established by Lassas and Uhlmann in [7] by means of a technique based on the analytic continuation.

This paper is a corrected and extended version of the preprint [3]. Our results are the following:

- we show a relationship between the Calderon problem and function algebras and give a new proof of Theorem 1 exploiting this relationship;
- a simple formula (see (1.6)) linking the DN-map to the Euler characteristic of the manifold is derived.

The paper is addressed to the specialists in inverse problems. Perhaps it is of some interest as an application of the commutative Banach algebras.

0.3. A bit of philosophy. Setting the goal to construct the manifold with a prescribed DN-map, one encounters the naive question: What type of "material" can it be constructed from?

A traditional situation in the boundary value inverse problems is the following: a domain $\Omega \subset \mathbf{R}^n$ *is given* whereas a function $c(x)$ on Ω (density, conductivity, etc.) has to be found through inverse data. Thus, we possess an initial reserve of points Ω, can mark an $x_0 \in \Omega$, and then discuss a possibility to determine $c(x_0)$.

Another situation occurs in the Calderon problem for manifolds: Ω is unknown in itself and has to be recovered, so that the question of "material" turns out to be a fundamental one. In similar dynamic inverse problem this role is played by a subset of the space-time cylinder $\Gamma \times [0, T]$ (the so-called pattern; see [1]). In [7] the manifold is constructed as a Riemann surface (set of germs) of a real analytic function found by the classical procedure of analytic continuation along paths from the boundary.

We propose to identify Ω to the spectrum (the set of multiplicative functionals or, the same, the space of maximal ideals) of some function algebra determined by inverse data. Roughly speaking, the "material" proposed is the set of the Dirac measures $\{\delta_{x_0} \,|\, x_0 \in \Omega\}$. As we show, it is the set which may be recovered through the DN-map.

0.4. The plan. The following basic facts are in use:

(i) (Ω, g) has the complex structure of a Riemann surface; this structure determines the class of metrics conformally equivalent to g;

(ii) the algebra $\mathcal{A}(\Omega)$ of functions continuous in Ω and holomorphic in $\operatorname{int}\Omega$ is nontrivial; functions $w \in \mathcal{A}(\Omega)$ (as local homeomorphisms $\Omega \mapsto \mathbf{C}$) determine the complex structure;

(iii) algebra $\mathcal{A}(\Omega)$ is generic: its (topologized) spectrum is homeomorphic to the manifold, $\operatorname{sp}\mathcal{A}(\Omega) \asymp \Omega$, whereas the algebra itself is identical to its Gelfand transform, $\hat{\mathcal{A}}(\Omega) \equiv \mathcal{A}(\Omega)$;

(iv) the algebra of traces $\mathcal{A}(\Gamma) := \{w|_\Gamma \,|\, w \in \mathcal{A}(\Omega)\}$ is isometrically isomorphic to $\mathcal{A}(\Omega)$ (through the map $\mathrm{tr} : w \mapsto w|_\Gamma$); the isometry yields $\operatorname{sp}\mathcal{A}(\Gamma) \asymp \operatorname{sp}\mathcal{A}(\Omega)$, $\hat{\mathcal{A}}(\Gamma) \equiv \hat{\mathcal{A}}(\Omega)$ that leads to $\operatorname{sp}\mathcal{A}(\Gamma) \asymp \Omega$ and $\hat{\mathcal{A}}(\Gamma) \equiv \mathcal{A}(\Omega)$;

(v) the algebra $\mathcal{A}(\Gamma)$ is determined by the DN-map Λ_g.

To solve the Calderon problem we use these facts in reverse order:

(α) from the operator Λ_g one recovers the trace algebra $\mathcal{A}(\Gamma)$;

(β) finding its spectrum and Gelfand transform we get $\Omega \asymp \operatorname{sp}\mathcal{A}(\Gamma)$ and $\mathcal{A}(\Omega) \equiv \hat{\mathcal{A}}(\Gamma)$;

(γ) using functions of algebra $\mathcal{A}(\Omega)$ we endow Ω with the complex structure;

(δ) introducing a metric g on Ω conformal to the complex structure we get the manifold (Ω, g) whose DN-map coincides with Λ_g by construction.

The procedure (α)$-$(δ) gives a canonical representative of the class of conformally equivalent manifolds which has the given DN-map. The assertion of Theorem 1 is a simple corollary of determinacy of this procedure.

0.5. A little more philosophy. Let us explain why we look at our approach as a variant of the BC-method. In dynamic problems the BC-method

(i) constructs a model (isometric copy) of the initial system through inverse data;

(ii) proposes a mechanism (the amplitude formulas of geometric optics) recovering the initial system through the model.

(See [1], [2].) The same is done in this paper: DN-map determines trace algebra $\mathcal{A}(\Gamma)$, which is a model of algebra $\mathcal{A}(\Omega)$; the Gelfand transform recovers $\mathcal{A}(\Omega)$ (together with manifold Ω). Moreover, there are some reasons to believe that the mechanism based on geometric optics formulas is in fact a kind of the Gelfand transform.

1. Harmonic functions and fields.

1.1. Harmonic functions. So, we have a 2-dimensional smooth orientable compact (Ω, g) with the boundary Γ. Just for simplicity we assume that Γ is homeomorphic to the circle (consists of one connected component).

Let $\operatorname{Harm}_g\Omega := \{u \,|\, \Delta_g\, u = 0 \quad \text{in } \operatorname{int}\Omega\}$ be the set of functions harmonic in Ω; in local coordinates one has

$$\frac{\partial}{\partial x^i}\det{}^{\frac{1}{2}} g(x)\, g^{ij}(x)\frac{\partial u(x)}{\partial x^j} = 0. \tag{1.1}$$

Solutions of the problem (0.1), (0.2) are said to be *potentials*; the family of potentials $\mathcal{U} := \{u^f \,|\, f \in C^\infty(\Gamma)\}$ coincides with $\operatorname{Harm}_g\Omega \bigcap C^\infty(\Omega)$.

1.2. Rotation operator. Introduce near Γ the semigeodesic coordinates γ, τ such that $ds^2 = Ed\gamma^2 + d\tau^2$ and $E|_\Gamma = 1$, so that γ is the natural parameter on Γ and $\tau(x) = \operatorname{dist}(x, \Gamma)$. The field $\frac{\partial}{\partial\gamma}$ fixes an orientation of Γ and an orientation of Ω; note that $\frac{\partial}{\partial\tau} = -\frac{\partial}{\partial\nu}$.

The orientation determines the continuous family of rotations $\Phi(x) \in \operatorname{End} T_x\Omega$ such that $\langle\Phi(x)\, a, \Phi(x)\, b\rangle = \langle a, b\rangle$, $\langle\Phi(x)\, a, a\rangle = 0$ for $a, b \in T_x\Omega$ and $\Phi(\gamma)\frac{\partial}{\partial\nu} = \frac{\partial}{\partial\gamma}$ on Γ. Note that $\Phi^2(x) = -\mathrm{id}$, $x \in \Omega$.

Let $\vec{\mathcal{L}}$ be the (real) space of the square integrable vector fields on (Ω, g) with the inner product $(a, b) = \int_\Omega d\Omega_x\, \langle a(x), b(x)\rangle$. The *rotation operator* $\Phi : \vec{\mathcal{L}} \mapsto \vec{\mathcal{L}}$, $(\Phi y)(x) := \Phi(x)y(x)$, $x \in \Omega$, is a unitary operator in $\vec{\mathcal{L}}$.

1.3. Harmonic fields. Let $\vec{\mathcal{H}} := \{h \in \vec{\mathcal{L}} \,|\, \operatorname{div} h = \operatorname{div}\Phi h = 0 \quad \text{in } \operatorname{int}\Omega\}$ be the (sub)space of *harmonic fields*. Harmonic fields are smooth in $\operatorname{int}\Omega$; the set $\vec{\mathcal{H}}^\infty := \vec{\mathcal{H}} \bigcap \vec{C}^\infty(\Omega)$ is dense in $\vec{\mathcal{H}}$.

The space $\vec{\mathcal{H}}$ contains the subspace of *potential fields* $\vec{\mathcal{E}} := \{\nabla u \in \vec{\mathcal{H}} \,|\, u \in \mathrm{Harm}_g\,\Omega \bigcap H^1(\Omega)\}$ ($H^1(\Omega)$ is the Sobolev class). The set $\vec{\mathcal{E}}^\infty := \vec{\mathcal{E}} \bigcap \vec{C}^\infty(\Omega) = \nabla\mathcal{U}$ is dense in $\vec{\mathcal{E}}$.

Recall in addition that every $h \in \vec{\mathcal{H}}$ is locally potential: for any $x_0 \in \Omega$ there exist a neighborhood U_{x_0} and a function $v \in \mathrm{Harm}_g\,U_{x_0}$ such that $h = \nabla v$ in U_{x_0}.

In what follows we denote by the common symbol "tr" the reduction of fields and functions given in Ω on the boundary Γ. The important fact is that the harmonic field is determined by its trace: $h \in \vec{\mathcal{H}}$ and $\mathrm{tr}\,h = 0$ implies $h = 0$ in Ω. Indeed, fixing an $x_0 \in \Gamma$, in U_{x_0} we have $h = \nabla v$ that leads to $\Delta_g v = 0$ and $\mathrm{tr}\,\nabla v = 0$, yielding $v = \mathrm{const}$ and $h = 0$ in U_{x_0} due to the uniqueness of the solution to the Cauchy problem for the second order elliptic equation. Covering the manifold with neighborhoods and applying the same uniqueness theorem, one easily obtains $h = 0$ everywhere in Ω.

Introduce the subspace $\vec{\mathcal{D}} := \vec{\mathcal{H}} \ominus \vec{\mathcal{E}}$ and list some of its known properties (see, e.g., [10]). This subspace has a finite dimension determined by topology of Ω:

$$\dim\vec{\mathcal{D}} \;=\; \dim\vec{\mathcal{H}}/\vec{\mathcal{E}} \;=\; \beta_1(\Omega) \;=\; 1-\chi(\Omega), \tag{1.2}$$

where β_1 is the first Betti number (dimension of the first homology group of the manifold) and χ is its Euler characteristic. Elements of the subspace are smooth: $\vec{\mathcal{D}} \subset \vec{C}^\infty(\Omega)$. The fields $b \in \vec{\mathcal{D}}$ are tangent on Γ; indeed, for any $f \in C^\infty(\Gamma)$ we have

$$0 = \int_\Omega d\Omega\,\langle \nabla u^f, b\rangle = \int_\Gamma d\gamma\, f\,\left\langle b, \frac{\partial}{\partial\nu}\right\rangle,$$

yielding $\langle b, \frac{\partial}{\partial\nu}\rangle = 0$ or, equivalently, $\mathrm{tr}\,b = \kappa\frac{\partial}{\partial\gamma}$ with a smooth $\kappa = \langle b, \frac{\partial}{\partial\gamma}\rangle$.

1.4. Topology from DN-map. Let $\dot{C}^\infty(\Gamma) := \{f \in C^\infty(\Gamma) \,|\, \int_\Gamma d\gamma\, f(\gamma) = 0\}$ be the subset of smooth functions with zero mean value. For $f \in \dot{C}^\infty(\Gamma)$ we denote by Jf the primitive function with zero mean value: $Jf \in \dot{C}^\infty(\Gamma)$, $\frac{d}{d\gamma}Jf = f$.

Recall the well-known properties of the DN-map: $\mathrm{Ker}\,\Lambda_g = \{\mathrm{const}\}$, $\mathrm{Ran}\,\Lambda_g \subset \dot{C}^\infty(\Gamma)$. Note also that by standard elliptic theory the operator $\Lambda_g J : L_2(\Gamma) \mapsto L_2(\Gamma)$, $\mathrm{Dom}\,\Lambda_g J = \dot{C}^\infty(\Gamma)$ is continuous.

Fix $f \in C^\infty(\Gamma)$; in accordance with the decomposition $\vec{\mathcal{H}} = \vec{\mathcal{E}} \oplus \vec{\mathcal{D}}$ the field $\Phi\nabla u^f \in \vec{\mathcal{H}}$ may be represented in the form

$$\Phi\nabla u^f = \nabla u^p + b^f \tag{1.3}$$

with $\nabla u^p \in \vec{\mathcal{E}}^\infty$ (p being determined by f) and $b^f \in \vec{\mathcal{D}}$; let $\mathrm{tr}\,b^f = \kappa^f\frac{\partial}{\partial\gamma}$.

LEMMA 1. (i) *The equality*

$$[\mathbf{1} + (\mathbf{\Lambda_g J})^\mathbf{2}]\frac{\mathbf{df}}{\mathbf{d}\gamma} = \mathbf{\Lambda_g J}\kappa^\mathbf{f} \tag{1.4}$$

holds.

(ii) *The inclusion $\Phi\nabla u^f \in \vec{\mathcal{E}}^\infty$ is equivalent to the relation*

$$[\mathbf{1} + (\mathbf{\Lambda_g J})^\mathbf{2}]\frac{\mathbf{df}}{\mathbf{d}\gamma} = \mathbf{0}. \tag{1.5}$$

(iii) *The equality*

$$\dim\,[\mathbf{1} + (\mathbf{\Lambda_g J})^\mathbf{2}]\,\dot{\mathbf{C}}^\infty(\mathbf{\Gamma}) = \mathbf{1} - \chi(\mathbf{\Omega}) \tag{1.6}$$

is valid.

Proof. (i) The traces of the left- and right-hand sides of (1.3) may be written as follows:

$$\operatorname{tr}\Phi\nabla u^f = \Phi\operatorname{tr}\nabla u^f = \Phi\left[\frac{df}{d\gamma}\frac{\partial}{\partial\gamma} + \Lambda_g f\frac{\partial}{\partial\nu}\right]$$

$$= \frac{df}{d\gamma}\Phi\frac{\partial}{\partial\gamma} + \Lambda_g f\,\Phi\frac{\partial}{\partial\nu} = \Lambda_g f\,\frac{\partial}{\partial\gamma} - \frac{df}{d\gamma}\frac{\partial}{\partial\nu} \tag{1.7}$$

and

$$\operatorname{tr}\left[\nabla u^p + b^f\right] = \left(\frac{dp}{d\gamma} + \kappa^f\right)\,\frac{\partial}{\partial\gamma} + \Lambda_g p\,\frac{\partial}{\partial\nu}. \tag{1.8}$$

Using (1.7) and (1.8) we get

$$\Lambda_g f = \frac{dp}{d\gamma} + \kappa^f, \qquad -\frac{df}{d\gamma} = \Lambda_g J\,\frac{dp}{d\gamma}$$

(note that the first equality yields $\kappa^f \in \dot{C}^\infty(\Gamma)$). Eliminating $\frac{dp}{d\gamma}$ one easily gets (1.4).

(ii) In accordance with (1.3) the inclusion $\Phi\nabla u^f \in \vec{\mathcal{E}}^\infty$ is equivalent to $b^f = 0$, which is equivalent to $\kappa^f = 0$ leading to (1.5).

(iii) Let P be the (orthogonal) projection in $\vec{\mathcal{H}}$ onto $\vec{\mathcal{D}}$; we will show that

$$P\Phi\vec{\mathcal{E}}^\infty = \vec{\mathcal{D}}. \tag{1.9}$$

Indeed, if $b \in \vec{\mathcal{D}},\ b \perp P\Phi\vec{\mathcal{E}}^\infty$, then for any smooth f one has

$$0 = \int_\Omega d\Omega\,\langle\Phi\nabla u^f, b\rangle = \int_\Omega d\Omega\,\langle\nabla u^f, \Phi^* b\rangle$$

$$= \int_\Gamma d\gamma\,u^f\,\left\langle\frac{\partial}{\partial\nu}, \Phi^* b\right\rangle = \int_\Gamma d\gamma\,f\,\left\langle\Phi\frac{\partial}{\partial\nu}, b\right\rangle = \int_\Gamma d\gamma\,f\,\left\langle\frac{\partial}{\partial\gamma}, b\right\rangle.$$

This leads to $\langle\frac{\partial}{\partial\gamma}, b\rangle = 0$, then $\operatorname{tr} b = \langle b, \frac{\partial}{\partial\gamma}\rangle\frac{\partial}{\partial\gamma} = 0$ (recall that b's are tangent to Γ), and, finally, to $b = 0$. Thus, the set $P\Phi\vec{\mathcal{E}}^\infty$ is dense in $\vec{\mathcal{D}}$, which implies (1.9) in view of $\dim\vec{\mathcal{D}} < \infty$.

Let f in (1.3) run over $C^\infty(\Gamma)$; due to (1.9) the corresponding projections b^f cover $\vec{\mathcal{D}}$. Since the map $b^f \mapsto \kappa^f$ is injective, one has $\dim\{\kappa^f \mid f \in C^\infty(\Gamma)\} = \dim\vec{\mathcal{D}}$. As is easy to see, the operator $\Lambda_g J$ is injective on $\dot{C}^\infty(\Gamma)$; therefore

$$\dim\vec{\mathcal{D}} = \dim\{\kappa^f \mid f \in C^\infty(\Gamma)\} = \dim\{\Lambda_g J\kappa^f \mid f \in C^\infty(\Gamma)\}.$$

Taking into account (1.2), the last equality, and (1.4) we get

$$1 - \chi(\Omega) = \dim\vec{\mathcal{D}} = \dim\left[\mathbf{1} + (\mathbf{\Lambda_g J})^{\mathbf{2}}\right]\frac{\mathbf{d}}{\mathbf{d}\boldsymbol{\gamma}}\mathbf{C}^\infty(\boldsymbol{\Gamma}),$$

which is equivalent to (1.6). The lemma is proved.

In the situation of the Calderon problem, Lemma 1 may be exploited as follows. Assume that Ω is a fortiori known to be homeomorphic to the sphere with handles and one removed disk. In this case $\chi(\Omega) = 1 - 2m$, where m is the number of handles, and one can find m by the formula

$$m = \frac{1}{2} \dim \left[\mathbf{1} + (\mathbf{\Lambda_g J})^{\mathbf{2}}\right] \dot{\mathbf{C}}^{\infty}(\mathbf{\Gamma}).$$

In particular, Ω is homeomorphic to the disk (so that $m = 0$) iff $(\Lambda_g J)^2 = -\mathbf{1}$ on $\dot{C}^{\infty}(\Gamma)$.

Given the DN-map and checking (1.5) one can select a subset of f's in $C^{\infty}(\Gamma)$ such that the "conjugate" fields $\Phi \nabla u^f$ are also potential. Later we'll use this opportunity.

2. Algebras.

2.1. CBA dictionary. We begin with minimal information concerning the commutative Banach algebras (CBAs); for details, see [6], [9].

A. The CBA is a (complex or real) Banach space $\mathcal{A}$ equipped with the multiplication operation ab satisfying $ab = ba$; $||ab|| \leq ||a||\,||b||$, $a, b \in \mathcal{A}$. Example: the algebra $C(X)$ of functions continuous on a (topological) compact X with the norm $||a|| = \max_{\Omega} |a(\cdot)|$; the subalgebras of $C(X)$ are called *function algebras*.

CBA is said to be *uniform* if it has the unit $e : ea = a$ and the relation $||a^2|| = ||a||^2$ holds. All the function algebras are uniform.

B. A subspace $I \neq \mathcal{A}$ is called *ideal* if $ja \in I$ for any $j \in I$, $a \in \mathcal{A}$. Ideal I is *maximal* if for any ideal $\tilde{I} \subset \mathcal{A}$ the relation $I \subset \tilde{I}$ implies $I = \tilde{I}$.

Let $\mathcal{I}$ be the set of maximal ideals of algebra $\mathcal{A}$. Every $I \in \mathcal{I}$ is closed; $\operatorname{codim} I = 1$; the quotient space $\mathcal{A}/I$ is isometrically isomorphic to the field of complex numbers $\mathbf{C}$.

C. Let $\mathcal{A}'$ be the space of continuous functionals on $\mathcal{A}$. A functional $\delta \in \mathcal{A}'$ is called *multiplicative* if $\delta(ab) = \delta(a)\delta(b)$. Example: the Dirac measure $\delta_{x_0} \in C'(X)$: $\delta_{x_0}(a) = a(x_0)$. The set of multiplicative functionals is denoted by $\mathcal{M}$.

D. There exists a canonical bijection between the sets $\mathcal{M}$ and $\mathcal{I}$: if $\delta \in \mathcal{M}$, then $I_\delta := \operatorname{Ker} \delta \in \mathcal{I}$; if $I \in \mathcal{I}$, then the projection $\delta_I : \mathcal{A} \mapsto \mathcal{A}/I = \mathbf{C}$ is an element of $\mathcal{M}$. In what follows we identify $\mathcal{M}$ to $\mathcal{I}$ through this bijection.

E. The *Gelfand transform* maps element $a \in \mathcal{A}$ into function $\hat{a}$ on $\mathcal{M}$ by the rule $\hat{a}(\delta) := \delta(a)$, $\delta \in \mathcal{M}$. The *Gelfand topology* is defined as the weakest topology on $\mathcal{M}$ in which all of $\hat{a}$ are continuous. The set $\mathcal{M}$ equipped with this topology is a compact; this compact is called the *spectrum* (or the maximal ideal space) of the algebra $\mathcal{A}$ and denoted by $\operatorname{sp} \mathcal{A}$.

The Gelfand transform $\hat{\mathcal{A}} := \{\hat{a} \,|\, a \in \mathcal{A}\}$ of algebra $\mathcal{A}$ is a subalgebra of the algebra $C(\operatorname{sp} \mathcal{A})$. If $\mathcal{A}$ is uniform, the map $a \mapsto \hat{a}$ turns out to be an isometric isomorphism (on its image): $(\alpha a + \beta b + cd)^{\wedge} = \alpha\hat{a} + \beta\hat{b} + \hat{c}\hat{d}$, $||\hat{a}|| = ||a||$ for all $\alpha, \beta \in \mathbf{C}$ and $a, b, c, d, \in \mathcal{A}$; in this case $\hat{\mathcal{A}}$ is a canonical realization of $\mathcal{A}$ in the form of function algebra.

F. An isomorphism $t : A(X) \mapsto B(Y)$ of two function algebras is called *spatial* if there exists a bijection $\beta : X \mapsto Y$ such that $tw = w \circ \beta^{-1}$. For a function algebra $\mathcal{A} \subset C(X)$, to each point $x_0 \in X$ one associates the Dirac measure $\delta_{x_0} \in \operatorname{sp} \mathcal{A}$, and because of this the natural embedding $\varepsilon : X \mapsto \operatorname{sp} \mathcal{A}$, $\varepsilon(x_0) = \delta_{x_0}$ occurs.

A function algebra $\mathcal{A} \subset C(X)$ is said to be *generic* if ε is a homeomorphism from X onto $\operatorname{sp} \mathcal{A}$. Generic algebra is spatially isomorphic to its Gelfand transform $\hat{\mathcal{A}}$: $\hat{w} = w \circ \varepsilon^{-1}$.

2.2. Algebra $\mathcal{A}(\Omega)$. The functions $u, u^* \in \mathrm{Harm}_g\,\Omega$ are said to be *conjugate* if they satisfy the Cauchy–Riemann conditions $\nabla u^*(x) = \Phi(x)\nabla u(x)$, $x \in \mathrm{int}\,\Omega$. Note the relation $(u^*)^* = -u$ (mod const). If Ω is homeomorphic to the disk, each harmonic u possesses the conjugate u^*; it is not the case if topology of Ω is non-trivial [5].

We assign a harmonic u to the set $\mathcal{P}$ if it has the conjugate u^* and $u, u^* \in \mathrm{Harm}_g\,\Omega \bigcap C(\Omega)$.

Let us define the family of complex functions $\mathcal{A}(\Omega) := \{w = u + iu^* \,|\, u \in \mathcal{P}\}$ and mention some of its properties [5]:

(i) in proper (isothermal) local coordinates x^1, x^2 each $w \in \mathcal{A}(\Omega)$ is a holomorphic function of $z = x^1 + ix^2$; functions of $\mathcal{A}(\Omega)$, as local homeomorphisms from Ω into $\mathbf{C}$, determine on Ω a complex structure of the Riemann surface;

(ii) for each $w \in \mathcal{A}(\Omega)$ $|w|$ attains maximum on Γ;

(iii) $\mathcal{A}(\Omega)$ is a closed subalgebra of the (complex) algebra $C(\Omega)$;

LEMMA 2. *The algebra $\mathcal{A}(\Omega)$ is generic.*

Proof. (1) The embedding $\varepsilon(\Omega) \subset \mathrm{sp}\,\mathcal{A}(\Omega)$ is a general fact (see section 2.1, item F) and one needs to prove the converse embedding only.

(2) Choose an $I \in \mathrm{sp}\,\mathcal{A}(\Omega)$; we are going to show that there exists a point $x_I \in \Omega$ such that $j(x_I) = 0$ for all $j \in I$. Assume the opposite: for any $x \in \Omega$ one can find $j_x \in I$ such that $j_x(x) \neq 0$. Due to continuity of j_x there exists a neighborhood $U_x \ni x$ such that j_x doesn't vanish in U_x. Since $\bigcup_x U_x$ covers Ω, we can select a finite subcovering $U_{x_1}, \ldots, U_{x_p}$; in this case the corresponding $j_{x_1}, \ldots, j_{x_p}$ have no common zeros in Ω.

(3) Due to smoothness of the boundary Γ the manifold (Ω, g) may be embedded into a larger manifold $(\Omega', g') : \Omega \subset \Omega', \mathrm{dist}_{g'}(\partial\Omega', \Omega) = r_0 > 0, g'|_\Omega = g$. One way to realize such an embedding is to use the semigeodesic coordinates, extending E on $\tau < 0$.

Denote $\Omega_r := \{x \in \Omega' \,|\, \mathrm{dist}_{g'}(x, \Omega) < r\}$, $0 < r < r_0$. Assign a function $w \in \mathcal{A}(\Omega)$ to the family $\mathcal{O}(\Omega)$ if it has a holomorphic continuation in Ω_r (with r depending on w). The family $\mathcal{O}(\Omega)$ is dense in $\mathcal{A}(\Omega)$ (see [4, p. 48, Corollary 2]).

(4) Since $\mathrm{codim}\, I = 1$, the set $I \bigcap \mathcal{O}(\Omega)$ is dense in I. Due to this one can approximate $j_{x_1}, \ldots, j_{x_p}$ by functions $\tilde{j}_{x_1}, \ldots, \tilde{j}_{x_p} \in I \bigcap \mathcal{O}(\Omega)$ so that $\tilde{j}_{x_k} \neq 0$ in U_{x_k} and $\tilde{j}_{x_1}, \ldots, \tilde{j}_{x_p}$ have no common zeros in Ω.

Since the set $\{\tilde{j}_{x_k}\}$ is finite, there exists $\Omega_r \supset \Omega$ such that $\tilde{j}_{x_1}, \ldots, \tilde{j}_{x_p}$ are defined and have no common zeros in Ω_r. Considering Ω_r as a noncompact Riemann surface one can find functions $g_1, \ldots, g_p$ holomorphic in Ω_r and satisfying the condition $g_1\tilde{j}_{x_1} + \cdots + g_p\tilde{j}_{x_p} = 1$ everywhere in Ω_r (see [5, p. 205]). Reducing functions on Ω we get $\tilde{g}_1\tilde{j}_{x_1} + \cdots + \tilde{g}_p\tilde{j}_{x_p} = 1$ in Ω for $\tilde{g}_k := g_k|_\Omega \in \mathcal{A}(\Omega)$ and $\tilde{j}_{x_k} \in I$, which yields the inclusion $1 \in I$. The last leads evidently to $I = \mathcal{A}(\Omega)$, which contradicts I to be a proper ideal.

Thus, all the functions belonging to I vanish in a point $x_I \in \Omega$ and the embedding $\varepsilon(\Omega) \supset \mathrm{sp}\,\mathcal{A}(\Omega)$ is established. So, $\varepsilon(\Omega) = \mathrm{sp}\,\mathcal{A}(\Omega)$, and it remains to show that ε is a homeomorphism.

(5) Identifying maximal ideals of $\mathcal{A}(\Omega)$ to multiplicative functionals we set $I_{x_0} \equiv \delta_{x_0}$, where δ_{x_0} is the Dirac measure. In terms of the bijection $\varepsilon : \Omega \mapsto \mathrm{sp}\,\mathcal{A}(\Omega)$, existence of which has been proved above, the Gelfand transform takes the form $\hat{w}(\delta) := \delta(w) = w(x_\delta) = (w \circ \varepsilon^{-1})(\delta)$ (i.e., is a spatial isomorphism).

Topologizing $\mathrm{sp}\,\mathcal{A}(\Omega)$ by Gelfand, we endow it with the weakest topology in which all $\hat{w} \in \hat{\mathcal{A}}(\Omega)$ are continuous. Since the corresponding topology on Ω determined by

functions $w \in \mathcal{A}(\Omega)$ coincides with the initial (metric) topology of (Ω, g), the bijection ε turns out to be a homeomorphism. The lemma is proved.

The density of $\mathcal{O}(\Omega)$ in $\mathcal{A}(\Omega)$ used in the proof provides one more important property of the algebra $\mathcal{A}(\Omega)$:

(iv) the smooth subalgebra $\mathcal{A}^\infty(\Omega) := \mathcal{A}(\Omega) \bigcap C^\infty(\Omega)$ is dense in $\mathcal{A}(\Omega)$.

2.3. Conforming metrics. The algebra $\mathcal{A}(\Omega)$ as well as the corresponding complex structure on Ω (see (i), section 2.2) are determined by the metric g, the structures corresponding to metrics g and ρg being identical since these metrics define one and the same reserve of harmonic functions.[2]

The converse is true in the following sense. We call a metric g *conforming* to the complex structure determined by the algebra $\mathcal{A}(\Omega)$ if $\mathrm{Re}\mathcal{A}(\Omega) \subset \mathrm{Harm}_g\,\Omega$. It can be shown that any two metrics conforming to the given structure are distinguished by a functional multiplier.

To construct a metric conforming to a given structure one can use the following standard trick. Let us choose an atlas $\{U_k, \varphi_k\} : \bigcup_k U_k \supset \Omega$; $\varphi_k : U_k \mapsto \mathbf{R}^2$, $\varphi_k(x) = \{\mathrm{Re}\, w_k(x), \mathrm{Im}\, w_k(x)\}, w_k \in \mathcal{A}(\Omega)$; let η_k be a subordinated partition of unity: $\eta_k \geq 0$, $\mathrm{supp}\,\eta_k \subset U_k$, $\sum_k \eta_k = 1$. Define on Ω the tensors $g_{ij}^{(k)} := \eta_k \delta_{ij}$. The tensor $g := \sum_k g^{(k)}$ defines a required metric.

Another way to find a conforming metric is to solve (1.1) with respect to the entries $\det{}^{\frac{1}{2}} g\, g^{ij}$ for sufficiently many $u \in \mathrm{Re}\,\mathcal{A}(\Omega)$, i.e., to recover the metric g up to a function multiplier.

2.4. Algebra $\mathcal{A}(\Gamma)$. The set of traces $\mathcal{A}(\Gamma) := \mathrm{tr}\,\mathcal{A}(\Omega) = \{\mathrm{tr}\, w \,|\, w \in \mathcal{A}(\Omega)\}$ is a closed subalgebra of the (complex) algebra $C(\Gamma)$. We list some of its properties.

(i) Since $\max_\Omega |w|$ for $w \in \mathcal{A}(\Omega)$ is attained at Γ, the equality $||w||_{\mathcal{A}(\Omega)} = ||\mathrm{tr}\, w||_{\mathcal{A}(\Gamma)}$ holds. Therefore, the mapping tr is an isometric isomorphism of algebras.

(ii) Since $\mathcal{A}(\Omega)$ is generic, whereas $\mathcal{A}(\Gamma)$ is isometrically isomorphic to $\mathcal{A}(\Omega)$, the Gelfand transform $\hat{\mathcal{A}}(\Gamma) \subset C(\mathrm{sp}\,\mathcal{A}(\Gamma))$ turns out to be spatially isomorphic to $\mathcal{A}(\Omega)$. The corresponding spatial isomorphism from $\mathcal{A}(\Omega)$ onto $\hat{\mathcal{A}}(\Gamma)$ is $\wedge\,\mathrm{tr}$,[3] whereas the corresponding space bijection is $\beta : \Omega \mapsto \mathrm{sp}\,\mathcal{A}(\Gamma)$, $\beta(x_0) = \mathrm{tr}\, I_{x_0}$.

(iii) Define the smooth subalgebra $\mathcal{A}^\infty(\Gamma) := \mathcal{A}(\Gamma) \bigcap C^\infty(\Gamma) = \mathrm{tr}\,\mathcal{A}^\infty(\Omega)$. Since $\mathcal{A}^\infty(\Omega)$ is dense in $\mathcal{A}(\Omega)$ (see (iv), section 2.2) $\mathcal{A}^\infty(\Gamma)$ turns out to be dense in $\mathcal{A}(\Gamma)$.

The following lemma plays a key role: it shows that the trace algebra $\mathcal{A}(\Gamma)$ is determined by the DN-map. Introduce the set

$$\mathcal{F} := \left\{ f \in C^\infty(\Gamma) \,|\, [\mathbf{1} + (\mathbf{\Lambda_g J})^\mathbf{2}] \frac{\mathbf{df}}{\mathbf{d\gamma}} = \mathbf{0} \right\}$$

of (real) functions satisfying (1.5). To each $f \in \mathcal{F}$ we associate the *conjugate function* f^* such that $\frac{df^*}{d\gamma} = \Lambda_g J \frac{df}{d\gamma}$ holds. As is easy to see, f^* is determined by f up to constant, $f^* \in \mathcal{F}$, and $f^{**} = -f \pmod{\mathrm{const}}$.

LEMMA 3. *The representation*

$$\mathcal{A}(\Gamma) = \mathrm{clos}_{C(\Gamma)}\, \{f + if^* \,|\, f \in \mathcal{F}\} \tag{2.1}$$

is valid.

[2] In dimension 2 the equalities $\Delta_g\, u = 0$ and $\Delta_{\rho g}\, u = 0$ are equivalent.

[3] Here and below (section 3.2) $\wedge$ denotes the Gelfand transform $\wedge\mathcal{A} := \hat{\mathcal{A}}$.

Proof. If $w = u + iu^* \in \mathcal{A}^\infty(\Omega)$, then $u = u^f$ with $f = \operatorname{tr} u$, $u^* = u^p$ with $p = \operatorname{tr} u^*$, and $\Phi\nabla u^f = \nabla u^p$. According to (ii) of Lemma 1 the last equality implies that $f, p \in \mathcal{F}$. Taking the traces in the same equality one easily gets $\frac{dp}{d\gamma} = \Lambda_g f = \Lambda_g J \frac{df}{d\gamma}$, so $p = f^*$. Hence $\operatorname{tr} w = f + ip = f + if^* \in \mathcal{A}^\infty(\Gamma)$, which establishes the inclusion $\mathcal{A}^\infty(\Gamma) \subset \{f + if^* \,|\, f \in \mathcal{F}\}$.

Take $f \in \mathcal{F}$ and its conjugate $f^* \in \mathcal{F}$. By virtue of Lemma 1(ii) the fields ∇u^f and ∇u^{f^*} are potential and satisfy

$$\operatorname{tr} \Phi\nabla u^f = \Lambda_g f \frac{\partial}{\partial\gamma} - \frac{df}{d\gamma}\frac{\partial}{\partial\nu} = \frac{df^*}{d\gamma}\frac{\partial}{\partial\gamma} + \Lambda_g f^* \frac{\partial}{\partial\nu} = \operatorname{tr} \nabla u^{f^*}.$$

Equality of the traces leads to equality of the fields: $\Phi\nabla u^f = \nabla u^{f^*}$ in Ω. Hence, the function $w = u^f + iu^{f^*}$ belongs to $\mathcal{A}^\infty(\Omega)$, whereas its trace $f + if^*$ belongs to $\mathcal{A}^\infty(\Gamma)$. Thus, we have $\{f + if^* \,|\, f \in \mathcal{F}\} \subset \mathcal{A}^\infty(\Gamma)$, which establishes the equality $\mathcal{A}^\infty(\Gamma) = \{f + if^* \,|\, f \in \mathcal{F}\}$.

Since $\mathcal{A}^\infty(\Gamma)$ is dense in $\mathcal{A}(\Gamma)$ one gets (2.1). The lemma is proved.

3. Solving the problem.

3.1. Construction of the manifold. So, let an operator Λ_g acting in (real) $C^\infty(\Gamma)$ be given and known to be the DN-map of a manifold. By results of Lee and Uhlmann [8] Λ_g determines the metric tensor at the boundary $g|_\Gamma$; therefore one can also assume the length element $d\gamma$ to be known. Due to the latter the class $\dot{C}^\infty(\Gamma)$ as well as the integration operator J acting in this class are at our disposal.

The following procedure gives a canonical representative of the class of manifolds whose DN-maps coincide with Λ_g.

Step 1. Select the set $\mathcal{F}$ in $C^\infty(\Gamma)$. Through $\mathcal{F}$ recover the algebra $\mathcal{A}(\Gamma)$ by the representation (2.1).

Step 2. Find $\operatorname{sp}\mathcal{A}(\Gamma) =: \Omega$ (see (ii), section 2.4). Identify the subset of ideals $\{I_{\gamma_0} \,|\, \gamma_0 \in \Gamma\} \subset \Omega$ to Γ by the rule $I_{\gamma_0} \equiv \gamma_0$.

Step 3. Find the Gelfand transform $\hat{\mathcal{A}}(\Gamma) =: \mathcal{A}(\Omega)$. Note that the "boundary" $\{I_{\gamma_0} \,|\, \gamma_0 \in \Gamma\}$ constructed in the previous step may also be identified as the Shilov boundary of $\mathcal{A}(\Omega)$ (see, e.g., [9]).

Step 4. Using functions $w \in \mathcal{A}(\Omega)$ as local homeomorphisms $\Omega \mapsto \mathbf{C}$ recover on Ω the complex structure. Its realifying determines on Ω a structure of the smooth 2-dimensional manifold with the boundary Γ.

Step 5. Equip Ω with a metric $\tilde{g}$ conforming to the complex structure (see section 2.3). Choose a positive function $\rho \in C^\infty(\Omega)$ such that the metric $g := \rho\tilde{g}$ satisfies $(\rho\tilde{g})_\Gamma = g|_\Gamma$.

The manifold (Ω, g) solves the Calderon problem: its DN-map coincides with Λ_g by construction.

3.2. Proof of Theorem 1. We recall that $\wedge$ denotes the Gelfand transform $\wedge\mathcal{A} = \hat{\mathcal{A}}$. The statement of the theorem is just a consequence of the well-determinacy of the procedure described above.[4] Indeed,

(1) if (Ω', g') and (Ω'', g'') have the common DN-map, the corresponding sets $\mathcal{F}$ determined by Λ_g coincide: $\mathcal{F}' = \mathcal{F}''$. Whence, by Lemma 3 the trace algebras coincide: $\mathcal{A}'(\Gamma) = \mathcal{A}''(\Gamma)$;

[4]Such a situation is quite usual for the BC-method.

(2) by the property (ii), section 2.4, the algebras $\mathcal{A}(\Omega')$ and $\mathcal{A}(\Omega'')$ are spatially isomorphic to the Gelfand transform of the common trace algebra:

$$\wedge' [\mathrm{tr}' \mathcal{A}(\Omega')] = \hat{\mathcal{A}}'(\Gamma) = \hat{\mathcal{A}}''(\Gamma) = \wedge'' [\mathrm{tr}'' \mathcal{A}(\Omega'')];$$

for the corresponding space bijections one has

$$\beta'(\Omega') = \mathrm{sp}\, \mathcal{A}'(\Gamma) = \mathrm{sp}\, \mathcal{A}''(\Gamma) = \beta''(\Omega''). \tag{3.1}$$

Therefore the algebras $\mathcal{A}(\Omega')$ and $\mathcal{A}(\Omega'')$ are spatially isomorphic (through the map $(\mathrm{tr}'')^{-1}\,(\wedge'')^{-1}\,\wedge'\,\mathrm{tr}'$), whereas the manifolds Ω' and Ω'' are connected through the bijection $\beta : \Omega' \mapsto \Omega''$, $\beta = {\beta''}^{-1}\beta'$;

(3) since the differentiable structures on Ω' and Ω'' are determined by the algebras $\mathcal{A}(\Omega')$ and $\mathcal{A}(\Omega'')$, the bijection β is a diffeomorphism. Due to (3.1) one has $\beta'(\gamma_0) = I'_{\gamma_0} = I''_{\gamma_0} = \beta''(\gamma_0)$, $\gamma_0 \in \Gamma$, which follows to $\beta(\gamma_0) = {\beta''}^{-1}(\beta'(\gamma_0)) = \gamma_0$, i.e., $\beta|_\Gamma = \mathrm{id}$;

(4) introduce on Ω' the (induced) metric $\tilde{g}''$ providing β to be isometry from $(\Omega', \tilde{g}'')$ onto (Ω'', g''). Since $\mathcal{A}(\Omega')$ and $\mathcal{A}(\Omega'')$ are isomorphic, the metric $\tilde{g}''$ (as well as g') turns out to be conforming to the structure determined by $\mathcal{A}(\Omega')$. Due to the last, one has $\tilde{g}'' = \rho g'$ with a $\rho \in C^\infty(\Omega')$. By $\beta|_\Gamma = \mathrm{id}$ one has $\tilde{g}''|_\Gamma = g''|_\Gamma$; taking into account $g'|_\Gamma = g''|_\Gamma = g|_\Gamma$ we obtain $\tilde{g}''|_\Gamma = g'|_\Gamma$, implying $\rho|_\Gamma = 1$.

So, β is an isometry from $(\Omega', \rho g')$ onto (Ω'', g''); i.e., (Ω', g') and (Ω'', g'') belong to one and the same conformal class. The theorem is proved.

Comments and acknowledgments. The first variant of this paper [3], published as a preprint in June 2002, doesn't contain references to the results of Lassas and Uhlmann: unfortunately, we hadn't known of their work [7]. We would like to apologize for this confusion and to thank G. Alessandrini and V. Isakov for informing us about this paper.

To the best of our knowledge, the relation $\mathrm{sp}\, \mathcal{A}(\Omega) \asymp \Omega$ was first established by J. Wermer [11] in the case of the Riemann surface with analytic Γ homeomorphic to the circle. If $\Omega \subset \mathbf{C}$, the algebra $\mathcal{A}(\Omega)$ is generic in a much more general situation (without assumptions on smoothness and connectedness of Γ, see, e.g., [6]). Perhaps, the same is valid for Riemann surfaces, so that, most probably, Lemma 2 is well known for specialists in function algebras. However, we didn't succeed in finding exact references.

Some relations between topology of Ω and the number of generators of the algebra $\mathcal{A}(\Omega)$ (and, hence, of $\mathcal{A}(\Gamma)$) are mentioned in [3]. This is also a way to extract topological properties from DN-map.

If $\Omega \subset \mathbf{C}$ is the unit disk, the operator $\Lambda_g J$ in fact coincides with the Hilbert transform on the unit circle. Perhaps this operator has an abstract analogue in a class of the uniform algebras with nontrivial Shilov boundary.

This paper was written during my stay at the University of Nantes in May 2002. I'm grateful to this university for the kind invitation and the support of my visit.

I'm much obliged to R. Novikov for hospitality, friendly help, and fruitful discussions. I'm deeply indebted to G. M. Henkin for valuable discussions; his consultations helped me to avoid a lot of mistakes. I would like to thank S. Belisheva for help in preparing the manuscript.

REFERENCES

[1] M.I. Belishev, *Boundary control in reconstruction of manifolds and metrics (the BC-method)*, Inverse Problems, 13 (1997), pp. R1–R45.

[2] M.I. Belishev, *Dynamical systems with boundary control: Models and characterization of inverse data*, Inverse Problems, 17 (2001), pp. 659–682.
[3] M.I. Belishev, *The Calderon Problem for Two-Dimensional Manifolds by the BC-Method*, PDMI Preprint 10/2002; available online from www.pdmi.ras.ru/preprint/2002/02-10.html.
[4] E. Bishop, *Subalgebras of functions on a Riemann surface*, Pacific J. Math., 8 (1958), pp. 29–50.
[5] O. Forster, *Lectures on Riemann Surfaces*, Springer-Verlag, New York, Heidelberg, Berlin, 1981.
[6] T.W. Gamelin, *Uniform Algebras*, Prentice-Hall, Englewood Cliffs, NJ, 1969.
[7] M. Lassas and G. Uhlmann, *On determining a Riemann manifold from the Dirichlet-to-Neumann map*, Ann. Sci. École Norm. Sup. (4), 34 (2001), pp. 771–787.
[8] J. Lee and G. Uhlmann, *Determining anisotropic real-analytic conductivities by boundary measurements*, Comm. Pure Appl. Math., 42 (1989), pp. 1097–1112.
[9] M.A. Naimark, *Normed Rings*, WN Publishing, Groningen, The Netherlands, 1970.
[10] G. Schwarz, *Hodge Decomposition—A Method for Solving Boundary Value Problems*, Lecture Notes in Math. 1607, Springer-Verlag, Berlin, 1995.
[11] J. Wermer, *The maximum principle for bounded functions*, Ann. of Math. (2), 69 (1959), pp. 598–604.

SIAM J. MATH. ANAL.
Vol. 35, No. 1, pp. 183–210

SOLVING TIME-HARMONIC SCATTERING PROBLEMS BASED ON THE POLE CONDITION I: THEORY*

THORSTEN HOHAGE†, FRANK SCHMIDT‡, AND LIN ZSCHIEDRICH‡

Abstract. The pole condition is a general concept for the theoretical analysis and the numerical solution of a variety of wave propagation problems. It says that the Laplace transform of the physical solution in radial direction has no poles in the lower complex half-plane. In the present paper we show that for the Helmholtz equation with a radially symmetric potential the pole condition is equivalent to Sommerfeld's radiation condition. Moreover, a new representation formula based on the pole condition is derived and used to prove existence, uniqueness, and asymptotic properties of solutions. This makes it possible to compute the far field of the solution without a Green function, which may not be known explicitly.

Key words. transparent boundary conditions, Laplace transform, Sommerfeld radiation condition

AMS subject classifications. 65N99, 35C10, 35C15, 35C20

DOI. 10.1137/S0036141002406473

1. Introduction.

Differential equations of Helmholtz type describe a large variety of time-harmonic wave propagation problems in acoustics, electromagnetics, and quantum mechanics. To formulate such problems properly on unbounded domains, a so-called radiation condition has to be imposed at infinity. Physically this condition implies that asymptotically no energy is transported towards the origin. The standard condition for bounded obstacles is *Sommerfeld's radiation condition* (cf. [19]). For scattering at rough surfaces or for inhomogeneous exterior domains containing wave guides, Sommerfeld's radiation condition is not valid, and it is often not obvious how to formulate an appropriate radiation condition. In some simple cases a radiation condition can be formulated by a *series representation* of the solution. For general rough surface scattering problems the so-called *upward propagating radiation condition* based on an integral representation of the solution is commonly used (cf. [22]). In this paper we propose a new radiation condition called the pole condition, which seems to be valid for a wide range of problems. Roughly speaking, a function satisfies the *pole condition* if its Laplace transform in the propagation direction has a holomorphic extension to the lower part of the complex plane. Here we show that for bounded obstacles and radially symmetric potentials the pole condition is equivalent to Sommerfeld's radiation condition.

The aim of this paper is not only to give a new proof of existence, uniqueness, and asymptotic properties of solutions to scattering problems based on the pole condition but also to lay the foundations of a new class of efficient numerical algorithm. We derive a set of equations, which can be solved numerically. A detailed discussion of numerical algorithms based on the pole condition will be published elsewhere [10].

*Received by the editors April 26, 2002; accepted for publication (in revised form) October 31, 2002; published electronically July 8, 2003.
http://www.siam.org/journals/sima/35-1/40647.html
†Institut für Numerische und Angewandte Mathematik, Universität Göttingen, Lotzestr. 16–18, D-37083 Göttingen, Germany (hohage@math.uni-goettingen.de). This author was supported by DFG grant DE293/7-1.
‡Zuse Institut Berlin (ZIB), Takustr. 7, D-14195 Berlin, Germany (frank.schmidt@zib.de, zschiedrich@zib.de). The third author was supported by DFG grant SCHM 1386/1-1.

For a preliminary report on numerical results we refer to [11]. For an analysis of the PML method based on the pole condition we refer to [9].

In many applications, one is interested in the behavior of the solution far away from the obstacle. Using a representation formula derived in this paper the solution can be evaluated at any given point in a cheap and stable manner. For problems for which neither a fundamental solution nor a series representation of the solution is known explicitly, this distinguishes our method from the standard methods (cf. [1, 2, 3, 7, 8, 12]).

The potential range of applications of our method is not restricted to the class of problems considered in this paper. A basic requirement is that the Laplace transform of the differential equation along a family of rays connecting the artificial boundary to infinity can be carried out analytically. The original motivation of this work was problems in integrated optics involving waveguides. For a numerical solution of such problems by methods based on the pole condition we refer to [11, 16].

The pole condition was first considered by the second author for problems with one space-like and one time-like variable. In [15, 17] the time-discretized Schrödinger equation is interpreted as a sequence of inhomogeneous Helmholtz problems. One-way wide-angle Helmholtz equations, ranging between the Schrödinger and the Helmholtz equations, were studied in [18]. With the results below, we hope to be able to carry over the analysis of these papers to time-dependent problems in arbitrary space dimensions.

2. Main results and outline of the paper. We consider partial differential equations of the form

$$\Delta u + (1 + p(|x|))\,\kappa^2 u = 0 \tag{2.1}$$

with a real-valued functions p in some exterior domain $\Omega \supset \{x \in \mathbb{R}^d : |x| > a_*\}$, $a_* > 0$. p is assumed to be analytic of the form $p(r) = \sum_{j=2}^{\infty} p_j r^{-j}$ and describes either a radially symmetric potential or a variation of the refractive index.

Let us motivate the pole condition by the simplest case $d = 1$, $p = 0$, and $a_* = 1$. Here (2.1) reduces to the ordinary differential equation $u'' + \kappa^2 u = 0$ with the general solution

$$u(1 + r) = C_1 e^{i\kappa r} + C_2 e^{-i\kappa r}, \qquad r > 0.$$

The term $C_1 e^{i\kappa r}$ corresponds to an outgoing wave, and $C_2 e^{-i\kappa r}$ to an incoming wave. The Laplace transform $\hat{u}_1(s) := \int_0^\infty e^{-sr} u(1 + r)\,\mathrm{d}r$ of $u(1 + \cdot)$, $\operatorname{Re} s > 0$, is given by

$$\hat{u}_1(s) = \frac{C_1}{s - i\kappa} + \frac{C_2}{s + i\kappa}.$$

This function, which has a holomorphic extension to $\mathbb{C} \setminus \{i\kappa, -i\kappa\}$, satisfies $\operatorname{res}_{i\kappa} \hat{u}_1 = C_1$ and $\operatorname{res}_{-i\kappa} \hat{u}_1 = C_2$. u is outgoing if and only if $\hat{u}_1$ has no pole at $-i\kappa$, i.e., if and only if $\operatorname{res}_{-i\kappa} \hat{u}_1 = 0$.

To avoid notational difficulties we assume that $d \geq 2$ from now on. Let us introduce the function

$$U(\rho, \hat{x}) := \rho^{\frac{d-1}{2}} u(\rho \hat{x}) \tag{2.2}$$

for $\rho > a_*$, $\hat{x} \in S^{d-1} := \{x \in \mathbb{R}^d : |x| = 1\}$, and its (shifted) Laplace transform

$$\hat{U}_a(s, \hat{x}) := \int_0^\infty e^{-sr} U(r + a, \hat{x})\,\mathrm{d}r \tag{2.3}$$

for $\operatorname{Re} s > 0$, $\hat{x} \in S^{d-1}$, and $a \geq a_*$. Note that the scaling factor in the definition of U is chosen such that $\|U(\rho, \cdot)\|_{L^2(S^{d-1})} = \|u\|_{L^2(\rho S^{d-1})}$ for all $\rho > a_*$.

DEFINITION 2.1 (pole condition). *A bounded function $u : \{x \in \mathbb{R}^d : |x| > a_*\} \to \mathbb{C}$ satisfies the pole condition if for some $a \geq a_*$ the function $\hat{U}_a(\cdot, \hat{x})$ defined by (2.2) and (2.3) has a holomorphic extension to the lower complex half-plane $\mathbb{C}^- := \{s \in \mathbb{C} : \operatorname{Im} s < 0\}$ for all $\hat{x} \in S^{d-1}$ such that the function $s \mapsto \int_{S^{d-1}} |\frac{\partial \hat{U}_a}{\partial s}(s, \hat{x})|^2 \, \mathrm{d}s(\hat{x})$ is bounded on compact subsets of $\mathbb{C}^-$.*

Remark 2.2. If the pole condition is satisfied for one $a \geq a_*$, it is satisfied for all $a \geq a_*$. This follows from the identity

$$\int_0^\infty e^{-sr} U(a + r, \hat{x}) \, \mathrm{d}r = \int_0^{b-a} e^{-sr} U(a + r, \hat{x}) \, \mathrm{d}r + e^{-s(b-a)} \int_0^\infty e^{-sr} U(b + r, \hat{x}) \, \mathrm{d}r \tag{2.4}$$

and the fact that both $s \mapsto \int_0^{b-a} e^{-sr} U(a + r, \hat{x}) \, \mathrm{d}r$ and $s \mapsto e^{-s(b-a)}$ are entire functions. Hence, the pole condition is a condition concerning the behavior of u at infinity but not the behavior of u on any compact set.

A similar condition without the scaling (2.2) was considered in [14]. We will show that a solution u to the differential equation (2.1) satisfies the pole condition if and only if it satisfies the Sommerfeld radiation condition

$$\lim_{\rho \to \infty} \rho^{\frac{d-1}{2}} \left(\frac{\partial u}{\partial \rho} - i\kappa u \right) = 0, \qquad \rho = |x|, \tag{2.5}$$

uniformly for all directions $\frac{x}{|x|}$.

The structure of the singularity is more complicated in general than in the simple example above. If a solution to (2.1) satisfies the pole condition, then $\hat{U}(\cdot, \hat{x})$ has an analytic extension not only to $\mathbb{C}_-$ but even to $\mathbb{C} \backslash \{i\kappa - t : t \geq 0\}$; i.e., we do not have an isolated singularity, but a singularity with a branch cut. For a holomorphic function $f : V \to \mathbb{C}$ defined on a domain $V \subset \mathbb{C}$ and $\sigma \in \overline{V}$ we define $\operatorname{Res}_\sigma f := \lim_{s \to \sigma, s \in V} (s - \sigma) f(s)$ if this limit exists. If f has an isolated pole of order 1 at σ, then $\operatorname{Res}_\sigma f = \operatorname{res}_\sigma f$. The functions

$$u_\infty(\hat{x}) = e^{-i\kappa a} \operatorname{Res}_{i\kappa} \hat{U}_a(\cdot, \hat{x}), \tag{2.6a}$$

$$\Psi_a(t, \hat{x}) = \frac{e^{-i\kappa a}}{2\pi i} \lim_{\epsilon \to 0} \left(\hat{U}_a(i\kappa - t - i\epsilon, \hat{x}) - \hat{U}_a(i\kappa - t + i\epsilon) \right) \tag{2.6b}$$

are well defined for $\hat{x} \in S^{d-1}$, $t > 0$, and a sufficiently large; i.e., u_∞ is independent of a and the limit in (2.6b) exists. It is a crucial result of our analysis that these functions determine the solution U completely via the representation formula

$$U(a + r, \hat{x}) = e^{i\kappa(a+r)} \left(u_\infty(\hat{x}) + \int_0^\infty e^{-tr} \Psi_a(t, \hat{x}) \, \mathrm{d}t \right), \qquad r \geq 0. \tag{2.7}$$

Let $\Delta_{S^{d-1}}$ denote the Laplace–Beltrami operator on S^{d-1} and define

$$\tilde{\Delta}_{\hat{x}} \varphi := \Delta_{S^{d-1}} \varphi + \left(\frac{(d-1)(3-d)}{4} \right) \varphi \tag{2.8}$$

for $\varphi \in C^2(S^{d-1})$. Then u_∞ and Ψ_a satisfy the Volterra integrodifferential equation

$$\left\{\check{p}_a(t) + te^{-at}\tilde{\Delta}_{\hat{x}}\right\} u_\infty(\hat{x}) + t(t-2i\kappa)\Psi_a(t,\hat{x}) + \int_0^t \left\{\check{p}_a(t-t_1) + (t-t_1)e^{-a(t-t_1)}\tilde{\Delta}_{\hat{x}}\right\} \Psi_a(t_1,\hat{x})\,\mathrm{d}t_1 = 0. \tag{2.9}$$

Here $\check{p}_a$ is the inverse Laplace transform of $p_a := p(a+\cdot)$ (cf. Lemma 4.1). If $p = 0$, then (2.9) can be converted to a differential equation by multiplying by e^{at} and differentiating twice with respect to t.

Given boundary data $f(\hat{x}) = U(a,\hat{x})$, (2.7) implies

$$u_\infty(\hat{x}) + \int_0^\infty \Psi_a(t,\hat{x})\,\mathrm{d}t = e^{-i\kappa a} f(\hat{x}). \tag{2.10}$$

We show that the system (2.9), (2.10) has a unique solution (u_∞, Ψ_a) if the radius a of the artificial boundary $\Gamma_a := \{x : |x| = a\}$ is chosen sufficiently large. Other boundary conditions can easily be taken care of by differentiating (2.7). It suffices to compute $\Psi_a(t,\hat{x})$ on a small interval $t \in [0,T]$ since $\Psi_a(t,\hat{x})$ decays exponentially as $t \to \infty$. Once u_∞ and Ψ_a are known, $U(\rho,\hat{x})$ can be evaluated for $\rho \geq a$ using (2.7).

The plan of this paper is as follows: In section 3 we introduce the Dirichlet-to-Neumann map on Γ_a and prove an existence and uniqueness theorem based on properties of this operator. In section 4 we derive an ordinary differential equation for the Fourier coefficients of $U(r,\cdot)$ and a corresponding Volterra integral equation for the Laplace transform of these functions. Existence and uniqueness of solutions to these integral equations is established in section 5. In the following section the main results of this paper are proved for single Fourier modes. As a simple consequence of a representation formula corresponding to (2.7) we derive asymptotic formulas for (generalized) Hankel functions for large arguments. In section 8 we construct the Dirichlet-to-Neumann map using Fourier series and show that it satisfies the assumptions of the existence and uniqueness theorem in section 3. Then, in section 9, we establish the formulas (2.7) and (2.9) and show the equivalence of the pole condition and Sommerfeld's radiation condition.

3. The Dirichlet-to-Neumann map on the artificial boundary. For simplicity, we assume that Ω is the complement of some sufficiently smooth compact set K contained in $\{x : |x| < a_*\}$ such that $p(|x|)$ is well defined and finite for $x \in \Omega$. Moreover, we assume that u satisfies the Neumann boundary conditions $\frac{\partial u}{\partial \nu} = f$ on the boundary ∂K. We could easily accommodate for more complicated situations, e.g., different boundary conditions or inhomogeneities in the interior of Γ_a.

From now on we assume without loss of generality (w.l.o.g.) that $\kappa = 1$. Since a and p are arbitrary, this can be achieved by the following rescaling: $\tilde{x} = \kappa x$, $\tilde{\rho} = \kappa\rho$, $\tilde{a} = \kappa a$, $\tilde{t} = \kappa^{-1}t$, $\tilde{s} = \kappa^{-1}s$, $\tilde{p}(\tilde{\rho}) = p(\rho)$, $\tilde{\check{p}}(\tilde{t}) = \kappa\check{p}(t)$, $\tilde{u}(\tilde{x}) = u(x)$, $\tilde{U}(\tilde{\rho},\hat{x}) = \kappa^{\frac{d-1}{2}} U(\rho,\hat{x})$, $\tilde{\hat{U}}(\tilde{s},\hat{x}) = \kappa^{\frac{d+1}{2}}\hat{U}(s,\hat{x})$, $\tilde{u}_\infty(\hat{x}) = \kappa^{\frac{d-1}{2}} u_\infty(\hat{x})$, $\tilde{\Psi}(\tilde{t},\hat{x}) = \kappa^{\frac{d+1}{2}}\Psi(t,\hat{x})$. For notational convenience we will drop the tildes in the following.

To arrive at a weak formulation, we multiply (2.1) by a function $-\overline{v}$ and integrate over $\Omega_a := \{x \in \Omega : |x| < a\}$. Formally applying Green's theorem yields

$$\int_{\Omega_a} \left(\nabla u \nabla \overline{v} - (1 + p(|x|))\, u\overline{v}\right)\,\mathrm{d}x - \int_{\Gamma_a \cup \partial K} \frac{\partial u}{\partial \nu}\overline{v}\,\mathrm{d}s = 0,$$

where the unit normal vector ν points to the exterior of Ω_a. Now we introduce a so-called *Dirichlet-to-Neumann map* $\mathrm{DtN} : H^{1/2}(\Gamma_a) \to H^{-1/2}(\Gamma_a)$ which maps the

Dirichlet data $u|_{\Gamma_a}$ of a solution u satisfying (2.1) and (2.5) to its Neumann data $\frac{\partial u}{\partial \nu}|_{\Gamma_a}$. Existence and uniqueness of such solutions in $\{x : |x| > a\}$ will be proved later. With the sesquilinear form $a : H^1(\Omega_a) \times H^1(\Omega_a) \to \mathbb{C}$,

$$a(u,v) := \int_{\Omega_a} \left(\nabla u \nabla \overline{v} - (1 + p(|x|))\, u\overline{v}\right) \mathrm{d}x - \int_{\Gamma_a} \mathrm{DtN} u \overline{v}\, \mathrm{d}s,$$

and the continuous antilinear functional $F : H^1(\Omega_a) \to \mathbb{C}$,

$$F(v) := \int_{\partial K} f \overline{v}\, \mathrm{d}s,$$

the variational problem reads

$$a(u,v) = F(v) \qquad \text{for all } v \in H^1(\Omega_a). \tag{3.1}$$

PROPOSITION 3.1. *Let* DtN *be an operator with the following properties:*

1. $\mathrm{DtN} : H^{1/2}(\Gamma_a) \to H^{-1/2}(\Gamma_a)$ *is linear and bounded.*
2. *There exists a compact operator* $L : H^{1/2}(\Gamma_a) \to H^{-1/2}(\Gamma_a)$ *such that the inequality* $\mathrm{Re} \int_{\Gamma_a} (-\mathrm{DtN} + L)\varphi\, \overline{\varphi}\, \mathrm{d}s \geq 0$ *holds for all* $\varphi \in H^{1/2}(\Gamma_a)$.
3. $\mathrm{Im} \int_{\Gamma_a} \mathrm{DtN}\varphi\, \overline{\varphi}\, \mathrm{d}s > 0$ *for all* $\varphi \in H^{1/2}(\Gamma_a)$, $\varphi \neq 0$.

Then the variational problem (3.1) *has a unique solution* u *for all right-hand sides* F, *and* u *depends continuously on* F.

Proof. As the proof is rather standard (cf., e.g., [2, Thm. 5.7] for a similar proof), we give only a brief sketch. Condition 1 ensures that the sesquilinear form a is well defined. Condition 2 is used to establish the Gårding inequality

$$\mathrm{Re}\, a(u,u) + c_2 \|u\|^2_{L^2(\Omega_a)} + \mathrm{Re}\, \langle L \,\mathrm{Tr}\, u, \mathrm{Tr}\, u\rangle_{L^2(\Gamma_a)} \geq c_1 \|u\|^2_{H^1(\Omega_a)}$$

for all $u \in H^1(\Omega_a)$ with constants $c_1, c_2 > 0$ and the trace operator $\mathrm{Tr} : H^1(\Omega_a) \to H^{1/2}(\Gamma_a)$. Since the embedding operator $H^1(\Omega_a) \hookrightarrow (H^1(\Omega_a))'$ and the operator $\mathrm{Tr}'\, L\, \mathrm{Tr} : H^1(\Omega_a) \to H^1(\Omega_1)$ are compact, it can be shown by the Lax–Milgram lemma and Riesz theory that the operator induced by the sesquilinear form a is Fredholm with index 0; i.e., uniqueness implies existence and stability. Let $u \in H^1(\Omega_a)$ satisfy $a(u,v) = 0$ for all $v \in H^1(\Omega_a)$. Taking the imaginary part of this equation and using condition 3, it follows that u has vanishing Cauchy data on Γ_a. Hence, by virtue of the Cauchy–Kovalevskaya theorem and elliptic regularity results, u must vanish everywhere. ∎

Usually the properties of DtN required in the previous proposition are proved using special properties of the Hankel functions (cf. [2, 12]). In the following we present a more systematic approach which also works for $p \neq 0$.

4. The Laplace transform of the separated differential equation. Let $\{(\varphi_j, \lambda_j) : j \in \mathbb{N}\}$ be a complete orthonormal system of eigenfunctions and eigenvalues of the operator $\tilde{\Delta}_{\hat{x}}$ defined in (2.8). φ_j may be chosen as trigonometric monomials for $d = 2$ and spherical harmonics for $d = 3$. Let $U_j(r) := \int_{S^1} U(r,\cdot)\, \overline{\varphi_j}\, \mathrm{d}s$ denote the Fourier coefficients of U. Using the formula

$$\Delta = \frac{\partial^2}{\partial \rho^2} + \frac{d-1}{\rho}\frac{\partial}{\partial \rho} + \frac{1}{\rho^2}\Delta_{S^{d-1}}$$

it follows after some simple computations that the Fourier coefficients $U_j(r)$ satisfy the differential equations

$$U_j''(\rho) + \left(1 + p(\rho) + \lambda_j \rho^{-2}\right) U_j(\rho) = 0. \tag{4.1}$$

Note that there is no term involving U_j' due to the scaling factor $\rho^{(d-1)/2}$.

Let $(\mathcal{L}f)(s) := \int_0^\infty e^{-s\rho} f(\rho)\,\mathrm{d}\rho$, $\operatorname{Re} s > 0$, denote the Laplace transform of a function f. In order to derive an equation for $\hat{U}_{j,a} := \mathcal{L}U_j(\cdot + a)$ we need the following lemma.

LEMMA 4.1. *Assume that the convergence radius of* $p(t^{-1}) = \sum_{m=1}^\infty p_m t^m$ *is greater than* $\frac{1}{a_p}$, $a_p \in (0,\infty)$, *and let* $a > a_p$. *Let*

$$\check{p}_a(s) := e^{-as} \sum_{m=1}^\infty \frac{p_m}{(m-1)!} s^{m-1} \tag{4.2}$$

be the inverse Laplace transform of $p(\cdot + a)$ *(i.e.,* $(\mathcal{L}\check{p}_a)(r) = p(r + a)$*), and let* $u \in C([0,\infty))$ *be a bounded function. Then*

$$\mathcal{L}\left(p(\cdot + a)u\right)(s) = \int_s^\infty \check{p}_a(s_1 - s)(\mathcal{L}u)(s_1)\,\mathrm{d}s_1 \tag{4.3}$$

for $\operatorname{Re} s > 0$. *Here* $\int_s^\infty f(s_1)\,\mathrm{d}s_1 := \int_0^\infty f(s + t)\,\mathrm{d}t$. *For all* $k = 0, 1, \ldots$ *there exists a constant* $C > 0$ *such that*

$$|\check{p}_a^{(k)}(s)| \le Ce^{-a\operatorname{Re} s + a_p|s|} \tag{4.4a}$$

for all $s \in \mathbb{C}$. *If* $p_1 = 0$, *then also*

$$|\check{p}_a(s)| \le C|s|e^{-a\operatorname{Re} s + a_p|s|}. \tag{4.4b}$$

Proof. We first prove by induction in $m \in \mathbb{N}$ that (4.3) is true for $p(t^{-1}) = t^m$. To show this for $m = 1$ we consider the function $f(t) := (t + a)^{-1}u(t)$. A simple computation shows that $\lim_{s\to\infty} \mathcal{L}f(s) = 0$ and

$$(\mathcal{L}u)(s) = \left(\mathcal{L}\left((\cdot + a)f\right)\right)(s) = a(\mathcal{L}f)(s) - (\mathcal{L}f)'(s).$$

On the other hand, the right-hand side of (4.3) with $\check{p}_a(s) = e^{-as}$ is the unique solution of this differential equation vanishing at ∞. Now assume that (4.3) holds true for $p(t^{-1}) = t^m$ with $m \le j$, $j \in \mathbb{N}$. Then

$$\begin{aligned}
\mathcal{L}\left(\frac{u}{(\cdot + a)^{j+1}}\right)(s) &= \int_s^\infty \frac{e^{a(s-s_1)}(s_1 - s)^{j-1}}{(j-1)!} \left(\mathcal{L}\left(\frac{u}{\cdot + a}\right)\right)(s_1)\,\mathrm{d}s_1 \\
&= \int_s^\infty \frac{e^{a(s-s_1)}(s_1 - s)^{j-1}}{(j-1)!} \int_{s_1}^\infty e^{a(s_1-s_2)}(\mathcal{L}u)(s_2)\,\mathrm{d}s_2\,\mathrm{d}s_1 \\
&= \int_s^\infty e^{a(s-s_2)}(\mathcal{L}u)(s_2) \int_s^{s_2} \frac{(s_1 - s)^{j-1}}{(j-1)!}\,\mathrm{d}s_1\,\mathrm{d}s_2 \\
&= \int_s^\infty \frac{e^{a(s-s_2)}(s_2 - s)^j}{j!}(\mathcal{L}u)(s_2)\,\mathrm{d}s_2.
\end{aligned}$$

So far we have proved (4.3) if $t \mapsto p(t^{-1})$ is a polynomial. It remains to consider the case that p is given by an infinite series. It follows from the definition of a_p that $C := \sup_{m\ge 0} |p_{m+1}|a_p^{-m} < \infty$. Hence,

$$\begin{aligned}
|\check{p}_a(s)| &\le e^{-a\operatorname{Re} s} \sum_{m=0}^\infty \frac{|p_{m+1}||s|^m}{m!} \le Ce^{-a\operatorname{Re} s} \sum_{m=0}^\infty \frac{a_p^m|s|^m}{m!} \\
&\le Ce^{-a\operatorname{Re} s + a_p|s|}.
\end{aligned}$$

The other estimates in (4.4) are derived in a similar manner. Since all partial sums in the definition of $\check{p}_a$ are bounded by the right-hand side of the previous inequality and since the series $p(t)$ converges uniformly for $|t| \geq a$, it can be shown by Lebesgue's dominated convergence theorem that

$$\begin{aligned}\int_0^\infty e^{-sr} p(r+a)u(r)\,\mathrm{d}r &= \lim_{M\to\infty} \int_0^\infty e^{-sr} \sum_{m=1}^{M} \frac{p_m}{(r+a)^m} u(r)\,\mathrm{d}r \\ &= \lim_{M\to\infty} \int_s^\infty e^{a(s-s_1)} \sum_{m=1}^{M} \frac{p_m (s-s_1)^{m-1}}{(m-1)!} (\mathcal{L}u)(s_1)\,\mathrm{d}s_1 \\ &= \int_s^\infty \check{p}_a(s_1-s)(\mathcal{L}u)(s_1)\,\mathrm{d}s_1. \qquad \square\end{aligned}$$

It follows from (4.1) and Lemma 4.1 that

$$\begin{aligned}&\int_s^\infty \left(\check{p}_a(s_1-s) + e^{-a(s_1-s)}(s_1-s)\lambda_j\right) \hat{U}_{j,a}(s_1)\,\mathrm{d}s_1 \\ &\qquad + (s^2+1)\hat{U}_{j,a}(s) = s\,U_j(a) + U_j'(a), \qquad \operatorname{Re} s > 0.\end{aligned} \tag{4.5}$$

5. The integral equation in the Laplace domain. In this and the following two sections we consider differential equations of the form

$$U''(a+r) + (1 + (\mathcal{L}P_a)(r))\,U(a+r) = 0, \qquad r > 0, \tag{5.1}$$

with an analytic function P_a of the form (4.2) with $p_1 = 0$ which satisfies the estimates (4.4) with $a > a_p$. Equations (4.1) are of this form. The dependence of the solution on λ_j will be discussed later. For studying the integral equation in the Laplace domain, it is useful to factor out the singularities of $\hat{U}_a$ at $\pm i$, i.e., to look at the function

$$w_a(s) = \hat{U}_a(s)(s^2+1). \tag{5.2}$$

In the following we often omit the index a in $\hat{U}_a$, w_a, and P_a. Due to Lemma 4.1 the function w satisfies the Volterra integral equation

$$w(s) + Jw(s) = sU(a) + U'(a) \tag{5.3}$$

with

$$(Jw)(s) := \int_s^\infty P(s_1-s)\frac{w(s_1)}{s_1^2+1}\,\mathrm{d}s_1. \tag{5.4}$$

Let us introduce the cuts $S_{\pm i} := \{\pm i + t : t < 0\}$ and $V := \mathbb{C}\backslash(S_i \cup S_{-i})$ (cf. Figure 5.1(a)). We define the metric on V by $d(s_1,s_2) := |s_1-s_2| + |\varphi(s_1)-\varphi(s_2)|$ with the function $\varphi : V \to \mathbb{R}$ given by $\varphi(s) := -\operatorname{Re} s$ if $\operatorname{Re} s \leq 0$ and $|\operatorname{Im} s| < 1$, $\varphi(s) := 0$ else. This metric is defined such that points on opposite sides of the cuts are far away from each other. Let $(\overline{V}, \overline{d})$ denote the completion of the metric space (V,d). Then $\overline{V}$ is the union of V and the set of points $s_\pm := \lim_{\epsilon\to 0,\epsilon>0} s \pm i\epsilon$ with $s \in S_i \cup S_{-i}$. For a continuous function $v : \overline{V} \to \mathbb{C}$ we can define a "jump function" $[v] : S_i \cup S_{-i} \to \mathbb{C}$ by

$$[v](s) := v(s_-) - v(s_+), \qquad s \in S_i \cup S_{-i}. \tag{5.5}$$

Note that $[v]$ is continuous on $S_i \cup S_{-i}$ with respect to the topology induced by the usual norm of $\mathbb{C}$.

We introduce the norm

$$\|w\|_X := \sup_{s\in\overline{V}} \frac{|w(s)|}{|s|^2+1}$$

and denote by X the space of all $w \in C(\overline{V})$ which are holomorphic in V and satisfy $w(s) = \mathrm{o}\big(|s|^2\big)$ uniformly for $|s| \to \infty$. X is equipped with the norm $\|\cdot\|_X$.

In the following we use the notation $|s|_1 := |\operatorname{Re} s| + |\operatorname{Im} s|$ for $s \in \mathbb{C}$. Moreover, we define the diamond-shaped domains $D_\pm := \{s \in \overline{V} : |s \mp i|_1 < \frac{1}{2}\}$.

LEMMA 5.1. *Let $0 < \alpha < 1$. Then there exists a constant c such that for all $w \in X$*

$$|(Jw)(s)| \le c \left(\sup_{\operatorname{Re} s_1 \ge \operatorname{Re} s} \frac{|w(s_1)|}{|s_1|^2+1} \right), \qquad s \in \overline{V}\backslash(D_+ \cup D_-), \tag{5.6a}$$

$$|(Jw)'(s)| \le c \left(\sup_{\operatorname{Re} s_1 \ge \operatorname{Re} s} \frac{|w(s_1)|}{|s_1|^2+1} \right), \qquad s \in \overline{V}\backslash(D_+ \cup D_-), \tag{5.6b}$$

$$|(Jw)(s) - (Jw)(\sigma)| \le c\|w\|_X\, d(s,\sigma)^\alpha, \qquad s, \sigma \in D_\pm, \tag{5.6c}$$

$$|(Jw)(s)| \le c\|w\|_X, \qquad s \in \overline{V}. \tag{5.6d}$$

Proof. Part (a). Since the operator J is defined by an integral over a holomorphic function (cf. (5.4)), we may deform the integration path in order to facilitate the proof. We choose the path $\gamma_s(t) := s + t$, $t \ge 0$, if it does not intersect with $D_+ \cup D_-$. Otherwise we set

$$\gamma_s(t) := s + t + i\psi_s(t), \qquad t \ge 0, \tag{5.7}$$

with a real-valued function ψ_s as shown in Figure 5.1(a). ψ_s is chosen such that γ_s does not intersect $D_+ \cup D_- \cup S_i \cup S_{-i}$, $\psi_s(0) = 0$, $|\psi_s| \le \frac{1}{2}$, $|\psi_s'| \le 1$, $\operatorname{meas}(\operatorname{supp}\psi_s) \le 1$, and $\lim_{t\to\infty} \psi(t) = 0$. We have

$$(Jw)(s) = \int_0^\infty P(t + i\psi_s(t)) \frac{|\gamma_s(t)|^2+1}{\gamma_s(t)^2+1} \frac{w(\gamma_s(t))}{|\gamma_s(t)|^2+1} \gamma_s'(t)\,\mathrm{d}t. \tag{5.8}$$

Due to (4.4a) there exists a constant $c > 0$ such that for all $t \ge 0$

$$\sup_{|\tau|\le \min(t,1/2)} |P(t+i\tau)|\,\mathrm{d}t \le ce^{-(a-a_p)t/2}. \tag{5.9}$$

To see this, choose T such that $-at + a_p|t + i/2| \le -(a-a_p)t/2$ for $t \ge T$. Then (5.9) holds true for $t \ge T$. By a compactness argument it is also true that $0 \le t \le T$. Here and in the following, c is a generic constant. Moreover, $\sup_{s\in\mathbb{C}\backslash(D_+\cup D_-)} \frac{|s|^2+1}{|s^2+1|} < \infty$. Hence,

$$\begin{aligned} |(Jw)(s)| &\le c \left(\int_0^\infty e^{-(a-a_p)t/2}\,\mathrm{d}t \right) \left(\sup_{\operatorname{Re} s_1 \ge \operatorname{Re} s} \frac{|w(s_1)|}{|s_1|^2+1} \right) \\ &= \frac{2c}{a-a_p} \left(\sup_{\operatorname{Re} s_1 \ge \operatorname{Re} s} \frac{|w(s_1)|}{|s_1|^2+1} \right). \end{aligned}$$

This implies (5.6a).

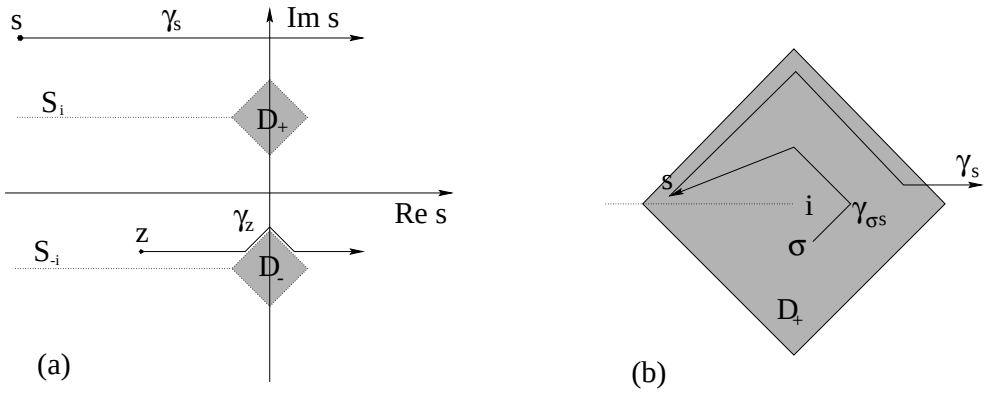


FIG. 5.1. *Integration paths in the proof of Lemma* 5.1.

Part (b). Differentiating (5.4) yields

$$(Jw)'(s) = -\int_s^\infty P'(s_1 - s)\frac{w(s_1)}{s_1^2+1}\,\mathrm{d}s_1 \tag{5.10}$$

since $P(0) = 0$. Now (5.6b) follows as above.

Part (c). W.l.o.g. we may assume that $s, \sigma \in D_+$ and that $|s-i|_1 \geq |\sigma - i|_1$. In this proof we say that s and σ are *on opposite sides of* S_i if $\operatorname{Re} s, \operatorname{Re}\sigma < 0$ and $\operatorname{Im}(s-i)\operatorname{Im}(\sigma-i) < 0$ or if s and σ are the limit of a sequence of such points. We first assume that s and σ are not on opposite sides of S_i. Let $\gamma_{\sigma s}$ be the shortest path from σ to s in D_+ such that $|s_1 - i|_1 \geq |\sigma - i|_1$ for all $s_1 \in \gamma_{\sigma s}$ (cf. Figure 5.1(b)). The length of this path can be estimated by $l(\gamma_{\sigma s}) \leq 3\delta$, where $\delta := |s - \sigma|$. Moreover, we choose the path γ_s from s to ∞ such that $|s_1 - i|_1 \geq |s - i|_1$ for all $s \in \gamma_s$ (cf. Figure 5.1(b)). We have

$$\begin{aligned}(Jw)(\sigma) - (Jw)(s) = &\int_{\gamma_{\sigma s}} \frac{P(s_1-\sigma)}{s_1^2+1} w(s_1)\,\mathrm{d}s_1 \\ &+ \int_{\gamma_s} \left(P(s_1-\sigma) - P(s_1 - s)\right)\frac{w(s_1)}{s_1^2+1}\,\mathrm{d}s_1.\end{aligned}$$

To estimate the integral over $\gamma_{\sigma s}$, we use the inequalities $|s| \leq |s|_1 \leq \sqrt{2}|s|$ and (4.4b):

$$\left|\frac{P(s_1 - s)}{(s_1+i)(s_1-i)}\right| \leq \sqrt{2}\left|\frac{1}{s_1+i}\right| \cdot \left|\frac{P(s-s_1)}{s-s_1}\right| \leq c.$$

Together with the bound on $l(\gamma_{\sigma s})$ this yields $\left|\int_{\gamma_{\sigma s}} \ldots\right| \leq c\delta\|w\|_X$.

The integral over γ_s is estimated by the mean value theorem:

$$\begin{aligned}&\left|\int_{\gamma_s} \left(P(s_1-\sigma) - P(s_1-s)\right)\frac{w(s_1)\,\mathrm{d}s_1}{s_1^2+1}\right| \\ &\leq \delta \int_{\gamma_s} \sup_{\lambda\in[0,1]} |P'(s_1 - \lambda s - (1-\lambda)\sigma)| \frac{|w(s_1)|\,|\,\mathrm{d}s_1|}{|s_1^2+1|}.\end{aligned}$$

To bound the integrand for $s_1 \in D_+ \cap \gamma_s$ we note that

$$\begin{aligned}|s_1 - i|_1 &\geq |s_1 - s|_1 - |s-i|_1 \geq |s_1 - s|_1 - |s_1 - i|_1, \\ 2|s_1 - i|_1 &\geq 2|s-i|_1 \geq |s-i|_1 + |\sigma - i|_1 \geq |s-\sigma|_1 \geq \delta\end{aligned}$$

due to the choice of γ_s. Adding these inequalities yields $4|s_1 - i|_1 \geq |s_1 - s|_1 + \delta$. Together with the estimates $|s_1 - s|_1 \geq \operatorname{Re}(s_1 - s)$ and $|s_1 + i|_1 \geq 1$ we obtain $|s_1^2 + 1| \geq \frac{1}{2}|s_1 + i|_1 |s_1 - i|_1 \geq \frac{1}{8}(\operatorname{Re}(s_1 - s) + \delta)$. Outside of D_+ the bound $\left|\frac{w(s_1)}{s_1^2+1}\right| \leq c\|w\|_X$ holds true. With (4.4a) and $t^* := \sup\{t \geq 0 : s + t \in D_+\}$ we obtain

$$\begin{aligned}\left|\int_{\gamma_s} \dots\right| &\leq c\delta\|w\|_X \left(\int_0^{t^*} \frac{\mathrm{d}t}{t+\delta} + \int_{t^*}^{\infty} e^{-(a-a_p)t/2}\,\mathrm{d}t\right) \\ &\leq c\delta\|w\|_X \left(\ln\delta + \frac{a-a_p}{2}\right) \leq c\delta^\alpha \|w\|_X.\end{aligned}$$

Since $\delta \leq d(s,\sigma)$, we have proved (5.6c) if s and σ are not on opposite sides of S_i. Otherwise, we obtain from our previous estimates that

$$\begin{aligned}|(Jw)(\sigma) - (Jw)(s)| &\leq |(Jw)(\sigma) - (Jw)(i)| + |(Jw)(i) - Jw(s)| \\ &\leq c(|\sigma - i|^\alpha + |i - s|^\alpha).\end{aligned}$$

We may assume w.l.o.g. that $\operatorname{Im}(\sigma - i) \leq 0$ and $\operatorname{Im}(s - i) \geq 0$. Then $d(\sigma, s) = |\operatorname{Re}\sigma| + |\sigma - s|$ and

$$\begin{aligned}&|\operatorname{Im}(\sigma - i)| + |\operatorname{Im}(i - s)| = |\operatorname{Im}(\sigma - s)| \leq d(\sigma, s), \\ &|\operatorname{Re}(s - i)| = |\operatorname{Re} s| \leq |\operatorname{Re}\sigma| + |\operatorname{Re}(s - \sigma)| \leq d(s, \sigma), \\ &|\operatorname{Re}(\sigma - i)| = |\operatorname{Re}\sigma| \leq d(s, \sigma).\end{aligned}$$

Inserting these inequalities in the previous inequality yields (5.6c).

Part (d). Inequality (5.6d) follows easily from (5.6a) and (5.6c). ∎

LEMMA 5.2. *The operator J is compact from X to X.*

Proof. It follows from (5.4) that Jw is holomorphic in V for $w \in X$ and from Lemma 5.1 that Jw is continuous in $\overline{V}$. Together with the estimate (5.6d) this shows that J maps X into X. To prove compactness, let $(w_n)_{n\in\mathbb{N}}$ be a sequence in X with $\|w_n\|_X \leq 1$ for all $n \in \mathbb{N}$. We have to show that the sequence $v_n := Jw_n$, $n \in \mathbb{N}$, has a convergent subsequence in X. Let us first consider the restrictions of v_n on some compact subset $K \subset \overline{V}$. Due to (5.6b) and (5.6c), the sequence $(v_n|_K)_{n\in\mathbb{N}}$ is equicontinuous on K with respect to the metric d. Hence, by the Arzelà–Ascoli theorem, there exists a subsequence of $(v_n)_{n\in\mathbb{N}}$ which converges with respect to the norm $\|\varphi\|_{\infty,K} := \sup_{s\in K} |\varphi(s)|$. In order to construct a subsequence which converges globally, we introduce the sets $K_j := \{s \in \overline{V} : |s| \leq j\}$ for $j \in \mathbb{N}$. It is easy to show that these sets are compact with respect to the metric d. By the argument above, there exists a subsequence $(v_{n_1(l)})_l$ which converges with respect to $\|\cdot\|_{\infty,K_1}$. Applying the same argument again, we get a subsequence $(v_{n_2(l)})_l$ which converges with respect to $\|\cdot\|_{\infty,K_2}$. Repeating this process of selecting subsequences, we arrive at an array $v_{n_j(l)}$ with the property that each row is a subsequence of the previous row. The diagonal subsequence $v_{n(l)} := v_{n_l(l)}$ converges to some function v with respect to the supremum norm on each K_j. In particular, $\lim_{l\to\infty} v_{n(l)}(s) = v(s)$ for all $s \in \overline{V}$. It remains to show that $\|v_{n(l)} - v\|_X \to 0$. Let $\epsilon > 0$. By virtue of (5.6d) there exists a constant $C > 0$ such that $|v_{n(l)}(s)| \leq C$ for all $s \in \overline{V}$ and $l \in \mathbb{N}$. Therefore,

$$\frac{|v(s) - v_{n(l)}(s)|}{1 + |s|^2} \leq \epsilon \qquad \text{for all } l \in \mathbb{N} \text{ and } |s| \geq \sqrt{\frac{2C}{\epsilon}}. \tag{5.11}$$

Let $J \geq \sqrt{2C/\epsilon}$, $J \in \mathbb{N}$. Since $v_{n(l)}$ converges to v with respect to $\|\cdot\|_{\infty,K_J}$, there exists $L \in \mathbb{N}$ such that

$$\sup_{s\in K_J} \frac{\big|v(s) - v_{n(l)}(s)\big|}{|s^2|+1} \leq \|v(s) - v_{n(l)}\|_{\infty,K_J} \leq \epsilon \tag{5.12}$$

for $l \geq L$. Putting (5.11) and (5.12) together yields $\|v - v_{n(l)}\|_X \leq \epsilon$ for $l \geq L$. ∎

PROPOSITION 5.3. *The integral equation* (5.3) *has a unique solution in* X *for all* $U(a), U'(a) \in \mathbb{C}$.

Proof. Let $w \in X$ satisfy the homogeneous equation $w + Jw = 0$. If we can show that $w = 0$, then the assertion follows from Riesz theory and Lemma 5.2. Inequality (5.6d) implies that $\|w\|_\infty < \infty$. Hence, there exists $s^* \in \mathbb{C}$ with $|\operatorname{Re} s^*| \geq \sigma := 2\sqrt{c/a^2}$ such that $|w(s)| \leq 2|w(s^*)|$ for all $s \in \mathbb{C}$ with $|\operatorname{Re} s| \geq \sigma$. It follows from (5.6a) that

$$\begin{aligned} |w(s^*)| = |(Jw)(s^*)| &\leq \sup_{\operatorname{Re} s_1 \geq \sigma} \frac{c}{|s_1^2|+1}|w(s_1)| \\ &< \frac{1}{4} \sup_{\operatorname{Re} s_1 \geq \sigma} |w(s_1)| \leq \frac{1}{2}|w(s^*)|, \end{aligned}$$

i.e., $w(s^*) = 0$. This, however, implies that $w(s) = 0$ for all $s \in \mathbb{C}$ with $\operatorname{Re} s \geq \sigma$, and since w is holomorphic in V and continuous in $\overline{V}$, $w(s) = 0$ for all $s \in \overline{V}$. ∎

6. The cut functions. In this section we study the cut functions

$$\psi_{a,\pm}(t) := \frac{[\hat{U}_a](\pm i - t)}{2\pi i \operatorname{Res}_{\pm i} \hat{U}_a}, \qquad t > 0 \tag{6.1}$$

(cf. (5.5)). Note that $\pm 2i \operatorname{Res}_{\pm i} \hat{U}_a = w(\pm i)$. In the next lemma we derive Volterra integral equations for $\psi_{a,\pm}$, which are uniquely solvable. This shows that $\psi_{a,\pm}$ can be defined without the assumption $\operatorname{Res}_i \hat{U}_a \neq 0$ and that, in fact, $\psi_{a,\pm}$ depends only on P_a but not on U.

LEMMA 6.1. *If* $\operatorname{Res}_{\pm i} \hat{U}_a \neq 0$, *the cut functions* $\psi_{a,\pm}(t)$ *defined by* (6.1) *for* $t > 0$ *satisfy the integral equations*

$$\psi_{a,+}(t) + \int_0^t \frac{P_a(t-t_1)}{t(t-2i)} \psi_{a,+}(t_1)\,\mathrm{d}t_1 = -\frac{P_a(t)}{t(t-2i)}, \tag{6.2a}$$

$$\psi_{a,-}(t) + \int_0^t \frac{P_a(t-t_1)}{t(t+2i)} \psi_{a,-}(t_1)\,\mathrm{d}t_1 = -\frac{P_a(t)}{t(t+2i)}. \tag{6.2b}$$

Proof. We will only prove (6.2a) since the proof of (6.2b) is analogous. Due to (5.3) and (5.2) we have

$$(s_\pm^2 + 1)\hat{U}(s_\pm) + \int_{\gamma_\pm^\epsilon} P(s_1 - s_\pm)\hat{U}(s_1)\,\mathrm{d}s_1 = s_\pm U(a) + U'(a)$$

for $s = i - t \in S_i$ and $\epsilon > 0$. The paths γ_+^ϵ and γ_-^ϵ are shown in Figure 6.1. Subtracting the equation with the $+$ sign from the equation with the $-$ sign yields

$$\begin{aligned} &t(t-2i)[\hat{U}](i-t) + \int_{-t}^{-\epsilon} P(t+t_1)[\hat{U}](i+t_1)\,\mathrm{d}t_1 \\ &+ \int_0^{2\pi} P(t - \epsilon e^{-i\varphi}) \frac{w(i - \epsilon e^{i\varphi})}{(-\epsilon e^{i\varphi})(2i - \epsilon e^{i\varphi})} (-i\epsilon) e^{i\varphi}\,\mathrm{d}\varphi = 0. \end{aligned}$$

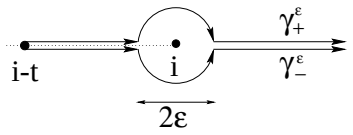


FIG. 6.1. *Integration path in the proof of Lemma* 6.1.

Since w is continuous at i, the last integral converges to $\pi P(t)w(i) = 2\pi i P(t)\operatorname{Res}_i \hat{U}$ as $\epsilon \to 0$. Dividing by $2\pi i(\operatorname{Res}_i \hat{U})t(t-2i)$ establishes (6.2a). ∎

Since both the kernel of the integral operators and the right-hand sides are bounded due to our assumptions on P, the Volterra integral equations (6.2) are uniquely solvable (cf., e.g., [13, Thm. 10.15]). From these integral equations we can deduce the following two lemmas concerning the behavior of $\psi_\pm$ near 0 and near ∞. Since

$$\psi_-(t) = \overline{\psi_+(t)} \tag{6.3}$$

the first lemma is only formulated for ψ_+.

LEMMA 6.2. *The function ψ_+ defined in* (6.1) *belongs to $C^\infty([0,\infty))$, and the derivatives of ψ_+ at* 0 *can be computed recursively as follows:*

$$\psi_+(0) = -\lim_{t\to 0} \frac{P(t)}{t(t-2i)}, \tag{6.4a}$$

$$\psi_+^{(k+1)}(0) = -\lim_{t\to 0} \frac{d^{k+1}}{dt^{k+1}} \left\{ \frac{P(t)}{t(t-2i)} \right\} + \frac{(k+1)!}{2i} \sum_{j=1}^{k+1} \frac{1}{(2i)^{k+1-j}(j+1)!} \sum_{n=1}^{j} P^{(n)}(0)\psi_+^{(j-n)}(0). \tag{6.4b}$$

Proof. Introducing $(Kv)(t) := \int_0^t \frac{P(t-t_1)}{t(t-2i)} v(t_1)\,\mathrm{d}t_1$, (6.2a) can be written as

$$\psi_+(t) + (K\psi_+)(t) = -\frac{P(t)}{t(t-2i)}, \qquad t > 0. \tag{6.5}$$

By repeated partial integration we obtain

$$\int_0^t P(t-t_1)\frac{t_1^j}{j!}\,\mathrm{d}t_1 = \sum_{l=1}^{\infty} P^{(l)}(0) \int_0^t \frac{(t-t_1)^l}{l!}\frac{t_1^j}{j!}\,\mathrm{d}t_1 = \sum_{l=1}^{\infty} \frac{P^{(l)}(0)t^{l+j+1}}{(l+j+1)!}.$$

Changing the order of integration and summation in the first equality is justified because the Taylor series of P converges uniformly. The right-hand side of the last equation is an analytic function in t. If $v(t) = \sum_{j=0}^{\infty} \frac{v^{(j)}(0)}{j!}t^j$ is a polynomial, then

$$\begin{aligned}(Kv)(t) &= \frac{1}{t(t-2i)} \int_0^t P(t-t_1) \sum_{j=0}^{\infty} \frac{v^{(j)}(0)}{j!} t_1^j\,\mathrm{d}t_1 \\ &= \frac{1}{t-2i} \sum_{l=1}^{\infty}\sum_{j=0}^{\infty} P^{(l)}(0)v^{(j)}(0)\frac{t^{l+j}}{(l+j+1)!}.\end{aligned}$$

Expanding $(t-2i)^{-1}$ in a power series and using the Cauchy product twice yields

$$(Kv)(t) = \frac{1}{-2i}\sum_{r=0}^{\infty}\left(\frac{t}{2i}\right)^r \sum_{m=1}^{\infty}\frac{t^m}{(m+1)!}\sum_{n=1}^{m}P^{(n)}(0)v^{(m-n)}(0) \tag{6.6}$$
$$= \frac{1}{-2i}\sum_{l=1}^{\infty}t^l\sum_{j=1}^{l}\left(\frac{1}{2i}\right)^{l-j}\frac{1}{(j+1)!}\sum_{n=1}^{j}P^{(n)}(0)v^{(j-n)}(0).$$

In particular, it can be seen that Kv is analytic at $t=0$.

We prove by induction that $\psi_+ \in C^n([0,\infty))$ for $n \in \mathbb{N}$. Note that the right-hand side of (6.5) is analytic at $t=0$ due to our assumptions on P. The case $n=0$ follows from the continuity and boundedness of the kernel of the integral operator K for $0 \le t_1 < t$. Equation (6.4a) is a consequence of $(K\psi_+)(0)=0$. Assume now that $\psi_+ \in C^k([0,\infty))$, $k \ge 0$. Then there exists a function R_k such that $\psi_+(t) = \sum_{j=0}^{k}\frac{\psi_+^{(j)}(0)}{j!}t^j + R_k(t)$ and $R_k(t) = \mathrm{o}\big(t^k\big)$ as $t \to 0$. We have

$$|(KR_k)(t)| \le \frac{1}{t|t-2i|}\int_0^t |P(t-t_1)|\,\mathrm{d}t_1 \sup_{0\le t_1\le t}|R_k(t_1)| = \mathrm{o}\big(t^{k+1}\big)$$

since $P(t) = \mathcal{O}(t)$ as $t \to 0$. Therefore, $KR_k \in C^{k+1}([0,\infty))$ and $(KR_k)^{(k+1)}(0)=0$. Now it follows from (6.5) and (6.6) that $\psi_+ \in C^{k+1}([0,\infty))$ and that $\psi_+^{(k+1)}(0)$ satisfies (6.4b). ∎

Since the first term on the right-hand side of (the analogue of) (2.4) does not contribute to $\operatorname{Res}_{\pm i}\hat{U}_a$ and $\psi_{a,\pm}$, the quantities

$$U_\infty^\pm := e^{\pm ia}\operatorname{Res}_{\pm i}\hat{U}_a \tag{6.7}$$

are independent of a, and

$$e^{at}\psi_{a,\pm}(t) = e^{bt}\psi_{b,\pm}(t) \tag{6.8}$$

for $b \ge a$. Alternatively, (6.8) follows from (6.2) using $e^{at}P_a(t) = e^{bt}P_b(t)$ (cf. (4.2)). Note that $\psi_{b,\pm}$ can be defined for all $b \in \mathbb{R}$ as a solution to (6.2) even if $\hat{U}_b$ does not exist.

LEMMA 6.3. *For all $a \in \mathbb{R}$ and $\epsilon > 0$ the cut functions satisfy*

$$|\psi_{a,\pm}(t)| = \mathcal{O}\Big(e^{-(a-a_p-\epsilon)t}\Big), \qquad t \to \infty. \tag{6.9}$$

Proof. Due to (6.8) it suffices to prove the lemma for $a = a_p$. It follows from (4.4a) with $k=0$ and (6.2) that

$$|\psi_{a_p,\pm}(t)| \le \frac{C}{|t(t-2i)|}\left(1 + \int_0^t |\psi_{a_p,\pm}(t_1)|\,\mathrm{d}t_1\right)$$

for all $t > 0$. Choosing t_* such that $\frac{C}{|t_*(t_*-2i)|} \le \epsilon$, it follows that

$$|\psi_{a_p,\pm}(t)| \le \Gamma + \int_{t_*}^t \epsilon|\psi_{a_p,\pm}(t_1)|\,\mathrm{d}t_1, \qquad t > t_*,$$

where $\Gamma := \epsilon(1+\int_0^{t_*}|\psi_{a_p,\pm}(t_1)|\,\mathrm{d}t_1)$. Now Gronwall's lemma (cf. [4]) implies $|\psi_{a_p,\pm}(t)| \le \Gamma e^{\epsilon(t-t_*)}$ for $t > t_*$. ∎

THEOREM 6.4. *The function U has a holomorphic extension to $\{\zeta \in \mathbb{C} : \operatorname{Re}\zeta > a\}$, and U and $U^{(k)}$ $(k \geq 1)$ satisfy the representation formulas*

$$\begin{aligned} U(z+a) &= U_\infty^+ e^{i(z+a)} \left(1 + \int_0^\infty e^{-tz}\psi_{a,+}(t)\,\mathrm{d}t\right) \\ &\quad + U_\infty^- e^{-i(z+a)} \left(1 + \int_0^\infty e^{-tz}\psi_{a,-}(t)\,\mathrm{d}t\right), \end{aligned} \tag{6.10a}$$

$$\begin{aligned} U^{(k)}(z+a) &= U_\infty^+ e^{i(z+a)} \left(i^k + \int_0^\infty (i-t)^k e^{-tz}\psi_{a,+}(t)\,\mathrm{d}t\right) \\ &\quad + U_\infty^- e^{-i(z+a)} \left((-i)^k + \int_0^\infty (-i-t)^k e^{-tz}\psi_{a,-}(t)\,\mathrm{d}t\right) \end{aligned} \tag{6.10b}$$

for $\operatorname{Re}(z) > 0$ *and* $a \geq a_p$.

Proof. Let $\gamma_1^R(t) := 1 + it$, $-R \leq t \leq R$ (cf. Figure 6.2(a)). Then

$$U(a+r) = \lim_{R\to\infty} \frac{1}{2\pi i} \int_{\gamma_1^R} e^{rs}\hat{U}_a(s)\,\mathrm{d}s$$

for $a > a_p$ and $r \geq 0$ by the inversion theorem for the Fourier transform. Now

$$\frac{1}{2\pi i} \int_{\gamma_1^R} e^{rs}\hat{U}_a(s)\,\mathrm{d}s = -\frac{1}{2\pi i} \int_{\gamma_2^R} e^{rs}\hat{U}_a(s)\,\mathrm{d}s$$

by virtue of Cauchy's integral theorem. Due to (5.2), (5.3), and (5.6d) the function $\hat{U}$ decays of order $|\hat{U}_a(s)| = \mathcal{O}\big(|s|^{-1}\big)$ as $|s| \to \infty$ uniformly for all directions. Therefore, the integrals from B to C, from D to E, and from F to A vanish asymptotically as $R \to \infty$. Moreover, given $\epsilon > 0$, there exists $\delta > 0$ such that $|\int_{\pi/2}^{\pi/2+\delta} e^{rs_R(\tau)}\hat{U}_a(s_R(\tau))s_R'(\tau)\,\mathrm{d}\tau| \leq \pi\epsilon$ for all $R \geq 1$, where $s_R(\tau) := i + R\exp(i\tau)$. Due to the exponential decay of the integrand, there exists $R_0 \geq 1$ such that $|\int_{\pi/2+\delta}^{\pi} e^{rs_R(\tau)}\hat{U}_a(s_R(\tau))s_R'(\tau)\,\mathrm{d}\tau| \leq \pi\epsilon$ for $R \geq R_0$. Both estimates together imply that $|(2\pi i)^{-1}\int_C^D e^{rs}\hat{U}_a(s)\,\mathrm{d}s| \leq \epsilon$ for $R \geq R_0$. Analogously, it can be shown that the integral from E to F tends to 0 as $R \to \infty$. A computation similar to that in the proof of Lemma 6.1 shows that the integrals around $\pm i$ converge to $\operatorname{Res}_{\pm i}\hat{U}_a e^{\pm ir}$ as $R \to \infty$. The integrals along $S_{\pm i}$ converge to $\operatorname{Res}_{\pm i}\hat{U}_a \int_0^\infty e^{r(\pm i - t)}\psi_{a,\pm}(t)\,\mathrm{d}t$ as $R \to \infty$. This yields (6.10a) for real $z \geq 0$. Differentiating (6.10a) and changing the order of differentiation and integration, which is possible by Lebesgue's dominated convergence theorem and (6.9), we obtain (6.10b). It is obvious that (6.10a) defines a holomorphic extension of U. ∎

By (6.10a) we can decompose any Fourier mode U satisfying (5.1) into an outgoing part and an incoming part. For a solution u to the full partial differential equation (2.1) a corresponding decomposition is not always possible. For example, the solution $u(x) = e^{i\kappa x_1}$ for $p = q = 0$ does not decay like $\mathcal{O}\big(\rho^{-(d-1)/2}\big)$. Since the Sommerfeld radiation condition implies such a behavior (cf. [2, sect. 2.2]) and since incoming solutions are complex conjugates of outgoing solutions, u cannot be decomposed into an outgoing and an incoming part. The reason that Theorem 6.4 does not carry over in full extent to the partial differential equation (2.1) is linked to the fact that the condition numbers of the matrices L_a in the next corollary increase exponentially with $|\lambda_j|$.

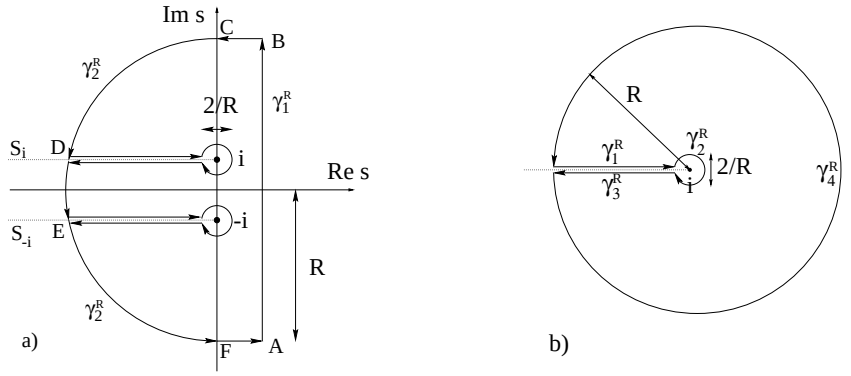


FIG. 6.2. *Integration paths in the proofs of Theorem* 6.4 *and Proposition* 6.6.

COROLLARY 6.5. *The matrix* L_a *defined by*

$$\begin{pmatrix} U(a) \\ U'(a) \end{pmatrix} = L_a \begin{pmatrix} U_\infty^+ \\ U_\infty^- \end{pmatrix}$$

via (6.10a) *and* (6.10b) *is regular. Hence, there exists* $(U(a), U'(a)) \neq 0$ *such that* $\operatorname{Res}_{-i} \hat{U}_a = 0$. *If* $\operatorname{Res}_{-i} \hat{U}_a = 0$, *then* $\hat{U}_a$ *is holomorphic in* $\mathbb{C}\backslash(S_i \cup \{i\})$; *i.e., it satisfies the pole condition.*

Proof. Let $L_a(U_\infty^+, U_\infty^-)^T = 0$. Since U solves a linear second order differential equation, $U(r) = U'(r) = 0$ implies that $U \equiv 0$, and hence $U_\infty^+ = U_\infty^- = 0$. Hence, L_a is regular. If $\operatorname{Res}_{-i} \hat{U}_a = 0$, then $[U_a](-i - t) = 0$, and therefore $[w](-i - t) = 0$ for all $t > 0$; i.e., w is continuous in the lower half-plane. Using Morera's theorem and a contour deformation around the cut S_{-i} it can be shown that w is holomorphic in $\mathbb{C}\backslash(S_i \cup \{i\})$. As $w(-i) = 0$, $\hat{U}_a$ is holomorphic in $\mathbb{C}\backslash(S_i \cup \{i\})$ as well. ∎

Finally, we need a representation formula for $\hat{U}_a$ in terms of the cut function.

PROPOSITION 6.6. *Let* $\operatorname{Res}_{-i} \hat{U}_a = 0$. *Then* $\hat{U}_a(s)$ *satisfies the representation formula*

$$\hat{U}_a(s) = -\frac{\operatorname{Res}_i \hat{U}_a}{i - s} - \int_0^\infty \frac{\operatorname{Res}_i \hat{U}_a \psi_{a,+}(t)}{i - t - s} \, \mathrm{d}t, \qquad s \in \mathbb{C}\backslash S_i. \tag{6.11}$$

Proof. Due to Corollary 6.5 the assumption $\operatorname{Res}_{-i} \hat{U}_a = 0$ implies that $\hat{U}_a$ is holomorphic in $\mathbb{C}\backslash(S_i \cup \{i\})$. Therefore, Cauchy's formula

$$\hat{U}_a(s) = \frac{1}{2\pi i} \int_{\gamma_1^R + \gamma_2^R + \gamma_3^R + \gamma_4^R} \frac{\hat{U}_a(s_1) \, \mathrm{d}s_1}{s_1 - s}$$

with the contour shown in Figure 6.2(b) holds true. The integral over γ_2^R converges to the first term on the right-hand side of (6.11). Recall from the proof of Theorem 6.4 that $\hat{U}_a(s) = \mathcal{O}(|s|^{-1})$ as $|s| \to \infty$ uniformly for all directions. Hence, the integral over γ_4^R tends to 0 as $R \to \infty$ since the integrand is of order $\mathcal{O}(|s|^{-2})$. Finally, the integrals over γ_1^R and γ_3^R yield the integral term in (6.11). ∎

7. Asymptotic expansion of the far field. The following result is a simple consequence of Theorem 6.4.

THEOREM 7.1. *Let $m \in \{0,1,2,\dots\}$, and assume that $U_\infty^+ = 0$ or $U_\infty^- = 0$, respectively. Then U and U' satisfy the asymptotic formulas*

$$U(z) = U_\infty^\pm e^{\pm iz}\left(1 + \sum_{l=0}^{m-1} \frac{\psi_{0,\pm}^{(l)}(0)}{z^{l+1}} + \mathcal{O}\Big(\frac{1}{|z|^{m+1}}\Big)\right), \tag{7.1}$$

$$U'(z) = U_\infty^\pm e^{\pm iz}\left(\pm i + \sum_{l=0}^{m-1} \frac{\pm i\psi_{0,\pm}^{(l)}(0) - l\psi_{0,\pm}^{(l-1)}(0)}{z^{l+1}} + \mathcal{O}\Big(\frac{1}{|z|^{m+1}}\Big)\right),$$

respectively, for $z \to \infty$ such that $|\arg z| \le \varphi < \frac{\pi}{2}$. Here $0 \cdot \psi_{0,\pm}^{(-1)}(0) := 0$.

Proof. Note that the integral term in (6.10a) is the Laplace transform of $\psi_{a,\pm}$. Due to (6.8) we may choose $a = 0$. Using the asymptotic formula

$$(\mathcal{L}f)(z) = \sum_{l=0}^{m-1} \frac{f^{(l)}(0)}{z^{l+1}} + \mathcal{O}\big(|z|^{-m-1}\big), \tag{7.2}$$

$z \to \infty, |\arg z| \le \varphi < \pi/2$, which holds for bounded functions $f \in C^m([0,\infty))$ (cf. [5, p. 47]), (6.10a), and Lemma 6.2, we immediately obtain (7.1). The asymptotic formula for U' follows analogously from (6.10b) and the identity $\frac{d^l}{dt^l}\left((\pm i - t)\psi_{0,\pm}(t)\right)\big|_{t=0} = \pm i\psi_{0,\pm}^{(l)}(0) - l\psi_{0,\pm}^{(l-1)}(0)$. ∎

As a special case of the previous theorem we reproduce the asymptotic formula for the Hankel functions for large arguments (cf. [21]).

COROLLARY 7.2. *The Hankel functions $H_j^{(1)}$ of the first kind of order j satisfy*

$$H_j^{(1)}(z) = \sqrt{\frac{2}{\pi z}} e^{i(z - \frac{j\pi}{2} - \frac{\pi}{4})}\left(\sum_{k=0}^{m}\left\{\prod_{l=1}^{k} \frac{j^2 - (l-\frac{1}{2})^2}{-2ilz}\right\} + \mathcal{O}\big(|z|^{-m-1}\big)\right)$$

for $z \to \infty$ such that $|\arg z| \le \varphi < \frac{\pi}{2}$ ($m \ge 0$).

Proof. With $P(t) = e^{-at}t(\frac{1}{4} - j^2)$, $U_\infty^+ = \sqrt{\frac{2}{\pi}}\exp\left(-i\frac{j\pi}{2} - i\frac{\pi}{4}\right)$ we get $H_j^{(1)}(\rho) = \rho^{-1/2}U(\rho)$. Using the identity $\frac{1}{t(t-2i)} = \frac{1}{-2it}\sum_{l=0}^{\infty} t^l(2i)^{-l}$ and (6.4) we obtain

$$\begin{aligned}
\psi_{0,+}(0) &= \frac{j^2 - \frac{1}{4}}{-2i},\\
\psi_{0,+}^{(k+1)}(0) &= \left(\frac{k+1}{2i} + \frac{1}{2i(k+2)}\left(\frac{1}{4} - j^2\right)\right)\psi_{0,+}^{(k)}(0)\\
&= \frac{1}{-2i(k+2)}\left(j^2 - \left(k + \frac{3}{2}\right)^2\right)\psi_{0,+}^{(k)}(0).
\end{aligned}$$

Now the assertion follows from (7.1). ∎

8. Spectral properties of the Dirichlet-to-Neumann map. Let $\mathcal{H}_j$ denote the solution to (4.1) with $U_\infty^+ = 1$ and $U_\infty^- = 0$. For the Helmholtz equation in $\mathbb{R}^2$ we have $H_j^{(1)}(\rho) = \sqrt{\frac{2}{\pi\rho}}\exp(-i\frac{j\pi}{2} - i\frac{\pi}{4})\mathcal{H}_{2j}(\rho)$. A complete orthonormal system in $L^2(\Gamma_a)$ is given by $\varphi_j^a(a\hat{x}) := a^{-\frac{d-1}{2}}\varphi_j(\hat{x})$ with the eigenfunctions φ_j of $\tilde{\Delta}_{\hat{x}}$ introduced in section 4. We expect that the solution to (2.1) satisfying the pole condition and

the boundary condition $\mathrm{Tr}_{\Gamma_a} u = f$ ($f \in H^{1/2}(\Gamma_a)$) is

$$u(\rho\hat{x}) = \sum_{j=1}^{\infty} \langle f, \varphi_j^a \rangle \, \rho^{-\frac{d-1}{2}} \frac{\mathcal{H}_j(\rho)}{\mathcal{H}_j(a)} \varphi_j(\hat{x}) \tag{8.1}$$

for $\rho \geq a$ and $\hat{x} \in S^{d-1}$. Since the Dirichlet-to-Neumann map $\mathrm{DtN} : H^{1/2}(\Gamma_a) \to H^{-1/2}(\Gamma_a)$ satisfies $(\mathrm{DtN} f)(\rho\hat{x}) = \frac{\partial}{\partial\rho} u(\rho\hat{x})|_{\rho=a}$, this leads to the definition

$$(\mathrm{DtN} f)(\rho\hat{x}) := \sum_{j=1}^{\infty} \mathrm{dtn}(\lambda_j) \, \langle f, \varphi_j \rangle \, \varphi_j(\hat{x}) \tag{8.2}$$

with the eigenvalues

$$\mathrm{dtn}(\lambda_j) = \frac{\left(\rho^{-\frac{d-1}{2}} \mathcal{H}_j(\rho)\right)' \big|_{\rho=a}}{a^{-\frac{d-1}{2}} \mathcal{H}_j(a)} = \frac{\mathcal{H}_j'(a)}{\mathcal{H}_j(a)} - \frac{d-1}{2a}. \tag{8.3}$$

The Sobolev norm on Γ_a of index $s \in \mathbb{R}$ is given by

$$\|f\|^2_{H^s(\Gamma_a)} = \|(I - \Delta_{\Gamma_a})^{s/2} f\|^2_{L^2(\Gamma_a)} \sim \sum_{j=1}^{\infty} (1 - \lambda_j)^s \, |\langle f, \varphi_j^a \rangle|^2$$

(cf. [20, Chap. 4]). Hence, properties 1–3 in Proposition 3.1 are equivalent to

$$|\mathrm{dtn}(\lambda_j)| = \mathcal{O}\left(\sqrt{|\lambda_j|}\right), \qquad j \to \infty, \tag{8.4a}$$

$$\mathrm{Re}(-\mathrm{dtn}(\lambda_j) + l_j) \geq 0 \qquad \text{for some sequence } |l_j| = \mathrm{o}\left(\sqrt{|\lambda_j|}\right), \tag{8.4b}$$

$$\mathrm{Im}\, \mathrm{dtn}(\lambda_j) > 0 \qquad \text{for all } j. \tag{8.4c}$$

LEMMA 8.1. *Let U satisfy the assumptions of sections 5–7, and let $U_\infty^- = 0$. Then*

$$\mathrm{Im}\, U'(\rho)\overline{U(\rho)} = \left|U_\infty^+\right|^2 \qquad \textit{for all } \rho \geq a. \tag{8.5}$$

Proof. Set $\tilde{p}(a + r) := (\mathcal{L}P)(r)$. Taking the imaginary part of

$$\begin{aligned} 0 &= \int_a^\rho (U'' + (1 + \tilde{p})U) \, \overline{U} \, \mathrm{d}\rho_1 \\ &= U'(\rho)\overline{U(\rho)} - U'(a)\overline{U(a)} + \int_a^\rho \left(-|U'|^2 + (1 + \tilde{p}) \, |U|^2\right) \mathrm{d}\rho_1 \end{aligned}$$

yields $\mathrm{Im}\, U'(\rho)\overline{U(\rho)} = \mathrm{Im}\, U'(a)\overline{U(a)} = \text{const}$. The constant can be evaluated using Theorem 7.1 by taking the limit $\rho \to \infty$. ∎

Equation (8.5) with $\rho = a$ implies (8.4c) after dividing by $|U(a)|^2$. Next, we will prove (8.4a) and (8.4b). Let $\nu_j = \sqrt{-\lambda_j}$, and let $\psi_{\nu,a,+}$ denote the solution to (6.2a) with $P(t) = \check{p}_a(t) - \nu_j^2 t e^{-at}$. Since $\lambda_j \to -\infty$ as $j \to \infty$, it follows that $\nu_j \to \infty$. For looking at this limit process, we may assume w.l.o.g. that $p_2 = \check{p}_a'(0) = 0$ by setting $\nu_j = \sqrt{-p_2 - \lambda_j}$. Multiplying (6.2a) by $t(t-2i)$, applying the Laplace transform, and

using the identity $\mathcal{L}(\int_0^t f(t-t_1)g(t_1)\,\mathrm{d}t_1) = (\mathcal{L}f)\cdot(\mathcal{L}g)$ yields the ordinary differential equation

$$\left(\partial_z^2 + 2i\partial_z + p(z) - \nu_j^2 z^{-2}\right) v(z;\nu_j) = 0 \tag{8.6}$$

for $v(z;\nu_j) := 1 + (\mathcal{L}\psi_{\nu,0,+})(z)$. Here and in the following we use the variables $z = \rho + i\sigma$ with $\rho, \sigma \in \mathbb{R}$. Since

$$\mathcal{H}_j(z) = e^{iz} v(z;\nu_j) \tag{8.7}$$

due to (6.10a), (8.6) can alternatively be derived immediately from (4.1).

In this section f' denotes the usual derivative of a holomorphic function f, whereas $\dot{f}$ denotes the partial derivative of f with respect to σ. By the chain rule, $\dot{f} = if'$, so $-\ddot{v} + 2\dot{v} + (p - \nu^2 z^{-2})v = 0$, where the argument (z) of p and the arguments (z,ν) of v have been omitted. Hence, the logarithmic derivative $\chi(z;\nu) := \frac{\dot{v}(z;\nu)}{v(z;\nu)}$ satisfies the Riccati differential equation

$$\dot{\chi}(z;\nu) + \chi^2(z;\nu) - 2\chi(z;\nu) = p(z) - \nu^2 z^{-2}. \tag{8.8}$$

It follows from Plancherel's theorem and (6.9) that

$$\int_{-\infty}^{\infty} |v(\rho + i\sigma;\nu) - 1|^2 \,\mathrm{d}\sigma = \frac{1}{2\pi}\|\psi_{\nu,\rho,+}\|_{L^2}^2 < \infty,$$
$$\int_{-\infty}^{\infty} |\dot{v}(\rho + i\sigma;\nu)|^2 \,\mathrm{d}\sigma = \frac{1}{2\pi}\|it\psi_{\nu,\rho,+}\|_{L^2}^2 < \infty.$$

Therefore, the Lebesgue measure of the sets $A_\epsilon(\rho,\nu) := \{\sigma : |v(\rho+i\sigma)-1| > \epsilon \text{ or } |\dot{v}(\rho+i\sigma)| > \epsilon\} < \infty$ is finite for all $\epsilon > 0$. Hence, for all $\rho \geq a$ and all $\nu \geq 0$ there exists a sequence σ_l such that $\sigma_l \notin A_{1/l}(\rho,\nu)$ and $\sigma_l > l$. This implies

$$\lim_{l\to\infty} \chi(\rho + i\sigma_l;\nu) = 0. \tag{8.9}$$

We now construct an approximation to $\chi(\rho + i\sigma;\nu)$ for $\sigma \geq 0$ by formal computations and then prove its validity. We rewrite (8.8) as $\dot{\chi} = -(\chi - 1 + \gamma_1)(\chi - 1 - \gamma_1)$ with $\gamma_1(z;\nu) := \sqrt{1 + p(z) - \nu^2 z^{-2}}$. Here and in the following we choose the negative real axis as the branch cut of the square root function. Neglecting the term $\dot{\chi}$ yields the two possible approximations $1 + \gamma_1$ and $1 - \gamma_1$. Only the latter of these approximations has the right behavior as $\sigma \to \infty$. The "error function" $\Delta_1 := \chi - 1 + \gamma_1$ satisfies the differential equation

$$\dot{\Delta_1} = \dot{\gamma}_1 - (\Delta_1 - 2\gamma_1)\Delta_1. \tag{8.10}$$

Since this equation has the same structure as (8.8), we can apply the same procedure as above to (8.10) and hopefully get a better approximation to χ. This process may be repeated recursively as follows: Set $\gamma_0 := 1$ and assume we have constructed a function γ_j $(j = 1, 2, \dots)$ such that

$$\chi = 1 - \gamma_j + \Delta_j, \tag{8.11}$$

where Δ_j satisfies the differential equation

$$\dot{\Delta_j} = -(\Delta_j - 2\gamma_j)\Delta_j + \dot{\gamma}_j - \dot{\gamma}_{j-1}. \tag{8.12}$$

This equation can be rewritten as $\dot{\Delta}_j = -(\Delta_j - \gamma_j - \gamma_{j+1})(\Delta_j - \gamma_j + \gamma_{j+1})$ with

$$\gamma_{j+1} := \sqrt{\gamma_j^2 + \dot{\gamma}_j - \dot{\gamma}_{j-1}}. \tag{8.13}$$

The function $\Delta_{j+1} := \Delta_j - \gamma_j + \gamma_{j+1}$ satisfies (8.11) and (8.12) with j replaced by $j+1$.

It turns out that the approximation of order $j = 2$ is the lowest that is sufficient for our purposes. In the appendix we establish the following bounds on $\Delta_2 = \chi - 1 + \gamma_2$.

LEMMA 8.2. *Given $0 < a < A < \infty$ there exist constants $\Gamma, N > 0$ such that for all $\rho \in [a, A]$ and all $\nu \geq N$*

$$|\Delta_2(\rho + i\sigma; \nu)| \leq \begin{cases} 2, & 0 \leq \sigma < \Gamma/\nu, \\ \Gamma/(\sigma\nu), & \Gamma/\nu \leq \sigma < \nu, \\ \Gamma/\sigma^2, & \nu \leq \sigma. \end{cases} \tag{8.14}$$

Moreover,

$$\gamma_2(\rho + i\sigma; \nu)^2 = 1 - \frac{\nu^2}{(\rho + i\sigma)^2}\left(1 + \mathcal{O}\left(\frac{1}{\nu}\right)\right) \tag{8.15}$$

as $\nu \to \infty$ uniformly for $\rho \in [a, A]$ and $\sigma \geq 0$.

It follows from (8.7) that $\frac{\mathcal{H}_j'(z)}{\mathcal{H}_j(z)} = i + \frac{v'(z;\nu_j)}{v(z;\nu_j)} = i(1 - \chi(z;\nu_j))$. Using (8.11), (8.14), and (8.15) we obtain

$$\frac{\mathcal{H}_j'(z)}{\mathcal{H}_j(z)} = i(\gamma_2(z;\nu_j) - \Delta_2(z;\nu_j)) = -\frac{\nu_j}{z} + \mathcal{O}(1) \tag{8.16}$$

as $j \to \infty$, which holds uniformly for z satisfying $\operatorname{Re} z \in [a, A]$ and $\operatorname{Im} z \in [0, S]$ for any $S \geq 0$. For $z = a$ this implies (8.4a) and (8.4b).

COROLLARY 8.3. *Given $a < R_2 < \infty$ and $S \geq 0$, there exist constants $C, N > 0$ such that*

$$\left|\frac{\mathcal{H}_j^{(l)}(\rho + i\sigma)}{\mathcal{H}_j(a + ib)}\right| \leq C\left(\frac{\nu_j}{|\rho + i\sigma|}\right)^l \left|\frac{a + ib}{\rho + i\sigma}\right|^{\nu_j} \tag{8.17}$$

for all $\nu_j \geq N$, $a \leq \rho \leq R_2$, $0 \leq b, \sigma \leq S$, and $l = 0, 1, 2$.

Proof. It follows from (8.16) that

$$\frac{\mathcal{H}_j(\rho + i\sigma)}{\mathcal{H}_j(a + ib)} = \exp\left(\int_{a+ib}^{\rho+i\sigma} \frac{\mathcal{H}_j'(\zeta_1)}{\mathcal{H}_j(\zeta_1)}\,\mathrm{d}\zeta_1\right) = \left(\frac{a + ib}{\rho + i\sigma}\right)^{\nu_j} \exp(\mathcal{O}(1)).$$

This implies (8.17) for $l = 0$. Together with (8.16) we obtain (8.17) for $l = 1$. The case $l = 2$ follows from differential equation (4.1). ∎

Let us summarize our results.

THEOREM 8.4. *For $f \in H^{1/2}(\Gamma_a)$ there exists a unique solution u to (2.1) in $\{x : |x| > a\}$ satisfying the pole condition and the boundary condition $\operatorname{Tr} u = f$ on Γ_a. u is given by the series (8.1), which converges uniformly on compact subsets of $\{x : |x| > a\}$ together with all its term-by-term derivatives of order ≤ 2. The corresponding Dirichlet-to-Neumann $\mathrm{DtN} f := \frac{\partial u}{\partial \nu}\big|_{\Gamma_a}$ is given by (8.2). It satisfies the assumptions of Proposition 3.1. Consequently, the variational problem (3.1) has a*

unique solution, which can be to extended to a solution of (2.1) *in the exterior domain such that the pole condition is satisfied.*

Proof. Assume that u is a solution for $f = 0$. We apply Lemma 8.1 to the Fourier modes U_j. Due to the pole condition we have $U_{j,\infty}^- = 0$. Since $U_j(a) = 0$, it follows from (8.5) that $U_{j,\infty}^+ = 0$. Corollary 6.5 implies that U_j has vanishing Cauchy data $U_j(a) = U_j'(a) = 0$. Hence $U_j = 0$. This proves uniqueness.

Convergence of the series (8.1) and its term-by-term derivatives as well as the series (8.2) follow from Corollary 8.3 and the estimates

$$\|\varphi_j\|_{C^l} \le C\|\varphi_j\|_{H^{l+d/2}} \le C\sqrt{1-\lambda_j}^{l+d/2} \tag{8.18}$$

($l = 0, 1, \dots$) on the eigenfunctions φ_j derived from Sobolev's embedding theorem on S^{d-1} (cf. [20, sect. 4.3]). Note that no division by zero can occur in the series (8.1) and (8.2) due to (8.5) since $\mathcal{H}_j(a) = 0$ would imply $\mathcal{H}_{j,\infty}^+ = 0$, which contradicts $\mathcal{H}_{j,\infty}^+ = 1$.

The assumptions of Proposition 3.1 have been established above using the equivalent formulation (8.4). The fact that we can extend the solution u^{int} to (3.1) by setting $f = \mathrm{Tr}_{\Gamma_a} u^{\mathrm{int}}$ follows from the Cauchy–Kovalevskaya theorem and elliptic regularity results. ∎

Finally, we establish an estimate for $|\mathcal{H}_j(a)|$ for large j. For the special case of Hankel functions it agrees with a well-known formula, which can be derived from the series representation of the Hankel functions (cf. [2, 21]).

PROPOSITION 8.5. *For any $a > a_p$ we have*

$$|\mathcal{H}_j(a)| = \exp\left(\nu_j \ln\frac{2\nu_j}{ea} + \mathcal{O}(\ln \nu_j)\right), \qquad j \to \infty. \tag{8.19}$$

Proof. It follows from the definition of χ and σ_l before (8.9) that

$$1 = \lim_{l\to\infty} |v(a + i\sigma_l; \nu_j)| = \exp\left(\mathrm{Re}\int_0^\infty \chi(a + i\sigma; \nu_j)\,\mathrm{d}\sigma\right)|v(a; \nu_j)|,$$

i.e., $|\mathcal{H}_j(a)| = \exp(-\mathrm{Re}\int_0^\infty \chi(a+i\sigma;\nu_j)\,\mathrm{d}\sigma)$. By virtue of Lemma 8.2, $\mathrm{Re}\int_0^\infty \Delta_2(a + i\sigma;\nu)\,\mathrm{d}\sigma = \mathcal{O}\big(\nu^{-1}\ln\nu\big)$ as $\nu \to \infty$. It can be seen from (8.15) that there exists a constant $C > 0$ such that $1 - \mathrm{Re}\,\gamma_2(a + i\sigma;\nu) \le 0$ and $1 - \mathrm{Re}\,\gamma_2 = (1 - \mathrm{Re}\sqrt{1-(\nu/z)^2})(1 + \mathcal{O}\big(\nu^{-1}\big))$ uniformly for $\sigma \ge C/\nu$ as $\nu \to \infty$. Moreover, we have $\int_0^{C/\nu} \mathrm{Re}\,|\gamma_2(a + i\sigma;\nu)|\,\mathrm{d}\sigma = \mathcal{O}(1)$. Hence,

$$\begin{aligned}
&\int_0^\infty \mathrm{Re}\,\chi(a + i\sigma;\nu)\,\mathrm{d}\sigma \\
&\quad= \int_{C/\nu}^\infty \left(1 - \mathrm{Re}\sqrt{1 - \frac{\nu^2}{(a+i\sigma)^2}}\right)\mathrm{d}\sigma\left(1 + \mathcal{O}\left(\frac{1}{\nu}\right)\right) + \mathcal{O}(1)\,.
\end{aligned}$$

Finally,

$$\begin{aligned}
&\int_{C/\nu}^\infty \left(1 - \mathrm{Re}\sqrt{1 - \frac{\nu^2}{(a+i\sigma)^2}}\right)\mathrm{d}\sigma = \mathrm{Re}\int_{a+iC/\nu}^{a+i\infty}\left(1 - \sqrt{1-\frac{\nu^2}{z^2}}\right)\frac{\mathrm{d}z}{i} \\
&= \mathrm{Re}\left(iz\left(-1 + \sqrt{1-\frac{\nu^2}{z^2}}\right) + \nu\ln\left(\frac{\nu}{z} - i\sqrt{1-\frac{\nu^2}{z^2}}\right)\right)\Big|_{z=a+iC/\nu}^{z=a+i\infty} \\
&= \nu - \nu\ln\frac{2\nu}{a} + \mathcal{O}(1)\,, \qquad \nu \to \infty.
\end{aligned}$$

This completes the proof. ∎

9. Equivalence to Sommerfeld's radiation condition. Crucial tools in the proofs of this section are the following uniform estimates of the cut functions.

LEMMA 9.1. *There exist constants $C, N \geq 0$ such that the estimates*

$$|\psi_{\nu,a,+}(t)| \leq C\nu^2 e^{-t(a-a_p)} \left(\frac{|\sqrt{t} + \sqrt{t-2i}|}{\sqrt{2}} \right)^{2\nu}, \tag{9.1a}$$

$$|\psi'_{\nu,a,+}(t)| \leq C \frac{\nu^3 e^{-t(a-a_p)}}{\sqrt{t|t-2i|}} \left(\frac{|\sqrt{t} + \sqrt{t-2i}|}{\sqrt{2}} \right)^{2\nu} \tag{9.1b}$$

hold true for all $t > 0$ and all $\nu \geq N$. If $p = 0$, then (9.1) is valid with $a_p = 0$, $C = \frac{1}{2}$, and $N = \sqrt{2}$.

Proof. Due to (6.8) it suffices to prove (9.1) for $a = 0$. Set $\check{p}_0(t) := e^{at}\check{p}_a(t)$ and recall that we have assumed w.l.o.g. that $\check{p}'_0(0) = 0$. Multiplying (6.2a) by $t(t-2i)$, differentiating twice, and dividing by $t(t-2i)$ yields the integrodifferential equation

$$\begin{aligned} \psi''_{\nu,0,+}(t) = {} & \frac{\nu^2 - 2}{t(t-2i)} \psi_{\nu,0,+}(t) - 4\frac{t-i}{t(t-2i)} \psi'_{\nu,0,+}(t) \\ & - \int_0^t \frac{\check{p}''_0(t-t_1)}{t(t-2i)} \psi_{\nu,0,+}(t_1)\, \mathrm{d}t_1 - \frac{\check{p}''_0(t)}{t(t-2i)}. \end{aligned} \tag{9.2}$$

Here $P(t) = -\nu^2 t + \check{p}_0(t)$ in (6.2a). We will derive bounds on the function

$$y(t) := \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} \psi_{\nu,0,+}(t) \\ \zeta_\nu(t)^{-1} \psi'_{\nu,0,+}(t) \end{pmatrix},$$

where $\zeta_\nu(t) := \frac{\sqrt{\nu^2-2}}{\sqrt{t(t-2i)}}$. The function ζ_ν is chosen such that both components of y have approximately the same size, i.e., such that ζ_ν approximates the logarithmic derivative $\psi'_{\nu,0,+}/\psi_{\nu,0,+}$. Using (9.2) and $\operatorname{Re}(\psi_{\nu,0,+}\overline{\psi'_{\nu,0,+}}) = \operatorname{Re}(\overline{\psi_{\nu,0,+}}\psi'_{\nu,0,+})$ we get

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}|y(t)|^2 = {} & \operatorname{Re}\left(\left(1 + \frac{\zeta_\nu^2}{|\zeta_\nu|^2}\right) \psi_{\nu,0,+}\overline{\psi'_{\nu,0,+}} - 3\frac{t-i}{t(t-2i)} \frac{|\psi'_{\nu,0,+}|^2}{|\zeta_\nu|^2} \right) \\ & - \operatorname{Re}\left(\frac{\check{p}''_0(t)}{t(t-2i)} \frac{\overline{\psi'_{\nu,0,+}}}{|\zeta_\nu|^2} + \int_0^t \frac{\check{p}''_0(t-t_1)}{t(t-2i)} \psi_{\nu,0,+}(t_1)\, \mathrm{d}t_1 \frac{\overline{\psi'_{\nu,0,+}}}{|\zeta_\nu|^2} \right). \end{aligned} \tag{9.3}$$

Using the identity $|1 + z^2| = 2\operatorname{Re} z$ with $z = \zeta_\nu/|\zeta_\nu|$ and the inequalities $2|ab| \leq |a|^2 + |b|^2$ and $\operatorname{Re}\zeta_\nu > 0$ we obtain

$$\left|1 + \frac{\zeta_\nu^2}{|\zeta_\nu|^2}\right| \cdot \left|\psi_{\nu,0,+}\overline{\psi'_{\nu,0,+}}\right| \leq \operatorname{Re}(\zeta_\nu) \left(|y_1|^2 + |y_2|^2\right).$$

Moreover, note that $\operatorname{Re}\frac{t-i}{t(t-2i)} = \frac{2+t^2}{t(t^2+4)} > 0$. If $p = 0$, this implies $\frac{1}{2}\frac{d}{dt}|y(t)|^2 \leq \operatorname{Re}(\zeta_\nu(t))|y(t)|^2$ and hence, by Gronwall's lemma,

$$|y(t)|^2 \leq \exp\left(2\int_0^t \operatorname{Re}\zeta_\nu(t_1)\, \mathrm{d}t_1 \right) |y(0)|^2.$$

Here $|y(0)| = |\psi_{\nu,0,+}(0)| = \frac{1}{2}\nu^2$ due to (6.4a). Using the indefinite integral $\int \big(t(t-2i)\big)^{-1/2}\,\mathrm{d}t = 2\ln(\sqrt{t}+\sqrt{t-2i})$ and the estimate $\sqrt{\nu^2-2} < \nu$ for $\nu \geq \sqrt{2}$, we obtain the assertion for the case $p = 0$.

Next we are going to derive a bound on $|y(t)|$ for $t \leq 1$ and general p. The inequality $-x^2 + 2xy - y^2 \leq 0$ for $x, y \in \mathbb{R}$ yields

$$-3\frac{2+t^2}{\nu^2|t-2i|}|\psi'_{\nu,0,+}|^2 + \frac{|\check{p}_0''(t)|}{\nu^2}|\psi'_{\nu,0,+}| \leq \frac{|t-2i|}{12(2+t^2)}\frac{|\check{p}_0''(t)|^2}{\nu^2}.$$

Moreover, $\int_0^t |\frac{\check{p}_0''(t-t_1)}{t(t-2i)}\psi_{\nu,0,+}(t_1)|\,\mathrm{d}t_1 \leq C\max_{0\leq t_1\leq t|}|y_1(t_1)|$ for $0 \leq t \leq 1$. Hence, (9.3) implies

$$\frac{1}{2}\frac{d}{dt}|y(t)|^2 < \mathrm{Re}(\zeta_\nu(t))|y(t)|^2 + \frac{C}{\nu^2} + \frac{C}{\nu}\max_{0\leq t_1\leq t}|y(t_1)|^2 \tag{9.4}$$

with a constant C independent of ν. Let $\varphi(t)$ be the solution to the initial value problem

$$\frac{1}{2}\varphi'(t) = \left(\frac{C}{\nu} + \mathrm{Re}(\zeta_\nu(t))\right)\varphi(t) + \frac{C}{\nu^2}, \qquad \varphi(0) = |y(0)|^2. \tag{9.5}$$

We claim that $|y(t)|^2 < \varphi(t)$ for $0 < t \leq 1$. Assume on the contrary that the set $M := \{0 < t \leq 1 : |y(t)|^2 \geq \varphi(t)\}$ is not empty, and let $t_* := \inf M$. Then $t_* > 0$ since $\frac{d}{dt}|y(0)|^2 < \varphi'(0)$. Moreover, it follows from the definition of t_* that $\varphi'(t_*) \leq \frac{d}{dt}|y(t_*)|^2$. On the other hand, $\max_{0\leq t\leq t_*}|y(t)|^2 \leq \max_{0\leq t\leq t_*}\varphi(t) = \varphi(t_*)$. Hence, $\frac{d}{dt}|y(t_*)|^2 < \varphi'(t_*)$ due to (9.4) and (9.5). This is a contradiction. Hence, $|y(t)|^2 < \varphi(t)$ for $0 < t \leq 1$. From the explicit solution

$$\begin{aligned}\varphi(t) = \exp\left(\int_0^t 2\left(\frac{C}{\nu} + \mathrm{Re}\,\zeta_\nu(t_2)\right)\mathrm{d}t_2\right)|y(0)|^2 \\ + \frac{2C}{\nu^2}\int_0^t \exp\left(\int_{t_1}^t 2\left(\frac{C}{\nu} + \mathrm{Re}\,\zeta_\nu(t_2)\right)\mathrm{d}t_2\right)\mathrm{d}t_1\end{aligned}$$

it follows that $\varphi(t) \leq C\nu^4\exp(2ta_p + \int_0^t 2\,\mathrm{Re}\,\zeta_\nu(t_2)\,\mathrm{d}t_2)$. This implies (9.1) for $0 < t \leq 1$.

Now we will prove (9.1) for $t \geq 1$. We may assume $|y(t)| \neq 0$. Otherwise (9.1) is trivially satisfied at t, and we have to apply the following argument separately on all intervals where $|y|$ does not vanish. Dividing (9.3) by $|y(t)|$ and using (4.4a) with $k = 2$ we obtain

$$|y(t)|' = \frac{1}{2|y(t)|}\frac{d}{dt}|y(t)|^2 \leq \mathrm{Re}(\zeta_\nu(t))|y(t)| + \frac{C}{t|(t-2i)\zeta_\nu(t)|}\eta(t)$$

with $\eta(t) := e^{a_p t}(1 + \int_0^t e^{-a_p t_1}|y(t_1)|\,\mathrm{d}t_1)$. Inserting the identities

$$|y(t)| = e^{a_p t}\left\{e^{-a_p t}\eta(t)\right\}' = \eta'(t) - a_p\eta(t), \tag{9.6a}$$

$$|y(t)|' = \eta''(t) - a_p\eta'(t) \tag{9.6b}$$

and using the estimate $\frac{C}{t|(t-2i)\zeta_\nu(t)|} \leq a_p\,\mathrm{Re}\,\zeta_\nu(t)$, which holds for all $t \geq 1$ and ν sufficiently large, we obtain $\eta'' \leq (a_p + \mathrm{Re}\,\zeta_\nu)\,\eta'$. This implies $\eta'(t) \leq \eta'(1)\exp(a_p(t-$

$1) + \int_1^t \operatorname{Re}\zeta_\nu(t_1)\,\mathrm{d}t_1)$ due to Gronwall's lemma. Now it follows from (9.6a) and (9.1) for $t = 1$ that $|y(t)| \le Ce^{a_p t}(\frac{|1+\sqrt{1-2i}|}{\sqrt{2}})^{2\nu}$ for $t > 1$. ∎

COROLLARY 9.2. *For $k \in \{0, 1, \dots\}$ and $m \in \{0, 1\}$ there exist constants $C, \sigma > 0$ such that for all $a > a_p$ and all $\nu_j \ge N$*

$$\frac{\|t^k \psi^{(m)}_{\nu_j,a,+}\|_{L^1}}{|\mathcal{H}_j(a)|} \le \frac{C}{(a-a_p)^{k+1-m}} \exp\left(\nu_j\left(\sigma + \ln\frac{a}{a-a_p}\right)\right). \tag{9.7}$$

Proof. We use the estimate

$$\frac{|\sqrt{t}+\sqrt{t-2i}|}{\sqrt{2}} \le \begin{cases} \gamma\sqrt{t}, & t \ge 1, \\ \gamma, & 0 \le t < 1 \end{cases}$$

with $\gamma := \frac{1+\sqrt{5}}{\sqrt{2}}$. Using Stirling's formula $\Gamma(x+1) = \exp(x\ln\frac{x}{e} + \mathcal{O}(\ln x))$ as $x \to \infty$, we obtain

$$\begin{aligned}
&\int_0^\infty e^{-t(a-a_p)} \left(\frac{|\sqrt{t}+\sqrt{t-2i}|}{\sqrt{2}}\right)^{2\nu} t^k \,\mathrm{d}t \\
&\le \gamma^{2\nu} \int_0^\infty e^{-(a-a_p)t} t^{\nu+k} \,\mathrm{d}t + \gamma^{2\nu} \int_0^1 e^{-(a-a_p)t} \,\mathrm{d}t \\
&\le \gamma^{2\nu} \left((a-a_p)^{-\nu-k-1}\Gamma(\nu+k+1) + 1\right) \\
&\le \frac{C}{(a-a_p)^{k+1}} \exp\left(\nu\ln\gamma^2 - \nu\ln(a-a_p) + \nu\ln\frac{\nu+k}{e} + \mathcal{O}(\ln\nu)\right).
\end{aligned}$$

Together with (9.1a) and (8.19) this implies (9.7) for $m = 0$. The case $m = 1$ follows analogously from (9.1b) using $\sqrt{t|t-2i|} \ge t$. ∎

THEOREM 9.3. *Assume that u satisfies (2.1) in $\{x \in \mathbb{R}^d : |x| > a_*\}$ such that $U(a_*, \cdot) \in L^2(S^{d-1})$ and that $\operatorname{Res}_{-i} \hat{U}_{j,a_*} = 0$ for all j. Let a be sufficiently large such that $\sigma + \ln\frac{a_*}{a-a_p} < 0$. Then the following are true:*

1. *The functions*

$$u_\infty(\hat{x}) := \sum_j \frac{U_j(a)}{\mathcal{H}_j(a)} \varphi_j(\hat{x}), \tag{9.8a}$$

$$\Psi_a(t, \hat{x}) := \sum_j \frac{U_j(a)}{\mathcal{H}_j(a)} \psi_{\nu_j,a,+}(t)\varphi_j(\hat{x}) \tag{9.8b}$$

are well defined for all $t \ge 0$ and $\hat{x} \in S^{d-1}$. Moreover, given $m \in \{0, 1\}$ and $k, l \in \{0, 1, \dots\}$, there exists a constant $C > 0$ such that

$$\|u_\infty\|_{C^l(S^{d-1})} \le C\|U(a_*, \cdot)\|_{L^2}, \tag{9.9a}$$

$$\int_0^\infty t^k \left\| \frac{\partial^m}{\partial t^m} \Psi_a(t, \cdot) \right\|_{C^l(S^{d-1})} \mathrm{d}t \le C\|U(a_*, \cdot)\|_{L^2}, \tag{9.9b}$$

and the series (9.8a) and (9.8b) converge with respect to all of these norms.

2. *The formulas (2.7) and (2.9) hold true. Equation (2.7) may be differentiated any number of times both with respect to ρ and $\hat{x}$, and integration and differentiation may be interchanged.*

Proof. Due to (6.10a) the assumption $\operatorname{Res}_{-i} \hat{U}_{j,a_*} = 0$ implies that

$$U_j(\rho) = \frac{U_j(a_*)}{\mathcal{H}_j(a_*)} \mathcal{H}_j(\rho) \tag{9.10}$$

for all $\rho \geq a_*$. It follows from Corollary 8.3 that

$$|U_j(a)| = |U_j(a_*)| \left| \frac{\mathcal{H}_j(a)}{\mathcal{H}_j(a_*)} \right| \leq C \, |U_j(a_*)| \left(\frac{a_*}{a} \right)^{\nu_j}. \tag{9.11}$$

Choose J such that $\nu_j \geq N$ for $j \geq J$, N given in Lemma 9.1. Using Corollary 9.2 and the bound (8.18) we obtain

$$\begin{aligned} &\sum_{j \geq J} |U_j(a)| \, \|\varphi_j\|_{C^l(S^{d-1})} \frac{\|t^k \psi^{(m)}_{\nu_j,a,+}\|_{L^1}}{|\mathcal{H}_j(a)|} \\ &\leq C \sum_{j \geq J} \frac{(1+\nu_j^2)^{l/2+d/4}}{(a-a_p)^{k+1-m}} \exp\left(\nu_j \Big(\sigma + \ln \frac{a}{a-a_p} + \ln \frac{a_*}{a} \Big) \right) |U_j(a_*)| \,. \end{aligned} \tag{9.12}$$

Using Lemma 6.3 for $j < J$ and applying Cauchy's inequality yields (9.9b). (Note that the L^2-norm and the right-hand side of (9.9b) could be replaced by any positive or negative Sobolev norm.) Inequality (9.9a) follows analogously from (8.18), (8.19), and (9.11).

To prove (2.7) for fixed $\hat{x} \in S^{d-1}$, we set $\psi_{a,+} = \psi_{\nu_j,a,+}$ and $U^+_\infty = \frac{U_j(a)}{\mathcal{H}_j(a)} \varphi_j(\hat{x})$ in (6.10a). Then $U(\rho)$ in (6.10a) is given by $U_j(\rho)\varphi_j(\hat{x})$ due to (9.10). Now the assertion follows by summing up over j and using (9.9). The differentiability properties of (2.7) are shown analogously using (6.10b) instead of (6.10a) and replacing $\varphi_j(\hat{x})$ by a derivative of φ_j at $\hat{x}$. Equation (2.9) follows in the same manner from (6.2a) multiplied by U^+_∞. ∎

Note that u_∞ may be interpreted as a delta peak of the cut function Ψ_a at $t = 0$. In other words, the formulas (2.7) and (2.9) remain valid if we formally replace $\Psi_a(t, \hat{x})$ by $\Psi_a(t, \hat{x}) + \delta_0(t) u_\infty(\hat{x})$ and then set $u_\infty = 0$.

THEOREM 9.4. *A bounded solution u to the differential equation* (2.1) *satisfies the Sommerfeld radiation condition* (2.5) *if and only if it satisfies the pole condition.*

Proof. Let us first assume that u satisfies the Sommerfeld radiation condition (2.5), and let U be defined by (2.2). Then

$$\frac{\partial}{\partial \rho} U(\rho, \hat{x}) - \left(i + \frac{d-1}{2\rho} \right) U(\rho, \hat{x}) = \mathrm{o}(1), \qquad \rho \to \infty, \tag{9.13}$$

uniformly for $\hat{x} \in S^{d-1}$. Therefore, the Fourier coefficients $U_j(\rho) := \langle U(\rho, \cdot), \varphi_j \rangle$ satisfy

$$U_j'(\rho) - \left(i + \frac{d-1}{2\rho} \right) U_j(\rho) = \mathrm{o}(1), \qquad \rho \to \infty.$$

By virtue of Theorem 7.1 this is equivalent to $\operatorname{Res}_{-i} \hat{U}_{j,a} = 0$. It follows that $U_j(\rho) = U_j(a)\mathcal{H}_j(\rho)/\mathcal{H}_j(a)$. Here a is chosen such that the assumption of Theorem 9.3 is satisfied. A comparison of Fourier coefficients shows that $\hat{U}_a(s, \hat{x}) = \sum_j \frac{U_j(a)}{\mathcal{H}_j(a)} \varphi_j(\hat{x}) \hat{\mathcal{H}}_{j,a}(s)$

for $\operatorname{Re} s > 0$ and $\hat{x} \in S^{d-1}$. We claim that for $\hat{x} \in S^{d-1}$ a holomorphic extension of $\hat{U}_a(\cdot, \hat{x})$ to $\mathbb{C}\backslash S_i$ is given by the function

$$s \mapsto -e^{-ia} \left(\frac{u_\infty(\hat{x})}{i-s} + \int_0^\infty \frac{\Psi_a(t,\hat{x})}{i-t-s} \,\mathrm{d}t \right). \tag{9.14}$$

Due to the estimates (9.9) this function is well defined and holomorphic in $\mathbb{C}\backslash S_i$. To show that it coincides with $\hat{U}_a(s,\hat{x})$ for $\operatorname{Re} s > 0$, we use Proposition 6.6 and the identity $\operatorname{Res}_i \hat{\mathcal{H}}_{j,a} = e^{-ia}$, which follows from (6.7) and the definition of $\mathcal{H}_j$. The boundedness of $s \mapsto \int_{S^{d-1}} |\frac{\partial}{\partial s} \hat{U}_a(s,\hat{x})|^2 \,\mathrm{d}\hat{x}$ follows from (9.9b) using Cauchy's inequality.

Now assume that u satisfies the pole condition, and let $\hat{U}_a(\cdot,\hat{x})$ be defined by (2.3). Using a standard corollary to Lebesgue's dominated convergence theorem and the boundedness assumption in the pole condition, it follows that the Fourier coefficients $\hat{U}_{j,a}(s) := \langle \hat{U}_a(s,\cdot), \varphi_j \rangle$ satisfy $\operatorname{Res}_{-i} \hat{U}_{j,a} = 0$. Differentiating (2.7) once and using a partial integration we get

$$\begin{aligned} \frac{\partial}{\partial \rho} U(\rho,\hat{x}) - iU(\rho,\hat{x}) &= -\int_0^\infty e^{-t(\rho-a)} t \Psi_a(t,\hat{x}) \,\mathrm{d}t \\ &= -\frac{1}{\rho-a} \int_0^\infty e^{-t(\rho-a)} \frac{\partial}{\partial t} \left\{ t\Psi_a(t,\hat{x}) \right\} \,\mathrm{d}t. \end{aligned}$$

By virtue of (9.9b), the integral term on the right-hand side of this equation is uniformly bounded for $\hat{x} \in S^{d-1}$. Since U is also uniformly bounded, this implies (9.13), which is equivalent to (2.5). □

We mention that (2.6) holds true with u_∞ and Ψ_a defined by (9.8). This follows from the Sokhotski–Plemelj formula (cf. [6]) and the fact that the function (9.14) coincides with $\hat{U}_a(\cdot,\hat{x})$.

We have constructed a solution (u_∞, Ψ_a) to the system (2.9), (2.10) if $f(\hat{x}) = U(a,\hat{x})$ and if the assumptions of Theorem 9.3 are satisfied. Uniqueness of this solution follows from the uniqueness of the corresponding system for each Fourier mode.

Appendix. Proof of Lemma 8.2. We may assume w.l.o.g. that $\Delta_2(\rho+i\sigma) \neq 0$ for all $\rho \in [a,A]$ and $\sigma \in \mathbb{R}$. Otherwise, if $\Delta_2(\rho_0+i\sigma_0) = 0$, then $\Delta_2(\rho_0+i\sigma;\nu) = 0$ for all $\sigma \in \mathbb{R}$ due to the uniqueness of initial value problems for (8.12), and then (8.14) is trivially satisfied for $\rho = \rho_0$. Our proof is based on the observation that the function $\sigma \mapsto |\Delta_2(\rho+i\sigma;\nu)|$ is decreasing at the point σ if and only if $\partial_\sigma(|\Delta_2(\rho+i\sigma;\nu)|^2) \leq 0$, if and only if $\operatorname{Re}(\dot{\Delta}_2/\Delta_2)(\rho+i\sigma;\nu) \leq 0$ (divide by $|\Delta_2(\rho+i\sigma;\nu)|^2$). Due to (8.12), $\sigma \mapsto |\Delta_2(\rho+i\sigma;\nu)|$ is decreasing at σ if and only if $\Delta_2(\rho+i\sigma;\nu) \in G(\rho+i\sigma;\nu)$, where

$$G(z;\nu) := \left\{ \delta \in \mathbb{C} : \operatorname{Re}\left[-\delta + 2\gamma_2(z;\nu) + \frac{\dot{\gamma}_2(z;\nu) - \dot{\gamma}_1(z;\nu)}{\delta} \right] \leq 0 \right\}.$$

Introducing the variable x for the expression in brackets and solving a quadratic equation for δ shows that $G = G^+ \cup G^-$ with

$$G^\pm := \left\{ \left(\gamma_2 - \frac{x}{2}\right) \left(1 \pm \sqrt{1 + \frac{\dot{\gamma}_2 - \dot{\gamma}_1}{(\gamma_2 - x/2)^2}} \right) : \operatorname{Re} x \leq 0 \right\}.$$

For the following arguments we introduce the strips $S_\lambda := \{\rho+i\sigma : a \leq \rho \leq A, \sigma \geq \lambda\}$ ($\lambda \geq 0$) in the complex plane. Note that there exist constants $C, N > 0$ such that

$$\frac{1}{|z\gamma_1|} = \frac{1}{\nu |\sqrt{\nu^{-2} z^2 (1+p) - 1}|} = \begin{cases} \frac{C}{\nu}, & \sigma \leq \nu, \\ \frac{C}{\sigma}, & \sigma > \nu \end{cases} \tag{A.1}$$

for all $z \in S_0$ and all $\nu \geq N$ and that

$$\dot\gamma_1 = i\frac{\nu^2}{z^3\gamma_1}\left(1+\frac{z^3\dot p}{2\nu^2}\right). \tag{A.2}$$

As $\gamma_2^2 - 1 = \gamma_1^2 - 1 - \dot\gamma_1$, (8.15) follows from (A.1) and (A.2). Using this and $\gamma_2 = \frac{i\nu}{z}\sqrt{1-\nu^{-2}z^2+\mathcal{O}(\nu^{-1})}$, it can be shown that there exist constants $C, N > 0$ such that

$$|z|\operatorname{Re}\gamma_2(z) \geq \begin{cases} C\nu\sigma, & 0 \leq \sigma < 1, \\ C\nu, & 1 \leq \sigma < \nu, \\ C\sigma, & \nu \leq \sigma \end{cases} \tag{A.3}$$

for all $\nu \geq N$ and all $z \in S_0$. As $\gamma_2 = \gamma_1\sqrt{1+\frac{\dot\gamma_1}{\gamma_1^2}}$, we have

$$\dot\gamma_2 - \dot\gamma_1 = \dot\gamma_1\left(\sqrt{1+\frac{\dot\gamma_1}{\gamma_1^2}}-1\right) + \left(1+\frac{\dot\gamma_1}{\gamma_1^2}\right)^{-1/2}\left(\frac{\ddot\gamma_1}{\gamma_1} - 2\frac{\dot\gamma_1^2}{\gamma_1^2}\right).$$

Since $\frac{\dot\gamma_1}{\gamma_1^2} = \mathcal{O}\big(\frac{1}{\nu}\big)$ uniformly for $z \in S_0$ due to (A.1) and (A.2), we have

$$\dot\gamma_1\left(\sqrt{1+\frac{\dot\gamma_1}{\gamma_1^2}}-1\right) = \frac{1}{2}\left(\frac{\dot\gamma_1}{\gamma_1}\right)^2 + \mathcal{O}\left(\frac{\dot\gamma_1^3}{\gamma_1^4}\right) = \mathcal{O}\big(|z|^{-2}\big).$$

Moreover,

$$\frac{\ddot\gamma_1}{\gamma_1} = \frac{3\nu^2}{z^2(z\gamma_1)^2}\left(1+\frac{z^4\ddot p}{\nu^2}\right) - \frac{2i}{z^2}\left(\frac{\nu\dot\gamma_1}{\gamma_1^2}\right)\left(\frac{\nu}{z\gamma_1}\right)\left(1+\frac{z^3\dot p}{2i\nu^2}\right) = \frac{\mathcal{O}(1)}{z^2}$$

uniformly for $z \in S_0$, so

$$|\dot\gamma_2 - \dot\gamma_1| = \mathcal{O}\big(|z|^{-2}\big) \tag{A.4}$$

uniformly for $z \in S_0$. Hence,

$$\left|\frac{\dot\gamma_2-\dot\gamma_1}{(\gamma_2 - x/2)^2}\right| = \mathcal{O}\left(\frac{1}{|z|^2(\operatorname{Re}\gamma_2)^2}\right) = \begin{cases} \mathcal{O}\big((\nu\sigma)^{-2}\big), & 0 \leq \sigma < 1, \\ \mathcal{O}\big(\nu^{-2}\big), & 1 \leq \sigma < \nu, \\ \mathcal{O}\big(\sigma^{-2}\big), & \nu \leq \sigma \end{cases}$$

uniformly for $z \in S_0$ and $\operatorname{Re} x \leq 0$ (cf. (A.3)). Now the Taylor formula $\sqrt{1+\epsilon} = 1 + \epsilon/2 + \mathcal{O}\big(\epsilon^2\big)$ $(\epsilon \to 0)$ implies that there exist constants $\Gamma, N > 0$ such that

$$\begin{aligned} &\left|\left(\gamma_2 - \frac{x}{2}\right)\left(1-\sqrt{1+\frac{\dot\gamma_2-\dot\gamma_1}{(\gamma_2-x/2)^2}}\right)\right| \\ &= \left|\left(\gamma_2-\frac{x}{2}\right)\left(1-1-\frac{1}{2}\frac{\dot\gamma_2-\dot\gamma_1}{(\gamma_2-x/2)^2} + \mathcal{O}\left(\left|\frac{\dot\gamma_2-\dot\gamma_1}{(\gamma_2-x/2)^2}\right|^2\right)\right)\right| \\ &\leq \frac{1}{2}\left|\frac{\dot\gamma_2-\dot\gamma_1}{\operatorname{Re}\gamma_2}\right| + \mathcal{O}\left(\frac{|\dot\gamma_2-\dot\gamma_1|}{|\operatorname{Re}\gamma_2|^3}\right) \leq \begin{cases} \Gamma(\sigma\nu)^{-1}, & \Gamma/\nu \leq \sigma < \nu, \\ \Gamma/\sigma^2, & \nu \leq \sigma \end{cases} \end{aligned}$$

for all $\nu \geq N$, $z \in S_{\Gamma/\nu}$, and $\operatorname{Re} x \geq 0$. Performing an analogous computation for $G^+(z;\nu)$, we obtain that for $\nu \geq N$

$$G^-(z;\nu) \subset \begin{cases} \{\zeta : |\zeta| \leq \Gamma/(\sigma\nu)\}, & \Gamma/\nu \leq \sigma < \nu, \\ \{\zeta : |\zeta| \leq \Gamma/\sigma^2\}, & \nu \leq \sigma, \end{cases} \tag{A.5a}$$

$$G^+(z;\nu) \subset \{\zeta : \operatorname{Re}\zeta \geq 1\}, \qquad \Gamma/\nu \leq \sigma. \tag{A.5b}$$

Now we are going to show that (A.5) and (8.9) imply (8.14) for $z \in S_{\Gamma/\nu}$. Let $\rho_0 \in [a, A]$. By virtue of (8.9) and the fact that $\lim_{\sigma\to\infty}(1 - \gamma_2(\rho_0 + i\sigma;\nu)) = 0$ for all ν, there exists a sequence $\Gamma/\nu = \sigma_0 < \sigma_1 < \cdots$ such that $\lim_{l\to\infty}\sigma_l = \infty$ and $|\Delta_2(\rho_0+i\sigma_l;\nu)| < 1/(l+1)$ for $l \geq 1$. We may also arrange that $\partial_\sigma|\Delta_2(\rho_0+i\sigma_l;\nu)| < 0$. Then the maximum of the function $\sigma \mapsto |\Delta_2(\rho_0 + i\sigma;\nu)|$ on the interval $[\sigma_l, \sigma_{l+1}]$ is attained at the point $\sigma_l^* \in [\sigma_l, \sigma_{l+1})$, and $\partial_\sigma|\Delta_2(\rho_0+i\sigma_l^*;\nu)| \leq 0$, i.e., $\Delta(\rho_0+i\sigma_l^*;\nu) \in G(\rho_0+i\sigma_l^*;\nu)$. If $\Delta(\rho_0+i\sigma_0;\nu) \in G^+(\rho_0+i\sigma_0;\nu)$, then, due to (A.5b) and the choice of the σ_l's, there exists a largest $\tilde\sigma \in (\sigma_l^*, \sigma_{l+1})$ such that $|\Delta_2(\rho_0 + i\tilde\sigma;\nu)| = \frac{1}{2}$, and $\Delta_2(\rho_0 + i\tilde\sigma;\nu) \in G(\rho_0 + i\sigma_*;\nu)$ since $\partial_\sigma|\Delta(\rho_0 + i\tilde\sigma;\nu)| \leq 0$. This contradicts (A.5). Hence, $\Delta(\rho_0 + i\sigma_l^*;\nu) \in G^-(\rho_0 + i\sigma_l^*;\nu)$, and (8.14) follows from (A.5a).

It remains to show (8.14) for $0 \leq \sigma \leq \Gamma/\nu$. In this case, there exist constants $C, N > 0$ such that

$$\operatorname{Re}\frac{\dot\Delta_2}{\Delta_2} = -\operatorname{Re}\Delta_2 + 2\operatorname{Re}\gamma_2 + \operatorname{Re}\frac{\dot\gamma_2 - \dot\gamma_1}{\Delta_2} \geq -C$$

if $\Delta_2(z;\nu)$ is in the annulus $1 \leq |\Delta_2(z;\nu)| \leq 2$ and $\nu \geq N$ (cf. (A.3) and (A.4)). Since

$$|\Delta_2(\rho + i\sigma;\nu)| = \left|\Delta_2\left(\rho + i\frac{\Gamma}{\nu};\nu\right)\right| \exp\left(\int_{\Gamma/\nu}^{\sigma} \operatorname{Re}\frac{\dot\Delta_2(\rho + i\sigma_1;\nu)}{\Delta_2(\rho + i\sigma_1;\nu)}\,\mathrm{d}\sigma_1\right)$$

and since $|\Delta_2(\rho + i\Gamma/\nu;\nu)| \leq 1$, it follows that $|\Delta_2(\rho+i\sigma;\nu)| \leq \exp(C(\Gamma/\nu - \sigma)) \leq 2$ for $0 \leq \sigma \leq \Gamma/\nu$ and $\nu \geq \max(N, C\Gamma/\ln 2)$. □

REFERENCES

[1] J. P. Bérenger, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 114 (1994), pp. 185–200.

[2] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering*, 2nd ed., Springer-Verlag, Berlin, Heidelberg, New York, 1997.

[3] L. Demkowicz and K. Gerdes, *Convergence of the infinite element methods for the Helmholtz equation in separable domains*, Numer. Math., 79 (1998), pp. 11–42.

[4] P. Deuflhard and F. Bornemann, *Scientific Computing with Ordinary Differential Equations*, Springer-Verlag, New York, 2002.

[5] G. Doetsch, *Handbuch der Laplace-Transformation*, Vol. 2, Birkhäuser-Verlag, Basel, Stuttgart, 1955.

[6] H. W. Engl, *Integralgleichungen*, Springer-Verlag, Vienna, 1997.

[7] B. Enquist and A. Majda, *Radiation boundary conditions for acoustic and elastic wave calculations*, Comm. Pure Appl. Math., 32 (1979), pp. 313–357.

[8] D. Givoli, *Numerical Methods for Problems in Infinite Domains*, Stud. Appl. Mech. 33, Elsevier, Amsterdam, 1992.

[9] T. Hohage, F. Schmidt, and L. Zschiedrich, *Solving time-harmonic scattering problems based on the pole condition.* II: *Convergence of the PML method*, SIAM J. Math. Anal., to appear.

[10] T. Hohage, F. Schmidt, and L. Zschiedrich, *Solving time-harmonic scattering problems based on the pole condition.* III: *Numerical algorithms*, in preparation.

[11] T. Hohage, F. Schmidt, and L. Zschiedrich, *A new method for the solution of scattering problems*, in Proceedings of the JEE'02 Symposium, Toulouse, France, 2002, B. Michielsen and F. Decavèle, eds., ONERA, Toulouse, France, pp. 251–256.

[12] F. Ihlenburg, *Finite Element Analysis of Acoustic Scattering*, Springer-Verlag, New York, 1998.
[13] R. Kress, *Linear Integral Equations*, Springer-Verlag, Berlin, Heidelberg, New York, 1989.
[14] F. Schmidt, *An Alternative Derivation of the Exact DtN-Map on a Circle*, Tech. Report SC 98-32, Zuse Institute Berlin, Berlin, 1998.
[15] F. Schmidt, *Construction of discrete transparent boundary conditions for Schrödinger-type equations*, Surveys Math. Indust., 9 (1999), pp. 87–100.
[16] F. Schmidt, *Solution of Interior-Exterior Helmholtz-Type Problems Based on the Pole Condition: Theory and Algorithms*, Habilitation thesis, Freie Universität Berlin, Berlin, 2001.
[17] F. Schmidt and P. Deuflhard, *Discrete transparent boundary conditions for the numerical solution of Fresnel's equation*, Comput. Math. Appl., 29 (1995), pp. 53–76.
[18] F. Schmidt, T. Friese, and D. Yevick, *Transparent boundary conditions for split-step Padé approximations of the one-way Helmholtz equation*, J. Comput. Phys., 168 (2000), pp. 696–719.
[19] A. Sommerfeld, *Die Greensche Funktion der Schwingungsgleichung*, Jahresber. Deutsch. Math.-Verein, 21 (1912), pp. 309–353.
[20] M. Taylor, *Partial Differential Equations*. I. *Basic Theory*, Springer-Verlag, New York, 1996.
[21] G. N. Watson, *Theory of Bessel Functions*, Cambridge University Press, London, 1922.
[22] B. Zhang and S. Chandler-Wilde, *Acoustic scattering by an inhomogeneous layer on a rigid plate*, SIAM J. Appl. Math., 58 (1998), pp. 1931–1950.

SIAM J. MATH. ANAL.
Vol. 35, No. 1, pp. 211–244

WELL-POSEDNESS OF VORTEX SHEETS WITH SURFACE TENSION*

DAVID M. AMBROSE†

Abstract. We study the initial value problem for two-dimensional, periodic vortex sheets with surface tension. We allow the upper and lower fluids to have different densities. Without surface tension, the vortex sheet is ill-posed: it exhibits the well-known Kelvin–Helmholtz instability. In the linearized equations of motion, surface tension removes the instability. It has been conjectured that surface tension also makes the full problem well-posed. We prove that this conjecture is correct using energy methods. In particular, for the initial value problem for vortex sheets with surface tension with sufficiently smooth data, it is proved that solutions exist locally in time, are unique, and depend continuously on the initial data. The analysis uses two important ideas from the numerical work of Hou, Lowengrub, and Shelley. First, the tangent angle and arclength of the vortex sheet are used rather than Cartesian variables. Second, instead of a purely Lagrangian formulation, a special tangential velocity is used in order to simplify the evolution equations. A special case of the result is well-posedness of water waves with surface tension; this is the first proof (with surface tension) which allows the wave to overturn.

Key words. vortex sheet, surface tension, Kelvin–Helmholtz instability

AMS subject classifications. 76B03, 76T99, 35Q35

DOI. 10.1137/S0036141002403869

1. Introduction.

The classical vortex sheet is the interface between two incompressible, inviscid, irrotational, density-matched two-dimensional fluids moving past each other, neglecting surface tension. In this situation, all of the vorticity is concentrated on the interface. At each time, the sheet can be viewed as a curve in the complex plane. The curve, z, is parameterized by a Lagrangian spatial variable, α. The curve evolves according to the Birkhoff–Rott integral (see page 141 of [Saf95]),

$$z_t^*(\alpha,t) = \frac{1}{2\pi i}\mathrm{PV}\int_{-\infty}^{\infty} \frac{\gamma(\alpha')}{z(\alpha,t)-z(\alpha',t)}\,d\alpha'. \tag{1.1}$$

The $*$ denotes complex conjugation; γ is the vortex sheet strength. Notice that γ is not a function of time. This problem has been well studied and has been found to be ill-posed in the usual sense (although it can be thought of as well-posed in analytic function spaces). In particular, it exhibits the well-known Kelvin–Helmholtz instability: in the linearization of the evolution equations about equilibrium, Fourier modes with high wave numbers grow without bound. Equation (1.1) neglects the effect of surface tension at the interface. Surface tension is a restoring force, and when surface tension is accounted for in the equations of motion, Fourier modes of high wave number remain bounded in the linearization. For this reason, it had been conjectured that surface tension makes the full problem well-posed [Bir62]. Taking this further, Beale, Hou, and Lowengrub demonstrated that even far from equilibrium, surface tension makes the linearized equations well-posed [BHL93]. Iguchi, Tanaka, and Tani have shown that the full problem is well-posed if the initial state is sufficiently close

*Received by the editors March 12, 2002; accepted for publication (in revised form) January 31, 2003; published electronically July 8, 2003. This work was partially supported by National Science Foundation grants DMS-9870091 and DMS-0102356.

http://www.siam.org/journals/sima/35-1/40386.html

†Department of Mathematics, Duke University, Durham, NC. Current address: Courant Institute of Mathematical Sciences, New York University, New York, NY 10012 (ambrose@cims.nyu.edu).

to the flat equilibrium and if the interface height is a function of horizontal position [ITT97]. The present work removes these restrictions, proving that the full problem is well-posed even for large initial data, confirming Birkhoff's conjecture.

Hou, Lowengrub, and Shelley have efficiently computed vortex sheets with surface tension by treating the part of the evolution equations related to surface tension carefully [HLS97]. They recast the evolution equations in two important ways. First, they choose to compute dependent variables which are naturally related to the surface tension. In particular, surface tension enters the evolution equations in the form $\gamma_t = \frac{1}{\mathrm{We}}\kappa_\alpha$, where κ is the curvature of the vortex sheet and We is the Weber number. The Weber number is a dimensionless parameter that is inversely proportional to the surface tension; the case without surface tension corresponds to $\mathrm{We} = \infty$. (Recall that without surface tension, $\gamma_t = 0$.) To simplify this curvature term in the evolution equations, Hou, Lowengrub, and Shelley describe the curve by its tangent angle and arclength rather than by the Cartesian variable, z. We use the notation s for arclength and θ for the tangent angle the curve forms with the horizontal. The strength of this choice of variables lies in the relationship $\kappa = \theta_\alpha / s_\alpha$.

Second, Hou, Lowengrub, and Shelley observe that while the normal velocity of the vortex sheet must be obtained from (1.1), the same is not true of the tangential velocity. By adding an arbitrary tangential velocity, T, to the evolution of z, the shape of the curve is not changed (i.e., the tangential velocity only serves to reparameterize the curve). A careful choice of T essentially reduces the problem by one dependent variable by making s_α independent of α. With this choice of T, if the curve is initially parameterized by arclength (normalized so that α is between 0 and 2π), then it will always remain that way.

The choice of tangential velocity also changes the evolution equation for γ. In addition to simplifying the curvature term, it also introduces new terms proportional to $T - \mathbf{W} \cdot \hat{\mathbf{t}}$. Here, $\mathbf{W}$ is the real form of (1.1) and $\hat{\mathbf{t}}$ is the tangent vector to the curve. In the purely Lagrangian formulation of the problem, T was equal to $\mathbf{W} \cdot \hat{\mathbf{t}}$, so these terms did not appear.

The main conclusion of this paper is that the vortex sheet with surface tension is well-posed in Sobolev spaces. In particular, this means that given periodic initial conditions $z(\cdot, 0) \in H^r$ and $\gamma(\cdot, 0) \in H^{r-3/2}$ for r large enough, there is some interval of time in which a solution to the vortex sheet evolution equations exists. This solution is unique, has the same regularity as the initial conditions, and depends continuously on the initial data. We first prove this for the case in which the upper and lower fluids have the same density, and then for the case of arbitrary densities.

In the proof, we use ideas from the numerical work of Hou, Lowengrub, and Shelley. That is, we use their tangential velocity and analyze the evolution of θ and L, the length of one period of the vortex sheet, rather than z. This simplifies the analysis considerably since many of the leading terms in the evolution equations are linear as functions of θ and γ.

The analysis uses energy methods. We first form approximations to the three evolution equations by convolving parts of the equations with approximations to the Dirac δ function. We then show that solutions to these approximated equations exist by using the Picard theorem for differential equations on Banach spaces; this requires proving that the time derivatives in the approximated equations are Lipschitz continuous.

We then define an energy function, E, for the approximate solutions. The energy function is related to the Sobolev norms of θ and γ and has no clear physical inter-

pretation. Much of the difficulty in the problem is to find the proper definition of the energy function. Finding the correct energy balance between θ and γ is very much complicated by the nonlinearities in the evolution equations. After finding the correct form of the energy, we estimate its growth and find

$$\frac{dE}{dt} \leq C_1 \exp\{C_2 E\} \tag{1.2}$$

for some positive constants C_1 and C_2.

Proving this estimate requires understanding various integral operators (including some singular integral operators), as well as estimating nonlinear terms. Also, the terms with the highest number of spatial derivatives need to be handled specially. This is similar to the numerical work. For computational reasons, Hou, Lowengrub, and Shelley introduced a small-scale decomposition (SSD) of the problem [HLS94]. That is, they identified terms which were unstable at small spatial scales and computed them by an implicit method while the remaining terms were treated explicitly. Their decomposition was

$$\theta_t = \frac{2\pi^2}{L^2} H(\gamma_\alpha) + P,$$

$$\gamma_t = \frac{2\pi}{L\mathrm{We}} \theta_{\alpha\alpha} + Q.$$

H is the Hilbert transform; P and Q represent all the terms which were computed explicitly. An important difference between this work and that of Hou, Lowengrub, and Shelley lies in the SSD. The two terms they identified as the most important for computational reasons also need to be treated carefully in the energy estimates. We identify an additional term which is significant to the analysis. Unlike the two principal terms in the SSD, this term is nonlinear as a function of θ and γ. The new term appears in the γ_t equation and is a consequence of the choice of tangential velocity. We write the new decomposition for γ_t as

$$\gamma_t = \frac{2\pi}{L\mathrm{We}} \theta_{\alpha\alpha} + \frac{2\pi^2}{L^2} \gamma H(\gamma\theta_\alpha) + \widetilde{Q}.$$

This raises an interesting question about the computing: Is there a benefit to changing the SSD in the numerical method to include this nonlinear term?

The estimate (1.2) implies that solutions of the approximate equations exist on a time interval independent of the mollification parameter, and they are uniformly bounded during that time interval. Another estimate then implies that solutions of the approximate equations form a Cauchy sequence (for different values of the mollification parameter) and thus converge to a strong solution of the original system.

A consequence of the existence proof is that development of a singularity in finite time implies that either the vortex sheet must intersect itself or the energy must blow up. These are the very kinds of singularities observed in [HLS97]. In their computations, vortex sheets developed two kinds of singularities when the surface tension was small: roll-up of the sheet and curvature singularities. Curvature singularities were also found by Siegel in an approximation to the vortex sheet with surface tension [Sie95]. (Curvature singularities correspond to the energy blowing up.) We do not address here whether or not singularities actually occur; this is a topic for future research.

After proving well-posedness in the case of fluids of the same density, we then prove the same results in the case of arbitrary densities. The proof is not significantly different from the earlier case. A special case of this result is when the upper fluid has density zero; this is the case of water waves with surface tension. Neglecting surface tension, Wu has proved well-posedness of water waves in both two and three dimensions [Wu97], [Wu99]. Iguchi has proved well-posedness of water waves with surface tension, but only when the height of the wave is a function of horizontal position [Igu01]. There is no such restriction in this paper.

In section 2, we rewrite and mollify the evolution equations for the vortex sheet with surface tension in the density-matched case. In section 3, we then state several technical lemmas we will need which demonstrate the boundedness of some useful operators. In section 4, we prove a priori estimates on the growth of solutions to the mollified initial value problem. We use these a priori estimates in section 5 to demonstrate existence and uniqueness of solutions to first the mollified problem and then the nonmollified problem. In section 6, we extend this result to the case in which the two fluids have different densities.

2. Evolution equations. In this section, we first discuss various operators that we will use throughout the paper. We then discuss in some detail the evolution equations for the exact vortex sheet with surface tension. Finally, we will apply mollifiers to the evolution equations to obtain an approximated set of evolution equations.

2.1. The Hilbert transform and associated operators. We begin with background information on some operators that we will use. The Hilbert transform, H, of a function $f : \mathbb{R} \to \mathbb{R}$ is defined by

$$Hf(\alpha) = \frac{1}{\pi} \mathrm{PV} \int_{-\infty}^{\infty} \frac{f(\alpha')}{\alpha - \alpha'} \, d\alpha'. \tag{2.1}$$

In Fourier space, H is a multiplier; that is, $\mathcal{F}(Hf)(\xi) = -i\mathrm{sgn}(\xi)\mathcal{F}f(\xi)$, where $\mathcal{F}$ is the Fourier transform. This implies that H is bounded from L^2 to L^2, and in fact $\|Hg\|_{L^2} = \|g\|_{L^2}$ and $H^2 g = -g$ whenever $\mathcal{F}g(0) = 0$. We will be concerned with functions which are periodic, and we will not want to consider the Hilbert transform as an integral over the entire real line. For 2π-periodic f we can sum over periodic images in (2.1) (see page 167 of [AF97]) to get

$$Hf(\alpha) = \frac{1}{2\pi} \mathrm{PV} \int_{0}^{2\pi} f(\alpha') \cot \frac{1}{2}(\alpha - \alpha') \, d\alpha'. \tag{2.2}$$

The Hilbert transform commutes with differentiation. We will sometimes denote differentiation by a subscript, and sometimes by application of an operator D. For example, f_α and $D_\alpha f$ both denote the derivative of f with respect to α. We define the operator Λ to be a derivative followed by the Hilbert transform: $\Lambda = HD_\alpha$. In Fourier space, we have $\mathcal{F}\Lambda f(\xi) = 2\pi|\xi|\mathcal{F}f(\xi)$. This implies that

$$\left(\int f^2 + f\Lambda f \, d\alpha \right)^{1/2}$$

is an equivalent norm to $\|f\|_{1/2}$. Also, Λ is self-adjoint.

We will also frequently use commutators of certain operators. We define $[H, f]g$ to be $H(fg) - fH(g)$. This is typically a smoothing operator; details are discussed in Lemma 3.7 and Corollary 3.8.

2.2. The exact evolution equations. Consider any curve $z(\alpha,t) = x(\alpha,t) + iy(\alpha,t)$ in the complex plane satisfying $z(\alpha+2\pi,t) = z(\alpha,t) + 2\pi$. Give the curve periodic normal velocity U and periodic tangential velocity T. Call the unit tangent and normal vectors to the curve $\hat{\mathbf{t}}$ and $\hat{\mathbf{n}}$, respectively. The tangent angle that the curve forms with the horizontal, θ, satisfies $\theta = \arctan(y_\alpha/x_\alpha)$. The derivative of arclength satisfies $s_\alpha^2 = x_\alpha^2 + y_\alpha^2$. Differentiating these two equations in time, we find that the evolution equations for θ and s_α are

$$\theta_t = \frac{1}{s_\alpha} U_\alpha + \frac{T}{s_\alpha}\theta_\alpha, \tag{2.3}$$

$$s_{\alpha t} = T_\alpha - \theta_\alpha U. \tag{2.4}$$

Also, the length of one period of the curve is $L(t) = \int_0^{2\pi} s_\alpha d\alpha$. This implies the evolution equation

$$L_t = \int_0^{2\pi} s_{\alpha t}\, d\alpha = -\int_0^{2\pi} \theta_\alpha U\, d\alpha. \tag{2.5}$$

For the vortex sheet, the normal velocity must be given by $U = \mathbf{W}\cdot\hat{\mathbf{n}}$, where $\mathbf{W}$ is the complex conjugate of the Birkhoff–Rott integral (1.1) expressed as a real vector:

$$\mathbf{W}(\alpha) = \frac{1}{2\pi}\mathrm{PV}\int \gamma(\alpha')\frac{(-(y(\alpha)-y(\alpha')), x(\alpha)-x(\alpha'))}{(x(\alpha)-x(\alpha'))^2 + (y(\alpha)-y(\alpha'))^2}\, d\alpha'.$$

In the Lagrangian formulation, $\mathbf{W} = (x_t, y_t)$.

We have some freedom in choosing T, so we use this freedom to make $s_{\alpha t}$ independent of α. Following [HLS94], we set $s_{\alpha t}(\alpha,t) = L_t(t)/2\pi$ in (2.4). Integrating to solve for T, we have

$$T(\alpha,t) = \int_0^\alpha \theta_{\alpha'} U \mathrm{d}\alpha' + \frac{\alpha}{2\pi} L_t. \tag{2.6}$$

We have set $T(0,t) = 0$ for simplicity. Now, if the curve is initially parameterized so that $s_\alpha = L/2\pi$, it will remain that way.

The time derivative of the vortex sheet strength can be found by combining Bernoulli's equation for potential flow with the Laplace–Young condition for the pressure jump across the interface (the pressure jump is proportional to curvature). The derivation is a more general form of the derivation in [BMO82]; details can be found in [Amb02]. The resulting evolution equation is

$$\gamma_t = \frac{1}{\mathrm{We}} D_\alpha\left(\frac{\theta_\alpha}{s_\alpha}\right) + D_\alpha\left(\frac{(T - \mathbf{W}\cdot\hat{\mathbf{t}})\gamma}{s_\alpha}\right). \tag{2.7}$$

Before attempting any analysis, we rewrite the evolution equations. We will need the geometric identity

$$\hat{\mathbf{t}}_\alpha \cdot \hat{\mathbf{n}} = \theta_\alpha. \tag{2.8}$$

Also, the following calculation is useful many times:

$$\begin{aligned} D_\alpha(T - \mathbf{W}\cdot\hat{\mathbf{t}}) &= T_\alpha - \mathbf{W}_\alpha\cdot\hat{\mathbf{t}} - \mathbf{W}\cdot\hat{\mathbf{t}}_\alpha \\ &= \theta_\alpha U + \frac{L_t}{2\pi} - \mathbf{W}_\alpha\cdot\hat{\mathbf{t}} - (\mathbf{W}\cdot\hat{\mathbf{n}})(\hat{\mathbf{t}}_\alpha\cdot\hat{\mathbf{n}}) = \frac{L_t}{2\pi} - \mathbf{W}_\alpha\cdot\hat{\mathbf{t}}. \end{aligned} \tag{2.9}$$

We have used (2.8) to see an important cancellation between the two θ_α terms. We will see soon that the $\mathbf{W}_\alpha \cdot \hat{\mathbf{t}}$ term which remains is similar to the Hilbert transform of θ_α, so the smoothness of the terms that canceled is not exactly the issue. The presence of the Hilbert transform in the term that remains is critical.

We are ready to begin rewriting the system of evolution equations (2.3), (2.5), (2.7). We expand U_α in (2.3):

$$\theta_t = \frac{2\pi}{L}(\mathbf{W}\cdot\hat{\mathbf{n}})_\alpha + \frac{2\pi}{L}T\theta_\alpha = \frac{2\pi}{L}(\mathbf{W}_\alpha\cdot\hat{\mathbf{n}}) + \frac{2\pi}{L}(T - \mathbf{W}\cdot\hat{\mathbf{t}})\theta_\alpha. \tag{2.10}$$

We now rewrite γ_t, using $s_\alpha = L/2\pi$ and (2.9):

$$\gamma_t = \frac{2\pi}{L}\frac{\theta_{\alpha\alpha}}{\mathrm{We}} + \frac{2\pi}{L}\left(\frac{L_t}{2\pi} - \mathbf{W}_\alpha\cdot\hat{\mathbf{t}}\right)\gamma + \frac{2\pi}{L}(T - \mathbf{W}\cdot\hat{\mathbf{t}})\gamma_\alpha. \tag{2.11}$$

We finish rewriting θ_t and γ_t by simplifying $\mathbf{W}_\alpha\cdot\hat{\mathbf{n}}$ and $\mathbf{W}_\alpha\cdot\hat{\mathbf{t}}$. To do this, we switch from real to complex notation. Define $\Phi : \mathbb{R}^2 \to \mathbb{C}$ to be the mapping $(a,b)\mapsto a+ib$. We denote by $z = x + iy$ the image under this mapping of our dependent variables (x, y). The following are then true:

$$\Phi(\hat{\mathbf{t}}) = \frac{2\pi z_\alpha}{L}, \qquad \Phi(\hat{\mathbf{n}}) = \frac{2\pi i z_\alpha}{L}. \tag{2.12}$$

Furthermore, the formula for a dot product is

$$\mathbf{a}\cdot\mathbf{b} = \mathrm{Re}\left(\Phi(\mathbf{a})\Phi(\mathbf{b})^*\right). \tag{2.13}$$

Under this mapping, we have for $\mathbf{W}$

$$\Phi(\mathbf{W})^* = \frac{1}{2\pi i}\mathrm{PV}\int \frac{\gamma(\alpha')}{z(\alpha) - z(\alpha')}\, d\alpha'. \tag{2.14}$$

We rewrite this slightly:

$$\Phi(\mathbf{W})^* = \frac{1}{2\pi i}\mathrm{PV}\int \left(\frac{\gamma(\alpha')}{z_\alpha(\alpha')}\right)\frac{z_\alpha(\alpha')}{z(\alpha) - z(\alpha')}\, d\alpha'.$$

We are interested in non-self-intersecting curves such that $z(\alpha) - \alpha$ is 2π-periodic, which implies

$$\frac{\gamma(\alpha)}{z_\alpha(\alpha)}\mathrm{PV}\int \frac{z_\alpha(\alpha')}{z(\alpha) - z(\alpha')}\, d\alpha' = 0.$$

(This can be seen by taking a branch of the logarithm.) Subtracting this from $\Phi(\mathbf{W})^*$ and taking a derivative, we get

$$\begin{aligned}
\Phi(\mathbf{W})^*_\alpha &= \frac{1}{2\pi i}D_\alpha\mathrm{PV}\int \left(\frac{\gamma(\alpha')}{z_\alpha(\alpha')} - \frac{\gamma(\alpha)}{z_\alpha(\alpha)}\right)\frac{z_\alpha(\alpha')}{z(\alpha) - z(\alpha')}\, d\alpha' \\
&= \frac{1}{2\pi i}\mathrm{PV}\int \left(\frac{\gamma(\alpha')}{z_\alpha(\alpha')} - \frac{\gamma(\alpha)}{z_\alpha(\alpha)}\right)\frac{-z_\alpha(\alpha')z_\alpha(\alpha)}{(z(\alpha) - z(\alpha'))^2}\, d\alpha' \\
&= \frac{-z_\alpha(\alpha)}{2\pi i}\mathrm{PV}\int \left(\frac{\gamma(\alpha')}{z_\alpha(\alpha')} - \frac{\gamma(\alpha)}{z_\alpha(\alpha)}\right) D_{\alpha'}\left(\frac{1}{z(\alpha) - z(\alpha')}\right)\, d\alpha'.
\end{aligned}$$

We integrate this by parts to get

$$\Phi(\mathbf{W})^*_\alpha(\alpha) = \frac{z_\alpha(\alpha)}{2\pi i}\mathrm{PV}\int\left(\frac{\gamma_\alpha(\alpha')}{z_\alpha(\alpha')} - \frac{\gamma(\alpha')z_{\alpha\alpha}(\alpha')}{z^2_\alpha(\alpha')}\right)\frac{1}{z(\alpha)-z(\alpha')}\,d\alpha'. \tag{2.15}$$

We approximate $z(\alpha)-z(\alpha')$ by the leading term from its Taylor series, $z_\alpha(\alpha')(\alpha-\alpha')$, to rewrite (2.15) as $\Phi(\mathbf{A}_1)^* + \Phi(\mathbf{R}_1)^* + \Phi(\mathbf{A}_2)^* + \Phi(\mathbf{R}_2)^*$, where

$$\Phi(\mathbf{A}_1)^* = \frac{z_\alpha(\alpha)}{2\pi i}\mathrm{PV}\int\frac{\gamma_\alpha(\alpha')}{z_\alpha(\alpha')}\left[\frac{1}{z_\alpha(\alpha')(\alpha-\alpha')}\right]\,d\alpha', \tag{2.16}$$

$$\Phi(\mathbf{A}_2)^* = -\frac{z_\alpha(\alpha)}{2\pi i}\mathrm{PV}\int\frac{\gamma(\alpha')z_{\alpha\alpha}(\alpha')}{z^2_\alpha(\alpha')}\left[\frac{1}{z_\alpha(\alpha')(\alpha-\alpha')}\right]\,d\alpha'. \tag{2.17}$$

The remainders from approximating $z(\alpha)-z(\alpha')$ are

$$\Phi(\mathbf{R}_1)^* = \frac{z_\alpha(\alpha)}{2\pi i}\mathrm{PV}\int\frac{\gamma_\alpha(\alpha')}{z_\alpha(\alpha')}\left[\frac{1}{z(\alpha)-z(\alpha')} - \frac{1}{z_\alpha(\alpha')(\alpha-\alpha')}\right]\,d\alpha', \tag{2.18}$$

$$\Phi(\mathbf{R}_2)^* = -\frac{z_\alpha(\alpha)}{2\pi i}\mathrm{PV}\int\frac{\gamma(\alpha')z_{\alpha\alpha}(\alpha')}{z^2_\alpha(\alpha')}\left[\frac{1}{z(\alpha)-z(\alpha')} - \frac{1}{z_\alpha(\alpha')(\alpha-\alpha')}\right]\,d\alpha'. \tag{2.19}$$

We rewrite $\mathbf{A}_1$ to show that it equals a smooth term plus a scalar multiple of $\hat{\mathbf{n}}$.

$$\Phi(\mathbf{A}_1)^* = \frac{z_\alpha}{2i}H\left(\frac{\gamma_\alpha}{z^2_\alpha}\right) = \frac{1}{2iz_\alpha}H(\gamma_\alpha) + \frac{z_\alpha}{2i}\left[H,\frac{1}{z^2_\alpha}\right](\gamma_\alpha).$$

We can write this as

$$\Phi(\mathbf{A}_1)^* = \frac{\pi}{L}\left(\frac{2\pi i z_\alpha}{L}\right)^* H(\gamma_\alpha) + \frac{z_\alpha}{2i}\left[H,\frac{1}{z^2_\alpha}\right](\gamma_\alpha).$$

We define $\mathbf{B}_1$ to be Φ^{-1} of the conjugate of the second term, and we use (2.12) to see that $\mathbf{A}_1 = \frac{\pi}{L}H(\gamma_\alpha)\hat{\mathbf{n}} + \mathbf{B}_1$. Similarly, we compute

$$\Phi(\mathbf{A}_2)^* = -\frac{z_\alpha}{2i}H\left(\frac{\gamma z_{\alpha\alpha}}{z^3_\alpha}\right) = -\frac{1}{2iz_\alpha}H\left(\frac{\gamma z_{\alpha\alpha}}{z_\alpha}\right) - \frac{z_\alpha}{2i}\left[H,\frac{1}{z^2_\alpha}\right]\left(\frac{\gamma z_{\alpha\alpha}}{z_\alpha}\right).$$

We look at $\mathbf{A}_2$ separately when taking the inner product with $\hat{\mathbf{t}}$ and $\hat{\mathbf{n}}$. We define $\mathbf{B}_2$ in the analogous way to $\mathbf{B}_1$. Using (2.12) and (2.13), we see that

$$\mathbf{A}_2\cdot\hat{\mathbf{t}} - \mathbf{B}_2\cdot\hat{\mathbf{t}} = \mathrm{Re}\left(-\frac{\pi}{iL}H\left(\frac{\gamma z_{\alpha\alpha}}{z_\alpha}\right)\right) = -\frac{\pi}{L}H\left(\gamma\mathrm{Re}\left\{\frac{2\pi z_{\alpha\alpha}}{L}\left(\frac{2\pi i z_\alpha}{L}\right)^*\right\}\right).$$

Using (2.12), (2.13), and (2.8), we see that this simplifies. In particular,

$$\mathrm{Re}\left\{\frac{2\pi z_{\alpha\alpha}}{L}\left(\frac{2\pi i z_\alpha}{L}\right)^*\right\} = \hat{\mathbf{t}}_\alpha\cdot\hat{\mathbf{n}} = \theta_\alpha.$$

This implies $\mathbf{A}_2\cdot\hat{\mathbf{t}} = -\frac{\pi}{L}H(\gamma\theta_\alpha) + \mathbf{B}_2\cdot\hat{\mathbf{t}}$. Similarly,

$$\mathbf{A}_2\cdot\hat{\mathbf{n}} = -\frac{\pi}{L}H\left(\gamma\hat{\mathbf{t}}_\alpha\cdot\hat{\mathbf{t}}\right) + \mathbf{B}_2\cdot\hat{\mathbf{n}}.$$

Since $\hat{\mathbf{t}}$ is a unit vector, $\hat{\mathbf{t}}_\alpha \cdot \hat{\mathbf{t}} = 0$, so $\mathbf{A}_2 \cdot \hat{\mathbf{n}} = \mathbf{B}_2 \cdot \hat{\mathbf{n}}$. We can now write

$$\mathbf{W}_\alpha \cdot \hat{\mathbf{n}} = \frac{\pi}{L} H(\gamma_\alpha) + \mathbf{m} \cdot \hat{\mathbf{n}}, \qquad \mathbf{W}_\alpha \cdot \hat{\mathbf{t}} = -\frac{\pi}{L} H(\gamma\theta_\alpha) + \mathbf{m} \cdot \hat{\mathbf{t}}, \tag{2.20}$$

where

$$\mathbf{m} = \mathbf{R}_1 + \mathbf{R}_2 + \mathbf{B}_1 + \mathbf{B}_2. \tag{2.21}$$

Finally, we can restate our initial value problem as

$$\theta_t = \boxed{\frac{2\pi^2}{L^2} H(\gamma_\alpha)} + \frac{2\pi}{L}(T - \mathbf{W} \cdot \hat{\mathbf{t}})\theta_\alpha + \frac{2\pi \mathbf{m} \cdot \hat{\mathbf{n}}}{L}, \tag{2.22}$$

$$\gamma_t = \boxed{\frac{2\pi}{L}\frac{\theta_{\alpha\alpha}}{\mathrm{We}} + \frac{2\pi^2 \gamma}{L^2} H(\gamma\theta_\alpha)} + \frac{2\pi}{L}(T - \mathbf{W} \cdot \hat{\mathbf{t}})\gamma_\alpha + \frac{\gamma(L_t - 2\pi \mathbf{m} \cdot \hat{\mathbf{t}})}{L}, \tag{2.23}$$

$$L_t = -\int_0^{2\pi} \theta_\alpha U \, d\alpha \tag{2.24}$$

subject to the initial conditions

$$\theta(\alpha, 0) = \theta_0(\alpha), \quad \gamma(\alpha, 0) = \gamma_0(\alpha), \quad L(0) = L_0. \tag{2.25}$$

The boxed terms are those which will need to be treated most carefully in the energy estimates.

2.3. The mollified equations. Our strategy for proving existence of solutions for the vortex sheet is as follows: First, we introduce mollifiers to the right-hand sides of the evolution equations. The purpose of this is to turn the right-hand sides into bounded functions of θ, γ, and L in certain Sobolev spaces. Thus, we need to include mollifiers only on terms which feature derivatives of θ and γ. The presence of the mollifiers will allow us to prove existence of solutions to the mollified equations for an amount of time which depends on the mollification parameter. We will need to know bounds on the mollified solutions which are independent of the mollification parameter, ε, in order to establish existence of solutions on a time interval independent of ε. We will then be able to pass to the limit as ε goes to zero and establish existence of solutions to the nonmollified equations.

We now state the mollified equations; we will define the new terms involved in these equations next.

$$\theta_t^\varepsilon = \boxed{\frac{2\pi^2}{L^{\varepsilon 2}} H\chi^\varepsilon(\gamma_\alpha^\varepsilon)} + \frac{2\pi}{L^\varepsilon}\chi^\varepsilon\left((T^\varepsilon - \mathbf{W}^\varepsilon \cdot \hat{\mathbf{t}}^\varepsilon)\chi^\varepsilon\theta_\alpha^\varepsilon\right) + \frac{2\pi \mathbf{m}^\varepsilon \cdot \hat{\mathbf{n}}^\varepsilon}{L^\varepsilon} + \mu^\varepsilon, \tag{2.26}$$

$$\gamma_t^\varepsilon = \boxed{\frac{2\pi\chi^\varepsilon\theta_{\alpha\alpha}^\varepsilon}{L^\varepsilon \mathrm{We}} + \frac{2\pi^2}{L^{\varepsilon 2}} H((\gamma^\varepsilon)^2\chi^\varepsilon\theta_\alpha^\varepsilon)} + \frac{2\pi}{L^\varepsilon}\chi^\varepsilon\left((T^\varepsilon - \mathbf{W}^\varepsilon \cdot \hat{\mathbf{t}}^\varepsilon)\chi^\varepsilon\gamma_\alpha^\varepsilon\right) + m_\gamma^\varepsilon, \tag{2.27}$$

$$L_t^\varepsilon = -\int_0^{2\pi} \theta_\alpha^\varepsilon U^\varepsilon \, d\alpha. \tag{2.28}$$

These equations are taken together with the initial conditions

$$\theta^\varepsilon(\alpha, 0) = \theta_0(\alpha), \quad \gamma^\varepsilon(\alpha, 0) = \gamma_0(\alpha), \quad L^\varepsilon(0) = L_0. \tag{2.29}$$

We remark that in some of the terms with the mollification operator, χ^ε, it appears once, while in other terms χ^ε appears twice. When χ^ε appears twice in a term, it is so that we can perform integration by parts in later estimates. In particular, this will be important at the end of the proof of Theorem 4.3.

We assume χ^ε to be a self-adjoint smoothing operator which commutes with derivatives and the Hilbert transform. For example, χ^ε could be convolution with an approximation to the Dirac mass. We need the following lemmas for χ^ε; proofs can be found on page 131 of [MB02]. In what follows, unless otherwise noted, all Sobolev spaces are of functions which are periodic with period 2π. The norm can thus be found by integrating over any interval of length 2π.

LEMMA 2.1. *For $f \in H^{m-k}$ and $k \in \{0, 1, 2, \dots\}$, we have $\chi^\varepsilon f \in H^m$ and*

$$\|\chi^\varepsilon f\|_m \leq \frac{c}{\varepsilon^k}\|f\|_{m-k}.$$

LEMMA 2.2. *For $f \in H^1$ and $\varepsilon,\ \varepsilon' > 0$,*

$$\left\|\chi^\varepsilon f - \chi^{\varepsilon'} f\right\|_0 \leq \max(\varepsilon, \varepsilon')\|f\|_1.$$

In the θ_t^ε equation, the μ^ε term has no counterpart in the nonmollified equation. It appears here to enforce the condition that $z^\varepsilon(\alpha) - \alpha$ be 2π-periodic. We do this because knowing that the tangent angle is periodic does not imply that the underlying curve is periodic. If we make the definition

$$P(t) = \int_\alpha^{\alpha+2\pi} z_\alpha^\varepsilon(\alpha', t)\ d\alpha',$$

then we would like to know that $P(t) = 2\pi$. (It is clear that P is independent of α since it is the integral of a periodic function over one period.) We assume that the initial data satisfies $P(0) = 2\pi$. We will choose μ such that $P'(t) = 0$. Let $\widetilde{\theta_t^\varepsilon} = \theta_t^\varepsilon - \mu^\varepsilon$. Since $z_\alpha = \frac{L}{2\pi}e^{i\theta}$, we see

$$P'(t) = \int_\alpha^{\alpha+2\pi}\left[\frac{L_t^\varepsilon z_\alpha^\varepsilon}{L^\varepsilon} + i\theta_t^\varepsilon z_\alpha^\varepsilon\right]d\alpha' = \int_\alpha^{\alpha+2\pi}\left[\frac{L_t^\varepsilon z_\alpha^\varepsilon}{L^\varepsilon} + i\widetilde{\theta_t^\varepsilon} z_\alpha^\varepsilon\right]d\alpha' + i\mu^\varepsilon\int_\alpha^{\alpha+2\pi} z_\alpha^\varepsilon\ d\alpha'.$$

This tells us that by defining μ^ε by

$$\mu^\varepsilon(t) = -\frac{\displaystyle\int_\alpha^{\alpha+2\pi}\left[\frac{L_t^\varepsilon z_\alpha^\varepsilon}{L^\varepsilon} + i\widetilde{\theta_t^\varepsilon} z_\alpha^\varepsilon\right]d\alpha'}{\displaystyle i\int_\alpha^{\alpha+2\pi} z_\alpha^\varepsilon\ d\alpha'}, \tag{2.30}$$

we guarantee that $P'(t) = 0$. This implies that $z^\varepsilon(\alpha) - \alpha$ is always 2π-periodic. We remark that for the nonmollified equations, it can be shown that $z(\alpha) - \alpha$ remains 2π-periodic. We refer the reader to [Amb02] for this calculation.

Before defining the mollified quantities $m^\varepsilon,\ T^\varepsilon$, etc., we must first define z^ε and $\mathbf{W}^\varepsilon$. We start with z^ε:

$$z^\varepsilon(\alpha, t) = z^\varepsilon(0, t) + \int_0^\alpha \frac{L^\varepsilon(t)}{2\pi}\cos(\theta^\varepsilon(\alpha', t)) + i\frac{L^\varepsilon(t)}{2\pi}\sin(\theta^\varepsilon(\alpha', t))\ d\alpha'.$$

Although it is frequently unimportant, we define $z^\varepsilon(0,t)$ for completeness. Since we have chosen $T(0,t) = 0$, we see that $z_t^\varepsilon(0,t) = (\mathbf{W}\cdot\hat{\mathbf{n}})\Phi(\hat{\mathbf{n}})(0,t)$; that is, $z^\varepsilon(0,t)$ evolves only by the normal velocity. Using (2.12) and (2.13), this means

$$z_t^\varepsilon(0,t) = \mathrm{Re}\left(\Phi(\mathbf{W}^\varepsilon)^*(0,t)\frac{2\pi i z_\alpha^\varepsilon(0,t)}{L^\varepsilon(t)}\right)\frac{2\pi i z_\alpha^\varepsilon(0,t)}{L^\varepsilon(t)}.$$

Integrating this in time yields $z^\varepsilon(0,t)$. Without loss of generality, we can choose $z^\varepsilon(0,0) = 0$. We introduce $z_d^\varepsilon(\alpha,t) = z^\varepsilon(\alpha,t) - z^\varepsilon(0,t)$ since we will usually not need $z(0,t)$. Notice that $D_\alpha z^\varepsilon = D_\alpha z_d^\varepsilon$. We will also use the similar notation $z_d(\alpha,t) = z(\alpha,t) - z(0,t)$, as necessary.

We now turn our attention to defining $\mathbf{W}^\varepsilon$. In order to do this, we first rewrite the nonmollified $\mathbf{W}$. Since we are looking for curves, z, with the property that $z(\alpha)-\alpha$ is 2π-periodic, we rewrite (2.14) so that the integral is over one period:

$$\Phi(\mathbf{W})^*(\alpha) = \frac{1}{4\pi i}\mathrm{PV}\int_{b-\pi}^{b+\pi}\gamma(\alpha')\left[\cot\frac{1}{2}(z(\alpha)-z(\alpha'))\right]\,d\alpha'. \tag{2.31}$$

Here, b is any real number. We rewrite this in order to isolate its most important part. In particular, we rewrite the quantity in brackets as

$$\left(\frac{1}{z_\alpha(\alpha')}\cot\frac{1}{2}(\alpha-\alpha')\right) + \left(\cot\frac{1}{2}(z(\alpha)-z(\alpha')) - \frac{1}{z_\alpha(\alpha')}\cot\frac{1}{2}(\alpha-\alpha')\right).$$

Then (2.31) becomes

$$\Phi(\mathbf{W})^*(\alpha) = \frac{1}{2i}H\left(\frac{\gamma}{z_\alpha}\right) + \mathcal{K}[z]\gamma, \tag{2.32}$$

where we define the integral operator $\mathcal{K}[z]$ as

$$\mathcal{K}[z]f(\alpha) = \frac{1}{4\pi i}\int_{b-\pi}^{b+\pi} f(\alpha')\left[\cot\frac{1}{2}(z_d(\alpha)-z_d(\alpha')) - \frac{1}{z_\alpha(\alpha')}\cot\frac{1}{2}(\alpha-\alpha')\right]\,d\alpha'. \tag{2.33}$$

Since we are interested in curves, z, which are non-self-intersecting, $\mathcal{K}[z]$ is not a singular integral operator. To take advantage of this, we write the cotangent as a function which is analytic at the origin plus a singular part:

$$\cot(w) = \frac{1}{w} + G(w). \tag{2.34}$$

This lets us rewrite $\mathcal{K}$ as $\mathcal{K}_1 + \mathcal{K}_2$, where

$$\mathcal{K}_1[z]f(\alpha) = \frac{1}{2\pi i}\int_{b-\pi}^{b+\pi} f(\alpha')\left[\frac{1}{z_d(\alpha)-z_d(\alpha')} - \frac{1}{z_\alpha(\alpha')(\alpha-\alpha')}\right]\,d\alpha', \tag{2.35}$$

$$\mathcal{K}_2[z]f(\alpha) = \frac{1}{4\pi i}\int_{b-\pi}^{b+\pi} f(\alpha')\left[G\left(\frac{1}{2}(z_d(\alpha)-z_d(\alpha'))\right) - \frac{1}{z_\alpha(\alpha')}G\left(\frac{1}{2}(\alpha-\alpha')\right)\right]\,d\alpha'. \tag{2.36}$$

We make definite choices of b in these integrals. For $\mathcal{K}_1$, we choose $b = \pi$ so that we integrate over $[0, 2\pi]$. For $\mathcal{K}_2$, we choose b so that the interval of integration avoids the

poles of G. From (2.34), we see that the poles of G are the nonzero integral multiples of π. Thus, we choose $b = \alpha$ in $\mathcal{K}_2$ so that the maximum difference between α and α' is π. This will be discussed further in the next section. We are now able to define $\mathbf{W}^\varepsilon$ similarly to (2.32):

$$\Phi(\mathbf{W}^\varepsilon)^*(\alpha) = \frac{1}{2i} H\left(\frac{\gamma^\varepsilon}{z_\alpha^\varepsilon}\right)(\alpha) + \mathcal{K}_1[z^\varepsilon]\gamma^\varepsilon(\alpha) + \mathcal{K}_2[z^\varepsilon]\gamma^\varepsilon(\alpha). \tag{2.37}$$

The definitions of the remaining mollified quantities are routine. The tangent and normal vectors to the mollified curve are given by

$$\Phi(\hat{\mathbf{t}}^\varepsilon) = \frac{2\pi z_\alpha^\varepsilon}{L^\varepsilon}, \qquad \Phi(\hat{\mathbf{n}}^\varepsilon) = \frac{2\pi i z_\alpha^\varepsilon}{L^\varepsilon}.$$

The mollified version of U is given by $U^\varepsilon = \mathbf{W}^\varepsilon \cdot \hat{\mathbf{n}}^\varepsilon$. Notice that by defining U^ε, we have completed the definition of L_t^ε in (2.28). T^ε is defined by replacing θ and U by their mollified versions in (2.6). $\mathbf{m}^\varepsilon$ is also defined much as before, with

$$\mathbf{m}^\varepsilon = \mathbf{R}_1^\varepsilon + \mathbf{R}_2^\varepsilon + \mathbf{B}_1^\varepsilon + \mathbf{B}_2^\varepsilon. \tag{2.38}$$

The terms $\mathbf{R}_1^\varepsilon$, $\mathbf{R}_2^\varepsilon$, $\mathbf{B}_1^\varepsilon$, and $\mathbf{B}_2^\varepsilon$ are defined the same way as the nonmollified versions but in terms of z^ε and γ^ε instead of z and γ.

In the mollified equations, we introduced m_γ^ε. This is because we would like both of the γ factors to be inside the Hilbert transform in the second of the important terms in (2.23). Thus, we define $\mathbf{m}_\gamma^\varepsilon$ to be the collection of terms

$$m_\gamma^\varepsilon = \frac{\gamma^\varepsilon(L_t^\varepsilon - 2\pi \mathbf{m}^\varepsilon \cdot \hat{\mathbf{t}}^\varepsilon)}{L^\varepsilon} - [H, \gamma^\varepsilon]\left(\frac{2\pi^2 \gamma^\varepsilon \chi^\varepsilon \theta_\alpha^\varepsilon}{{L^\varepsilon}^2}\right). \tag{2.39}$$

3. Preliminaries. We will need to use a variety of routine estimates for integral operators and other functions in terms of θ and γ. We begin with an estimate for the composition of two functions; see page 11 of [Tay96] for the proof.

LEMMA 3.1. *If F is smooth and u is in $H^k \cap L^\infty$, then $\|F(u)\|_k \le c(1 + \|u\|_k)$. The constant, c, depends on $|F^{(j)}u|_\infty$ for j between 0 and k.*

The next lemma gives a bound for z in terms of θ and L. This is important since z occurs frequently in our integral operators but not in our energy function.

LEMMA 3.2. *Let s be a positive integer. If $\theta \in H^{s-1}$, then $z_d \in H^s$ and $\|z_d\|_s \le cL(1 + \|\theta\|_{s-1})$.*

Proof. We use the fact that $D_\alpha z_d = D_\alpha z$, and we use the following equivalent norm for $z_d \in H^s$:

$$\|z_d\|_s = \sum_{j=0}^{s} \|D_\alpha^j z_d\|_0 = \|z_d\|_0 + \sum_{j=0}^{s-1} \|D_\alpha^j z_\alpha\|_0 = \|z_d\|_0 + \|z_\alpha\|_{s-1}.$$

Recall the definition

$$z_d(\alpha, t) = \int_0^\alpha \frac{L(t)}{2\pi} e^{i\theta(\alpha', t)}\, d\alpha'.$$

Clearly, the H^0 norm of this is bounded by a constant times L. To bound $\|z_\alpha\|_{s-1}$, we use Lemma 3.1 with the formula $z_\alpha = \frac{L}{2\pi} e^{i\theta}$ to see that $\|z_\alpha\|_{s-1} \le cL(1 + \|\theta\|_{s-1})$. Notice that c is independent of θ since $|e^{i\theta}|_\infty = 1$. This concludes the proof of the lemma. ∎

We will also need a standard interpolation lemma for Sobolev spaces. The Sobolev spaces we use are for periodic functions.

LEMMA 3.3. *If $f \in H^s$, and m is a real number such that $s > m > 0$, then*

$$\|f\|_m \ \leq \ c\|f\|_s^{m/s}\|f\|_0^{1-m/s}.$$

Proof. We use Hölder's inequality for series in the form

$$\sum_{\xi=-\infty}^{\infty} a_\xi b_\xi w_\xi \leq \left(\sum_{\xi=-\infty}^{\infty} a_\xi^p w_\xi\right)^{1/p} \left(\sum_{\xi=-\infty}^{\infty} b_\xi^q w_\xi\right)^{1/q},$$

where $\frac{1}{p} + \frac{1}{q} = 1$. We use as weights $w_\xi = (\mathcal{F}f(\xi))^2$. We calculate

$$\begin{aligned} (3.1) \qquad & \sum_{\xi=-\infty}^{\infty} (1+|\xi|^2)^m (\mathcal{F}f(\xi))^2 \\ & \leq \left(\sum_{\xi=-\infty}^{\infty} (1+|\xi|^2)^s (\mathcal{F}f(\xi))^2\right)^{m/s} \left(\sum_{\xi=-\infty}^{\infty} (\mathcal{F}f(\xi))^2\right)^{1-m/s}. \end{aligned}$$

This completes the proof. ∎

In simplifying the integral operator $\mathcal{K}_1$, we find divided differences to be very useful. We have the integral representation formulae for the divided differences q_1 and q_2:

$$q_1(\alpha,\alpha') = \frac{z_d(\alpha) - z_d(\alpha')}{\alpha - \alpha'} = \int_0^1 z_\alpha\left(t\alpha + (1-t)\alpha'\right)\, dt,$$

$$q_2(\alpha,\alpha') = \frac{z_d(\alpha) - z_d(\alpha') - z_\alpha(\alpha)(\alpha-\alpha')}{(\alpha-\alpha')^2} = \int_0^1 (t-1) z_{\alpha\alpha}\left((1-t)\alpha + t\alpha'\right)\, dt.$$

The next lemma gives bounds for the divided differences in terms of z. Rather than include a proof, we refer the reader to [BHL93]. (The proof makes use of the integral representations above.) The version of the lemma in [BHL93] was not for periodic functions, but that is not an important difference.

LEMMA 3.4. *Let $z_d \in H^m[a,b]$ and $k \leq m-1$. Then in either α or α', $q_1 \in H^{m-1}[a,b]$ and $q_2 \in H^{m-2}[a,b]$. Furthermore, there are the bounds*

$$\|q_1\|_{m-1} \ \leq \ c\|z_d\|_m, \qquad \|q_2\|_{m-2} \ \leq \ c\|z_d\|_m.$$

Remark. This lemma is true for any function in H^m, not only for z_d.

It now becomes important that we consider only curves z which are non-self-intersecting. In particular, when we specify that z is non-self-intersecting, we require that it satisfy the following estimate for some $\bar{c} > 0$ and for all α and α':

$$(3.2) \qquad \bar{c} < \left|\frac{z_d(\alpha) - z_d(\alpha')}{\alpha - \alpha'}\right|.$$

In addition to ruling out self-intersections, this condition also guarantees that the curve z has no cusps.

We also begin to consider only triples (θ, γ, L) which lie in a bounded set. To do this, we define the energy for a triple (θ, γ, L). Our energy function is

$$E(t) = E^0(t) + L^2(t) + \sum_{j=2}^{r} E^j(t), \tag{3.3}$$

where $E^0(t) = \|\theta\|_0^2 + \|\gamma\|_0^2$ and

$$E^j(t) = \frac{1}{2}\int (D_\alpha^{j-1}\theta)^2 + \frac{\pi \mathrm{We}}{L}(D_\alpha^{j-2}\gamma)\Lambda(D_\alpha^{j-2}\gamma) + \frac{\gamma^2\pi^2\mathrm{We}^2}{L^2}(D_\alpha^{j-2}\gamma)^2 \, d\alpha \tag{3.4}$$

for $j \geq 2$. We write $E^j = E_1^j + E_2^j + E_3^j$. We make the assumptions that

$$2\pi \, < \, L, \qquad E(t) \, < \, \bar{d}, \tag{3.5}$$

where $\bar{d}$ is some positive constant. This implies that the H^{r-1} norm of θ, the $H^{r-3/2}$ norm of γ, and the size of L are all bounded above by a constant. We have also assumed that L is bounded below by 2π. We make this assumption for technical reasons (related to the proof of Theorem 5.1). That L is bounded below by 2π is automatically true for any nontrivial curve z such that $z(\alpha) - \alpha$ is 2π-periodic. By Lemma 3.2, (3.5) also implies that the H^r norm of z_d is bounded by a constant. When choosing $\bar{c}$ and $\bar{d}$, we make sure that $\bar{c}$ is small enough and $\bar{d}$ is large enough so that conditions (3.2) and (3.5) are satisfied by the initial data.

Remark on minimal regularity. In the definition of the energy function (3.3), we have introduced an integer r. In what follows, we will be assuming the initial data θ_0 to be in H^{r-1} and γ_0 to be in $H^{r-3/2}$. We will sometimes make an assumption that "r is large enough." This means that we are assuming $r > k$, where k is an absolute constant. We do not determine here the minimal value of k which is necessary to make our proofs hold. This is because this would surely not be sharp.

We prove now that the operator $\mathcal{K} = \mathcal{K}_1 + \mathcal{K}_2$ defined in (2.35) and (2.36) is smooth.

LEMMA 3.5. *Let s be an integer such that $s \geq 2$. If $z_d \in H^s$, then $\mathcal{K}[z] : H^1 \to H^{s-1}$ and, in particular, there are positive constants C_1 and C_2 such that*

$$\|\mathcal{K}[z]f\|_{s-1} \leq C_1\|f\|_1 \exp\{C_2\|z_d\|_s\}.$$

Similarly, $\mathcal{K}[z] : H^0 \to H^{s-2}$, and $\|\mathcal{K}[z]f\|_{s-2} \leq C_1\|f\|_0 \exp\{C_2\|z_d\|_s\}$.

Proof. We will deal with $\mathcal{K}_1$ and $\mathcal{K}_2$ separately. We begin by taking $s-1$ derivatives of $\mathcal{K}_1[z]f$.

$$D_\alpha^{s-1}\mathcal{K}_1[z]f(\alpha) = \frac{1}{2\pi i}\int_0^{2\pi} f(\alpha')D_\alpha^{s-1}\left[\frac{1}{z_d(\alpha) - z_d(\alpha')} - \frac{1}{z_\alpha(\alpha')(\alpha - \alpha')}\right] d\alpha'.$$

We apply one of the $s-1$ derivatives to the quantity in brackets.

$$D_\alpha^{s-1}\mathcal{K}_1[z]f(\alpha) = \frac{1}{2\pi i}\int_0^{2\pi} f(\alpha')D_\alpha^{s-2}\left[-\frac{z_\alpha(\alpha)}{(z_d(\alpha) - z_d(\alpha'))^2} + \frac{1}{z_\alpha(\alpha')(\alpha - \alpha')^2}\right] d\alpha'.$$

We rearrange the factors of z_α and write the quantity in brackets as an α'-derivative.

$$D_\alpha^{s-1}\mathcal{K}_1[z]f(\alpha) = \frac{1}{2\pi i}\int_0^{2\pi} \frac{f(\alpha')}{z_\alpha(\alpha')}D_\alpha^{s-2}D_{\alpha'}\left[-\frac{z_\alpha(\alpha)}{z_d(\alpha) - z_d(\alpha')} + \frac{1}{\alpha - \alpha'}\right] d\alpha'.$$

We integrate this by parts and estimate it, recognizing that the quantity in brackets is a ratio of divided differences.

$$|D_\alpha^{s-1}\mathcal{K}_1[z]f(\alpha)| = \left|\frac{1}{2\pi i}\int_0^{2\pi} D_{\alpha'}\left(\frac{f(\alpha')}{z_\alpha(\alpha')}\right) D_\alpha^{s-2}\left[\frac{z_\alpha(\alpha)}{z_d(\alpha)-z_d(\alpha')} - \frac{1}{\alpha-\alpha'}\right]\,d\alpha'\right|$$

$$\leq c\left\|\frac{f}{z_\alpha}\right\|_1 \left\|\frac{q_2}{q_1}\right\|_{s-2} \leq c\|f\|_1 \left\|\frac{1}{z_\alpha}\right\|_1 \|q_2\|_{s-2} \left\|\frac{1}{q_1}\right\|_{s-2}. \tag{3.6}$$

Using (3.2) and Lemma 3.1, we estimate $\|1/q_1\|_{s-2} \leq c(1+\|q_1\|_{s-2})$. Similarly, since $|z_\alpha| = L/2\pi$, we use Lemma 3.1 to estimate

$$\left\|\frac{1}{z_\alpha}\right\|_1 \leq c(1+\|z_\alpha\|_1). \tag{3.7}$$

To finish the estimate for $\mathcal{K}_1$, we use Lemma 3.4 and (3.6) to conclude

$$\|\mathcal{K}_1[z]f\|_{s-1} \leq C_1\|f\|_1 \exp\{C_2\|z_d\|_s\}. \tag{3.8}$$

We calculate $D_\alpha^{s-1}\mathcal{K}_2[z]f$. It is

$$\frac{1}{2\pi i}\int_{\alpha-\pi}^{\alpha+\pi} f(\alpha')D_\alpha^{s-1}\left[G\left(\frac{1}{2}(z_d(\alpha)-z_d(\alpha'))\right) - \frac{1}{z_\alpha(\alpha')}G\left(\frac{1}{2}(\alpha-\alpha')\right)\right]\,d\alpha'. \tag{3.9}$$

The largest number of derivatives which fall on z_d is $s-1$, and we assumed $z_d \in H^s$. We can estimate the factor of $1/z_\alpha$ in H^{s-1} similarly to (3.7). It is important to notice that the poles of G have been avoided. That is, $z_d(\alpha)-z_d(\alpha')$ cannot approach any multiple of 2π because of the periodicity of $z(\alpha)-\alpha$ and the condition (3.2). By (3.5), we know that z_d is in a bounded set, and we also know that G is analytic. So, we can use Lemma 3.1 to conclude that

$$\|\mathcal{K}_2[z]f\|_{s-1} \leq C_1\|f\|_0 \exp\{C_2\|z_d\|_s\}. \tag{3.10}$$

Combining (3.8) and (3.10), we have proved the lemma. □

We also need a Lipschitz-type estimate for $\mathcal{K}$. We state the lemma here, but we refer the reader to [Amb02] for the proof.

LEMMA 3.6. *Given z and z' in H^r which satisfy* (3.2) *and* (3.5), *we have the Lipschitz estimate*

$$\|\mathcal{K}[z]f - \mathcal{K}[z']f\|_1 \leq c\|\theta-\theta'\|_1\|f\|_1.$$

The proof is routine. It involves considering $\mathcal{K}_1$ and $\mathcal{K}_2$ individually. For $\mathcal{K}_1$, Lipschitz estimates on the divided differences q_1 and q_2 are used. For $\mathcal{K}_2$, Lipschitz estimates for the function G defined in (2.34) are used.

We will frequently need estimates regarding the smoothness of the commutator of the Hilbert transform and multiplication by a smooth function.

LEMMA 3.7. *For $\psi \in H^s$, the operator $[H,\psi]$ is bounded from H^0 to H^{s-1}. Also, $[H,\psi]$ is bounded from H^{-1} to H^{s-2}. For $i=0$ or $i=-1$, we have*

$$\|[H,\psi]f\|_{s-1+i} \leq c\|f\|_i\|\psi\|_s.$$

Proof. We begin by writing $[H,\psi]$ as an integral operator:

$$[H,\psi]f(\alpha) = \frac{1}{2\pi i}\int_b^{b+2\pi} f(\alpha')\Big(\psi(\alpha')-\psi(\alpha)\Big)\cot\left(\frac{1}{2}(\alpha-\alpha)\right)\,d\alpha'.$$

Since the functions are periodic, we may choose $b = \alpha$, so that the maximum difference between α and α' is π. We can write the kernel as

$$\left(\frac{\psi(\alpha) - \psi(\alpha')}{\alpha - \alpha'}\right)\left((\alpha - \alpha')\cot\frac{1}{2}(\alpha - \alpha')\right).$$

The first part of this product is a divided difference, and the second part is an analytic function on the domain in which we are interested. The lemma now follows from Lemma 3.4 and from the generalized Young's inequality [Fol95]. For $g \in H^{-1}$, we integrate by parts once; otherwise, the proof is similar. □

COROLLARY 3.8. *For $s \geq 3$ and $\psi \in H^s$, the operator $[H, \psi]$ is bounded from H^{s-2} to H^s. For $s \geq 4$ and $\psi \in H^{s-1/2}$, $[H, \psi]$ is bounded from H^{s-2} to $H^{s-1/2}$. For $i = 0$ or $i = -1/2$, we have the estimates $\|[H, \psi]f\|_{s+i} \leq c\|f\|_{s-2}\|\psi\|_{s+i}$.*

Proof. Let $g \in H^{s-2}$. We compute $D_\alpha^s[H, \psi]g$ by using the product rule for $s - 2$ of the derivatives.

$$\begin{aligned} D_\alpha^s[H, \psi]g &= D_\alpha^2(D_\alpha^{s-2}(H(\psi g) - \psi H(g))) \\ &= D_\alpha^2 \sum_{k=0}^{s-2} \binom{s-2}{k} \left[H((D_\alpha^k\psi)(D_\alpha^{s-k-2}g)) - (D_\alpha^k\psi)(HD_\alpha^{s-k-2}g)\right]. \end{aligned}$$

We consider the summands separately in the cases $k = 0$, $k = 1$, and $k > 1$. The simplest case is $k > 1$, for then the highest number of derivatives on g is $s - 4$ and the highest number of derivatives on ψ is $s - 2$. The products of these terms are thus in H^2, which is fortunate since we are taking two derivatives of the sum. Thus, the H^0 norms of these terms are bounded by $c\|\psi\|_s\|g\|_{s-2}$. When $k = 0$, we have

$$D_\alpha^2 H(\psi D_\alpha^{s-2}g) - D_\alpha^2(\psi H D_\alpha^{s-2}(g)) = D_\alpha^2[H, \psi](D_\alpha^{s-2}(g)).$$

Lemma 3.7 applies to this term, so its H^0 norm is bounded by $c\|\psi\|_3\|g\|_{s-2}$. The $k = 1$ term is similarly bounded by $c\|\psi\|_3\|g\|_{s-2}$. This proves the first part of the corollary. The proof for $\psi \in H^{s-1/2}$ is similar. □

The final lemma we will need is about commutator operators associated with taking a large number of derivatives. This lemma, though, relies on another lemma about norms of products of functions in Sobolev spaces.

LEMMA 3.9. *Suppose $s > 1/2$ and $s \geq s' \geq 0$. If $f \in H^s$ and $g \in H^{s'}$, then $fg \in H^{s'}$ with the estimate*

$$\|fg\|_{s'} \leq c\|f\|_s\|g\|_{s'}.$$

We do not prove this lemma here. Instead, we refer the reader to page 366 of [Bea81]. The version of the lemma in [Bea81] is for functions on $\mathbb{R}^3$, but the proof of Lemma 3.9 is essentially the same. We are now prepared to give our final lemma.

LEMMA 3.10. *If k is a positive integer such that $k \geq 2$, s is a positive real number, $f \in H^{s+k}$, and $g \in H^{s+k-1}$, then the following estimate for commutators of derivatives holds:*

$$\|D_\alpha^k(fg) - fD_\alpha^k g\|_s \leq c\|f\|_{s+k}\|g\|_{s+k-1}.$$

Proof. We first notice that the product rule implies that $D_\alpha^k(fg) - fD_\alpha^k g = \sum_{j=1}^k \binom{k}{j}(D_\alpha^{k-j}g)(D_\alpha^j f)$. We now use the triangle inequality:

$$\left\|D_\alpha^k(fg) - fD_\alpha^k g\right\|_s \leq \sum_{j=1}^k \binom{k}{j} \left\|(D_\alpha^{k-j}g)(D_\alpha^j f)\right\|_s. \tag{3.11}$$

Lemma 3.9 implies that each of these individual summands can be bounded by $c\|f\|_{s+k}\|g\|_{s+k-1}$. This completes the proof, but we illustrate this with an example. If $k = 3$ and $s = 1/2$, then (3.11) reads

$$\begin{aligned}(3.12)\quad &\left\|(D_\alpha^3 f)g + 3(D_\alpha^2 f)(D_\alpha g_\alpha) + 3(D_\alpha f)(D_\alpha^2 g)\right\|_{1/2} \\ &\qquad \leq \left\|(D_\alpha^3 f)g\right\|_{1/2} + 3\left\|(D_\alpha^2 f)(D_\alpha g)\right\|_{1/2} + 3\left\|(D_\alpha f)(D_\alpha^2 g)\right\|_{1/2}.\end{aligned}$$

Lemma 3.9 implies that the first term on the right-hand side of (3.12) can be bounded by $c\|f\|_{7/2}\|g\|_{3/2}$, the second can be bounded by $c\|f\|_{7/2}\|g\|_{3/2}$ (or by $c\|f\|_{5/2}\|g\|_{5/2}$), and the third can be bounded by $c\|f\|_{5/2}\|g\|_{5/2}$. The sum of these can then be bounded by $c\|f\|_{7/2}\|g\|_{5/2}$. □

4. A priori estimates. We wish to consider solutions of the initial value problem which lie in a particular Banach space, $H^{r-1} \times H^{r-3/2} \times \mathbb{R}$. We call this space $\mathcal{B}$. Our solutions, (θ, γ, L), will be contained in an open subset, $\mathcal{O}$, of $\mathcal{B}$; in particular, $\mathcal{O}$ is the subset of $\mathcal{B}$ in which the conditions (3.2) and (3.5) hold. Notice that this open set depends on the particular choice of $\bar{c}$ and $\bar{d}$. The norm of $(\theta, \gamma, L) \in \mathcal{B}$ is defined to be $\|\theta\|_{r-1} + \|\gamma\|_{r-3/2} + |L|$.

The goal of this section is to provide a bound on the growth of the energy of a solution to the mollified initial value problem. We begin by providing routine estimates of quantities related to θ, γ, and L. These estimates hold for both the mollified and for the nonmollified quantities. Throughout this section, all constants are independent of α, t, and ε, but they may depend on r, $\bar{c}$, $\bar{d}$, or the Weber number.

LEMMA 4.1. *If $|z_\alpha(\alpha, t)| = L(t)/2\pi$, $L(t) > 2\pi$, and $z_\alpha \in H^{r-1}([a, b])$, then $1/z_\alpha$ is in H^{r-1} with the estimate*

$$\left\|\frac{1}{z_\alpha}\right\|_{r-1} \leq c(1 + \|z_\alpha\|_{r-1}),$$

where c depends on r and on the interval $[a, b]$.

Proof. We use Lemma 3.1 with $F(u) = 1/u$. The assumption that

$$\left|\frac{1}{z_\alpha}\right| = \frac{2\pi}{L} < 1$$

implies that $|F^{(j)}u|_\infty < 1$ for all j. Thus, we have proved the lemma and the constant c is independent of both α and t. □

We will use Lemma 4.1 at some points in the proof of Lemma 4.2.

LEMMA 4.2. *For $(\theta, \gamma, L) \in \mathcal{O}$, the following bounds hold:*

$$\|\mathbf{W}\|_0 \leq C_1\|\gamma\|_1 \exp\{C_2\|z_d\|_1\}, \tag{4.1}$$

$$|L_t| \leq C\|\theta\|_1\|\mathbf{W}\|_0, \tag{4.2}$$

$$\|\mathbf{m}\|_{r-1} \leq C_1\|\gamma\|_{r-2} \exp\{C_2\|z_d\|_r\}, \tag{4.3}$$

$$\|m_\gamma\|_{r-3/2} \leq C\|\gamma\|_{r-3/2}(|L_t| + \|\mathbf{m} \cdot \hat{\mathbf{t}}\|_{r-1} + \|\gamma\|_{r-2}\|\theta\|_{r-1}), \tag{4.4}$$

$$\|D_\alpha(T - \mathbf{W} \cdot \hat{\mathbf{t}})\|_{r-2} \leq C(|L_t| + \|\theta\|_{r-1}\|\gamma\|_{r-2} + \|\mathbf{m} \cdot \hat{\mathbf{t}}\|_{r-1}), \tag{4.5}$$

$$|\mu^\varepsilon(t)| \leq c(|L_t^\varepsilon| + \|\gamma^\varepsilon\|_1 + \|T^\varepsilon - \mathbf{W}^\varepsilon \cdot \hat{\mathbf{t}}^\varepsilon\|_0\|\theta^\varepsilon\|_1 + \|\mathbf{m}^\varepsilon\|_0). \tag{4.6}$$

Proof. The bounds (4.1), (4.2), (4.3), (4.4), and (4.5) hold for both the mollified and for the nonmollified quantities. In (4.6), however, we make the ε-dependence explicit because there is no nonmollified counterpart to μ^ε.

Proof. We use the representation $\Phi(\mathbf{W})^* = \frac{1}{2i} H(\frac{\gamma}{z_\alpha}) + \mathcal{K}[z]\gamma$ to estimate $\|\mathbf{W}\|_0$. In making this estimate, we make use of Lemma 3.5. Thus,

$$\|\mathbf{W}\|_0 \;\le\; c_1 \left(\|\gamma\|_0 \left| \frac{1}{z_\alpha} \right|_\infty + \|\gamma\|_1 \exp\{c_2 \|z_d\|_1\} \right).$$

To estimate $|1/z_\alpha|_\infty$, we use that $|z_\alpha| = L/2\pi$ and $2\pi < L$. This implies the estimate $\|\mathbf{W}\|_0 \;\le\; c_3 \|\gamma\|_1 \exp\{c_2 \|z_d\|_1\}$.

We turn to L_t. From (2.28) we easily compute

$$|L_t| \le \int_0^{2\pi} |\theta_\alpha| |U| \; d\alpha \;\le\; c \|\theta\|_1 \|\mathbf{W} \cdot \hat{\mathbf{n}}\|_0 \;\le\; c \|\theta\|_1 \|\mathbf{W}\|_0 |\hat{\mathbf{n}}|_\infty.$$

Since $\hat{\mathbf{n}}$ is a unit vector, $|\hat{\mathbf{n}}|_\infty = 1$. This implies the bound (4.2).

We see from (2.38) that $\mathbf{m}$ is made up of two kinds of terms: commutators and integral remainder operators. We write $\mathbf{m} = (\mathbf{m} \cdot \hat{\mathbf{t}})\hat{\mathbf{t}} + (\mathbf{m} \cdot \hat{\mathbf{n}})\hat{\mathbf{n}}$, and bound the normal and tangential parts separately. We use Corollary 3.8 on the commutators and Lemma 3.5 on the integral remainder operators. More specifically, in the tangential part, the commutator is

$$\mathbf{B}_1 \cdot \hat{\mathbf{t}} + \mathbf{B}_2 \cdot \hat{\mathbf{t}} = \mathrm{Re} \left(\frac{z_\alpha^2}{2iL} \left[H, \frac{1}{z_\alpha^2} \right] \left(\gamma_\alpha - \frac{\gamma z_{\alpha\alpha}}{z_\alpha} \right) \right).$$

By Corollary 3.8, we can bound its H^{r-1} norm by

$$c \|z_\alpha\|_{r-1}^2 \left\| \frac{1}{z_\alpha} \right\|_{r-1} \left\| \frac{\gamma z_{\alpha\alpha}}{z_\alpha} - \gamma_\alpha \right\|_{r-3}.$$

By Lemma 4.1, $\|1/z_\alpha\|_{r-1} \le c(1 + \|z_\alpha\|_{r-1})$. Similarly,

$$\left\| \frac{\gamma z_{\alpha\alpha}}{z_\alpha} - \gamma_\alpha \right\|_{r-3} \le c \|\gamma\|_{r-3} (1 + \|z_\alpha\|_{r-2})^2 + \|\gamma\|_{r-2}$$

By applying Lemma 3.5 to the other part of $\mathbf{m} \cdot \hat{\mathbf{t}}$, which is $\mathbf{R}_1 \cdot \hat{\mathbf{t}} + \mathbf{R}_2 \cdot \hat{\mathbf{t}}$, we see that it is bounded in H^{r-1} by

$$\left(\left\| \frac{\gamma_\alpha}{z_\alpha} \right\|_1 + \left\| \frac{\gamma z_{\alpha\alpha}}{z_\alpha^2} \right\|_1 \right) C_1 \exp\{C_2 \|z_d\|_r\} \|z_d\|_r.$$

This is in turn bounded by

$$\left(\|\gamma\|_2 (1 + \|z_d\|_2) + \|\gamma\|_1 (1 + \|z_d\|_3)^3 \right) C_1 \exp\{C_2 \|z_d\|_r\} \|z_d\|_r.$$

As long as $r \ge 4$, we can put all of this into one bound, namely,

$$\|\mathbf{m} \cdot \hat{\mathbf{t}}\|_{r-1} \le C_1 \|\gamma\|_{r-2} \exp\{C_2 \|z_d\|_r\}.$$

The estimate for $\mathbf{m} \cdot \hat{\mathbf{n}}$ is similar, and this leads to the bound (4.3). Note that the bound on m_γ also requires the use of Corollary 3.8.

The bound on $D_\alpha(T - \mathbf{W} \cdot \hat{\mathbf{t}})$ is immediate from (2.9) and (2.20). To bound $|\mu^\varepsilon|$, we begin with the denominator in (2.30). Using (3.2), we have

$$\left| iL^\varepsilon \int_\alpha^{\alpha+2\pi} z_\alpha^\varepsilon(\alpha') \; d\alpha' \right| \;\ge\; 2\pi |\bar{c} L^\varepsilon| \;\ge\; 4\pi^2 \bar{c}. \tag{4.7}$$

Next, we look at the numerator of $|\mu^\varepsilon|$; it can be expressed as

$$\left| \int_\alpha^{\alpha+2\pi} \left[\frac{L_t^\varepsilon z_\alpha^\varepsilon}{L^\varepsilon} + \frac{2\pi i z_\alpha^\varepsilon}{L^\varepsilon} \left(\frac{\pi}{L^\varepsilon} (H\chi^\varepsilon \gamma_\alpha^\varepsilon) + \chi^\varepsilon [(T^\varepsilon - \mathbf{W}^\varepsilon \cdot \hat{\mathbf{t}}^\varepsilon)\chi^\varepsilon \theta_\alpha^\varepsilon] + \mathbf{m}^\varepsilon \cdot \hat{\mathbf{n}}^\varepsilon \right) \right] d\alpha' \right|.$$

Using Lemma 2.1 and the fact that $|z_\alpha^\varepsilon / L^\varepsilon| < c$, we bound this by

$$c(|L_t^\varepsilon| + \|\gamma_\alpha^\varepsilon\|_0 + \|T^\varepsilon - \mathbf{W}^\varepsilon \cdot \hat{\mathbf{t}}^\varepsilon\|_0 \|\theta_\alpha^\varepsilon\|_0 + \|\mathbf{m}\|_0).$$

Combined with (4.7), this clearly implies the estimate (4.6). □

We are now ready to prove the main estimate.

THEOREM 4.3. *Let $(\theta, \gamma, L) \in C([0,T); \mathcal{O})$ be a solution to the mollified initial value problem for the vortex sheet with surface tension* ((2.26), (2.27), (2.28), *and* (2.29)). *If the initial data θ_0 is in H^{r-1} and γ_0 is in $H^{r-3/2}$, where the integer r is large enough, then there exists a time $T^* > 0$ and positive constants C_1 and C_2 such that for all t satisfying $t \le T < T^*$,*

$$E(t) \le -\frac{\ln(e^{-C_2 E(0)} - C_1 C_2 t)}{C_2}. \tag{4.8}$$

The time T^* depends only upon the initial data, r, the Weber number, and the set $\mathcal{O}$.

Proof. Recall the definition of the energy in (3.3) and (3.4). We compute the time derivative of the energy. Our conclusion will be that for some positive constants C_1 and C_2,

$$\frac{dE}{dt} \le C_1 \exp\{C_2 E\}.$$

Demonstrating this bound will prove the theorem since this differential inequality has a solution until time $T^* = \frac{e^{-C_2 E(0)}}{C_1 C_2}$, and until that time, the bound (4.8) holds.

We first compute $\frac{dE_1^j}{dt}$. From the definition of E_1^j, we have

$$\frac{dE_1^j}{dt} = \int (D_\alpha^{j-1}\theta)(D_\alpha^{j-1}\theta_t)\, d\alpha.$$

Plugging in from (2.26) for θ_t, this is

$$\begin{aligned} \frac{dE_1^j}{dt} = {} & \int (D_\alpha^{j-1}\theta) \left(\frac{2\pi^2}{L^2} H\chi^\varepsilon (D_\alpha^j \gamma) \right) d\alpha \\ & + \int (D_\alpha^{j-1}\theta) D_\alpha^{j-1} \left(\chi^\varepsilon \left(\frac{2\pi}{L} (T - \mathbf{W}\cdot\hat{\mathbf{t}})\chi^\varepsilon D_\alpha \theta \right) \right) d\alpha \\ & + \int 2\pi (D_\alpha^{j-1}\theta) \left(\frac{D_\alpha^{j-1}(\mathbf{m}\cdot\hat{\mathbf{n}})}{L} + \mu \right) d\alpha. \end{aligned} \tag{4.9}$$

We rewrite the second integral in (4.9) simply by adding and subtracting. This is because the most important term from that integral comes when all of the $j-1$ derivatives fall on θ:

$$\begin{aligned} & \int (D_\alpha^{j-1}\theta) D_\alpha^{j-1} \left(\chi^\varepsilon \left(\frac{2\pi}{L} (T - \mathbf{W}\cdot\hat{\mathbf{t}})\chi^\varepsilon D_\alpha \theta \right) \right) d\alpha \\ & \qquad = \int (D_\alpha^{j-1}\theta)\chi^\varepsilon \left(\frac{2\pi}{L} (T - \mathbf{W}\cdot\hat{\mathbf{t}})\chi^\varepsilon D_\alpha^j \theta \right) d\alpha \\ & \quad + \int (D_\alpha^{j-1}\theta)\chi^\varepsilon \left(\frac{2\pi}{L} \left[D^{j-1}((T - \mathbf{W}\cdot\hat{\mathbf{t}})\chi^\varepsilon D_\alpha \theta) - (T - \mathbf{W}\cdot\hat{\mathbf{t}})\chi^\varepsilon D_\alpha^j \theta \right] \right) d\alpha. \end{aligned} \tag{4.10}$$

We introduce the notation

$$\frac{dE_1^j}{dt} = P_1^j + V_1^j + B_1^j + B_2^j,$$

where P_1^j is the first integral on the right-hand side of (4.9), V_1^j is the first term on the right-hand side of (4.10), B_1^j is the second term on the right-hand side of (4.10), and B_2^j is the third integral on the right-hand side of (4.9). Notice that P_1^j corresponds to the boxed term in (2.26). P_1^j will cancel with another term we will see soon; this will be the primary cancellation and will demonstrate the usefulness of the surface tension in this problem. V_1^j is a transport term which we will discuss in more detail later. Notice that if we were using the purely Lagrangian tangential velocity ($T = \mathbf{W} \cdot \hat{\mathbf{t}}$), there would be no transport term. Both of the B terms are bounded in terms of our energy, as we will now show.

To demonstrate that the B terms are bounded, we make frequent use of Lemma 4.2. We start with B_1^j. The factor $D_\alpha^{j-1}\theta$ that multiplies the difference in B_1^j can clearly be bounded by $cE^{1/2}$. Using Lemma 3.10, we bound the quantity in brackets by $c\|T - \mathbf{W} \cdot \hat{\mathbf{t}}\|_{j-1}\|\theta\|_{j-1}$. By Lemma 4.2 and the definition of the energy function, we can conclude that $B_1^j \leq C_1 \exp\{C_2 E\}$.

We estimate B_2^j by $cE^{1/2}\|\mathbf{m} \cdot \hat{\mathbf{n}} + \mu\|_{j-1}$. Recall that j is at most r. The Sobolev algebra property implies $\|\mathbf{m} \cdot \hat{\mathbf{n}} + \mu\|_{r-1} \leq c\|\mathbf{m}\|_{r-1}\|\hat{\mathbf{n}}\|_{r-1} + c|\mu|$. Lemma 4.2 tells us that $\|\mathbf{m}\|_{r-1} \leq C_1\|\gamma\|_{r-2}\exp\{C_2\|z_d\|_r\}$, and by the definition of $\hat{\mathbf{n}}$ we can bound $\|\hat{\mathbf{n}}\|_{r-1}$ by $\|z_d\|_r$. Also, Lemma 3.2 gives the bound $\|z_d\|_r \leq cL(1+\|\theta\|_{r-1})$. Similarly, we could use Lemma 4.2 to bound $|\mu|$ in terms of the norms of θ and γ. These estimates combine to yield the estimate $B_2^j \leq C_1 \exp\{C_2 E\}$.

We next take the time derivative of E_2^j. This is

$$\frac{dE_2^j}{dt} = \frac{\pi \mathrm{We}}{L}\int (D_\alpha^{j-2}\gamma_t)\Lambda(D_\alpha^{j-2}\gamma)\, d\alpha - \frac{\pi \mathrm{We} L_t}{L^2}\int (D_\alpha^{j-2}\gamma)\Lambda(D_\alpha^{j-2}\gamma)\, d\alpha.$$

We plug in for γ_t from (2.27). We use the following notation:

$$\frac{dE_2^j}{dt} = P_2^j + S_1^j + V_2^j + B_3^j + B_4^j + B_5^j,$$

where

$$P_2^j = \int \frac{2\pi^2}{L^2}(\chi^\varepsilon D_\alpha^j \theta)(\Lambda D_\alpha^{j-2}\gamma)\, d\alpha,$$

$$S_1^j = \int \frac{2\pi^3 \mathrm{We}}{L^3} H(\gamma^2 \chi^\varepsilon D_\alpha^{j-1}\theta)(\Lambda D_\alpha^{j-2}\gamma)\, d\alpha, \tag{4.11}$$

and

$$V_2^j = \int \left(\chi^\varepsilon \left(\frac{2\pi^2 \mathrm{We}}{L^2}(T - \mathbf{W} \cdot \hat{\mathbf{t}})\chi^\varepsilon D_\alpha^{j-1}\gamma\right)\right)\Lambda(D_\alpha^{j-2}\gamma)\, d\alpha.$$

Notice that P_2^j and S_1^j correspond to the boxed terms in (2.27). The primary cancellation that we mentioned earlier is between P_1^j and P_2^j. If there were no surface tension in the problem ($\mathrm{We} = \infty$), P_2^j would not be here. S_1^j will be involved in the

secondary cancellation, along with a term from E_3^j which we will soon see. S_1^j is the reason we need to include the unusual term E_3^j in the energy function at all. V_2^j is a transport term which we will deal with later. Again, if we were using the purely Lagrangian tangential velocity ($T = \mathbf{W} \cdot \hat{\mathbf{t}}$), the transport term would not be present.

We look now at the more easily bounded terms.

$$B_3^j = \int \frac{2\pi^3 \mathrm{We}}{L^3} \Lambda(D_\alpha^{j-2}\gamma) H \left[D_\alpha^{j-2}(\gamma^2 \chi^\varepsilon \theta_\alpha) - (\gamma^2 \chi^\varepsilon D_\alpha^{j-1}\theta) \right] \, d\alpha. \tag{4.12}$$

Like S_1^j, this corresponds to the second of the boxed terms in (2.27). This term is very similar to B_1^j. The next term,

$$B_4^j = \int \frac{2\pi^2 \mathrm{We}}{L^2} \chi^\varepsilon \left[D_\alpha^{j-2}((T - \mathbf{W} \cdot \hat{\mathbf{t}})\gamma_\alpha) - (T - \mathbf{W} \cdot \hat{\mathbf{t}}) D_\alpha^{j-1}\gamma \right] \Lambda(D_\alpha^{j-2}\gamma) \, d\alpha, \tag{4.13}$$

is related to V_2^j and is also very similar to B_1^j. Both B_3^j and B_4^j can be bounded in terms of the energy by a straightforward application of Lemmas 3.10 and 4.2. Finally, the term

$$B_5^j = \int \frac{\pi \mathrm{We}}{L} (D_\alpha^{j-2} m_\gamma) \Lambda(D_\alpha^{j-2}\gamma) - \frac{\pi \mathrm{We}}{L^2} L_t (D_\alpha^{j-2}\gamma) \Lambda(D_\alpha^{j-2}\gamma) \, d\alpha$$

can be bounded by a straightforward application of Lemma 4.2, as was B_2^j.

We can now see the primary cancellation:

$$P_1^j + P_2^j = \int \frac{2\pi^2}{L^2} \left[(D_\alpha^{j-1}\theta)(H\chi^\varepsilon(D_\alpha^j \gamma)) + (\chi^\varepsilon D_\alpha^j \theta)(\Lambda D_\alpha^{j-2}\gamma) \right] \, d\alpha. \tag{4.14}$$

Using the facts that χ^ε is self-adjoint and commutes with D_α, that $\Lambda = HD_\alpha$, and integrating by parts on the second of the two terms, we see that

$$P_1^j + P_2^j = \int \frac{2\pi^2}{L^2} \left[(D_\alpha^{j-1}\theta)(H\chi^\varepsilon(D_\alpha^j \gamma)) - (D_\alpha^{j-1}\theta)(H\chi^\varepsilon D_\alpha^j \gamma) \right] \, d\alpha = 0.$$

We remark here that this cancellation occurs because of the surface tension term in the evolution equation for γ. Without this cancellation, we would not be able to bound the growth of the norm of the solution.

Continuing, we take the time derivative of E_3^j. As we said earlier, the only reason to include E_3^j in the energy is to cancel S_1^j. Using (2.27) to substitute for γ_t, we write

$$\frac{dE_3^j}{dt} = S_2^j + B_6^j,$$

where

$$S_2^j = \int \frac{\gamma^2 \pi^2 \mathrm{We}^2}{L^2} (D_\alpha^{j-2}\gamma) \left(D_\alpha^{j-2} \left(\frac{2\pi \chi^\varepsilon \theta_{\alpha\alpha}}{L\mathrm{We}} \right) \right) \, d\alpha \tag{4.15}$$

will be the term which is involved in this secondary cancellation. This corresponds to the first of the boxed terms in (2.27). The rest of the terms are more easily bounded.

They are

$$B_6^j = \int \left[\frac{\gamma^2\pi^2 \mathrm{We}^2}{L^2}(D_\alpha^{j-2}\gamma)\left(D_\alpha^{j-2}\left(\frac{2\pi^2}{L^2}H(\gamma^2\chi^\varepsilon\theta_\alpha)\right)\right) \right.$$
$$+\frac{\gamma^2\pi^2 \mathrm{We}^2}{L^2}(D_\alpha^{j-2}\gamma)\left(D_\alpha^{j-2}\chi^\varepsilon\left(\frac{2\pi}{L}(T-\mathbf{W}\cdot\hat{\mathbf{t}})\chi^\varepsilon\gamma_\alpha\right)\right)$$
$$+\frac{\gamma^2\pi^2 \mathrm{We}^2}{L^2}(D_\alpha^{j-2}\gamma)(D_\alpha^{j-2}m_\gamma)$$
$$\left.+\frac{\gamma\gamma_t\pi^2 \mathrm{We}^2}{L^2}(D_\alpha^{j-2}\gamma)^2 - \frac{2\gamma^2\pi^2 \mathrm{We}^2}{L^3}L_t(D_\alpha^{j-2}\gamma)^2\right] d\alpha.$$

The bounds for the B terms are very similar to those done earlier and are based on Lemma 4.2. We will not perform these estimates here. We do remark that the term γ_t on the last line can be bounded in terms of the energy because it includes only low derivatives of θ and γ.

We can now see the secondary cancellation. From (4.11) and (4.15) we have

$$S_1^j + S_2^j = \int \frac{2\pi^3 \mathrm{We}}{L^3}H(\gamma^2\chi^\varepsilon D_\alpha^{j-1}\theta)(\Lambda D_\alpha^{j-2}\gamma) + \frac{\gamma^2\pi^2 \mathrm{We}^2}{L^2}(D_\alpha^{j-2}\gamma)\left(D_\alpha^{j-2}\left(\frac{2\pi\chi^\varepsilon\theta_{\alpha\alpha}}{L\mathrm{We}}\right)\right) d\alpha. \tag{4.16}$$

We rewrite the first term using the fact that Λ is self-adjoint and that $H\Lambda = -D_\alpha$. We slightly rearrange the second term.

$$S_1^j + S_2^j = \int \frac{2\pi^3 \mathrm{We}}{L^3}\left[- D_\alpha\Big(\gamma^2(\chi^\varepsilon D_\alpha^{j-1}\theta)\Big)(D_\alpha^{j-2}\gamma) + \gamma^2(\chi^\varepsilon D_\alpha^j\theta)(D_\alpha^{j-2}\gamma)\right] d\alpha.$$

We expand the first D_α using the product rule. The cancellation occurs, and we are left with

$$S_1^j + S_2^j = -\int \frac{4\pi^3\gamma\gamma_\alpha \mathrm{We}}{L^3}[(\chi^\varepsilon D_\alpha^{j-1}\theta)(D_\alpha^{j-2}\gamma)]\, d\alpha.$$

This can clearly be bounded by cE^2 (since r is large enough).

Next we turn our attention to the transport terms, V_1^j and V_2^j. Recall that V_1^j is defined as the first integral on the right-hand side of (4.10). We rewrite V_1^j using the fact that χ^ε is self-adjoint.

$$V_1^j = \int (\chi^\varepsilon D_\alpha^{j-1}\theta)(\chi^\varepsilon D_\alpha^j\theta)\left(\frac{2\pi}{L}(T-\mathbf{W}\cdot\hat{\mathbf{t}})\right) d\alpha.$$

Since the terms with θ form a perfect derivative, we integrate by parts. (This is why we placed χ^ε twice in the corresponding term in (2.26).)

$$V_1^j = -\frac{1}{2}\int (\chi^\varepsilon D_\alpha^{j-1}\theta)^2\left(\frac{2\pi}{L}D_\alpha(T-\mathbf{W}\cdot\hat{\mathbf{t}})\right) d\alpha.$$

This is clearly bounded in terms of the energy.

The transport term V_2^j is a little different because of the presence of the Λ operator.

$$V_2^j = \frac{2\pi^2 \mathrm{We}}{L^2} \int (T - \mathbf{W}\cdot\hat{\mathbf{t}}) \chi^\varepsilon D_\alpha^{j-1} \gamma \Lambda(\chi^\varepsilon D_\alpha^{j-2} \gamma) \, d\alpha. \tag{4.17}$$

In general, if we have an integral of the form $\int g f_\alpha \Lambda(f) \, d\alpha$, we rewrite it using the fact that the adjoint of H is $-H$.

$$\int g f_\alpha \Lambda(f) \, d\alpha = -\int H(g f_\alpha) f_\alpha \, d\alpha.$$

We then pull the factor of g through the Hilbert transform to get

$$\int g f_\alpha \Lambda(f) \, d\alpha = -\int g \Lambda(f) f_\alpha \, d\alpha - \int ([H,g] f_\alpha) f_\alpha \, d\alpha.$$

We add $\int g f_\alpha \Lambda(f) \, d\alpha$ to both sides and we find that

$$\int g f_\alpha \Lambda(f) \, d\alpha = -\frac{1}{2} \int ([H,g] f_\alpha) f_\alpha \, d\alpha.$$

We integrate this by parts and apply Lemma 3.7 to bound this by $c\|g\|_3 \|f\|_0^2$. Making the appropriate choice of f and g, (4.17) can then be bounded in terms of the energy (since r is large enough).

We have now proven that

$$\frac{dE}{dt} \leq C_1 \exp\{C_2 E\}.$$

As we remarked at the beginning of the proof, this proves the theorem. □

5. Existence of solutions. In this section, we demonstrate existence of solutions to the mollified initial value problem. We then show that these solutions converge (as the mollification parameter tends to zero) to a solution of the nonmollified problem. We demonstrate that this solution to the nonmollified problem is unique and has the same regularity as the initial data.

THEOREM 5.1. *Given $(\theta_0, \gamma_0, L_0) \in \mathcal{O}$, there exists a unique solution to the mollified initial value problem for the vortex sheet with surface tension* ((2.26), (2.27), (2.28), (2.29)) *which satisfies the non-self-intersection condition* (3.2) *and the boundedness condition* (3.5). *There exists a time $T^\varepsilon > 0$ such that the solution, $(\theta^\varepsilon, \gamma^\varepsilon, L^\varepsilon)$, is in $C^1([0,T^\varepsilon); \mathcal{O})$. T^ε depends on ε, r, $\mathcal{O}$, and $\|(\theta_0, \gamma_0, L_0)\|_{\mathcal{B}}$.*

Proof. Define $\mathbf{F}_\varepsilon : \mathcal{B} \to \mathcal{B}$ by letting its three components F_ε^1, F_ε^2, and F_ε^3 be the right-hand sides of (2.26), (2.27), and (2.28), respectively. (It is clear that the range of $\mathbf{F}_\varepsilon$ is contained in $\mathcal{B}$ from the presence of the mollifiers in the evolution equations and from Lemma 4.2.) Let two triples, $X = (\theta, \gamma, L)$ and $X' = (\theta', \gamma', L')$, be in $\mathcal{O}$. Since the mollified equations can be viewed as a system of ODEs on a Banach space, we use the Picard theorem to prove existence of solutions (see page 78 of [Zei86]). To do this, we need to show the bound $\|\mathbf{F}_\varepsilon(X) - \mathbf{F}_\varepsilon(X')\|_{\mathcal{B}} \leq c\|X - X'\|_{\mathcal{B}}$. Here, the constant c is allowed to depend on ε.

Using the triangle inequality in the obvious way, we break $\|\mathbf{F}_\varepsilon(X) - \mathbf{F}_\varepsilon(X')\|_{\mathcal{B}}$ into manageable pieces. We begin by looking at the part of $\|F_\varepsilon^1(X) - F_\varepsilon^1(X')\|_{\mathcal{B}}$ which

comes from the first term in (2.26):

$$\left\| \frac{2\pi^2}{L^2} H\chi^\varepsilon(\gamma_\alpha) - \frac{2\pi^2}{L'^2} H\chi^\varepsilon(\gamma'_\alpha) \right\|_{r-1} \leq \left\| \frac{2\pi^2}{L^2} H\chi^\varepsilon(\gamma_\alpha - \gamma'_\alpha) \right\|_{r-1} + \left\| \left(\frac{2\pi^2}{L^2} - \frac{2\pi^2}{L'^2} \right) H\chi^\varepsilon(\gamma'_\alpha) \right\|_{r-1}.$$

Using Lemma 2.1 and the bounds (3.5), we can bound this by $\frac{c}{\varepsilon^2}\|X - X'\|_{\mathcal{B}}$. The second term in (2.26) contributes

$$\left\| \chi^\varepsilon \left(\frac{2\pi}{L}(T - \mathbf{W}\cdot\hat{\mathbf{t}})\chi^\varepsilon\theta_\alpha - \frac{2\pi}{L'}(T' - \mathbf{W}'\cdot\hat{\mathbf{t}}')\chi^\varepsilon\theta'_\alpha \right) \right\|_{r-1} \leq \frac{c}{\varepsilon^{r-1}} \left\| \frac{2\pi}{L}(T - \mathbf{W}\cdot\hat{\mathbf{t}})\chi^\varepsilon\theta_\alpha - \frac{2\pi}{L'}(T' - \mathbf{W}'\cdot\hat{\mathbf{t}}')\chi^\varepsilon\theta'_\alpha \right\|_0. \tag{5.1}$$

We add and subtract twice to bound this by a constant (which depends on ε) times

$$\left\| \left(\frac{2\pi}{L} - \frac{2\pi}{L'} \right)(T - \mathbf{W}\cdot\hat{\mathbf{t}})\chi^\varepsilon\theta_\alpha \right\|_0 + \left\| \frac{2\pi}{L'}((T - \mathbf{W}\cdot\hat{\mathbf{t}}) - (T' - \mathbf{W}'\cdot\hat{\mathbf{t}}'))\chi^\varepsilon\theta_\alpha \right\|_0 + \left\| \frac{2\pi}{L'}(T' - \mathbf{W}'\cdot\hat{\mathbf{t}}')\chi^\varepsilon(\theta_\alpha - \theta'_\alpha) \right\|_0. \tag{5.2}$$

Notice that the first of these three terms involves the difference $L - L'$, and the third involves the difference $\theta - \theta'$. Because of the bounds (3.5), Lemma 4.2, and Lemma 2.1 we can bound these two terms by $c\|X - X'\|_{\mathcal{B}}$. So, we only need to pay special attention to

$$\left\| (T - \mathbf{W}\cdot\hat{\mathbf{t}}) - (T' - \mathbf{W}'\cdot\hat{\mathbf{t}}') \right\|_0.$$

We can easily bound this by $c\|X - X'\|_{\mathcal{B}}$ using (2.6), (2.28), (2.37), and Lemma 3.6. We have now concluded that (5.1) is bounded by $c\|X - X'\|_{\mathcal{B}}$. Continuing in this manner, we get that

$$\|\mathbf{F}_\varepsilon(X) - \mathbf{F}_\varepsilon(X')\|_{\mathcal{B}} \leq c\|X - X'\|_{\mathcal{B}},$$

where the constant depends on ε. Since $\mathbf{F}_\varepsilon$ is Lipschitz, we know that solutions to the mollified equations exist for at least a short time. □

Using Theorem 4.3, we can improve this to demonstrate existence of solutions for some amount of time independent of ε.

COROLLARY 5.2. *There exists a time $T > 0$ such that solutions of the mollified initial value problem, $(\theta^\varepsilon, \gamma^\varepsilon, L^\varepsilon)$, are in $C^1([0,T);\mathcal{O})$. T does not depend on ε.*

Proof. A consequence of the Picard theorem for (autonomous) ODEs on a Banach space (see page 78 of [Zei86]) is that solutions may be continued as long as the solution does not leave an open ball. This is analogous to the standard continuation theorem for ODEs on $\mathbb{R}^n$. Recall that we chose the constants $\bar{c}$ and $\bar{d}$ such that the initial data are in the set $\mathcal{O}$. We need to check that solutions $(\theta^\varepsilon, \gamma^\varepsilon, L^\varepsilon)$ cannot leave the open set $\mathcal{O}$ until some time which does not depend on ε. Note that since $z^\varepsilon(\alpha) - \alpha$ is 2π-periodic, the condition $L^\varepsilon > 2\pi$ is automatically satisfied. By (4.8), we see that if the initial data satisfy (3.5), then the solution satisfies (3.5) as long as

$$-\frac{\ln(e^{-C_2E(0)} - C_1C_2t)}{C_2} < \bar{d}. \tag{5.3}$$

Since we have imposed the condition (5.3) at time $t = 0$ on the data, and since the left-hand side is a continuous function of time, there is a time T_1 such that (5.3) is satisfied for all $t < T_1$.

For the condition (3.2), we know again that it is satisfied by the initial data. Recall that we have given the name q_1 to the quantity inside the absolute value on the right-hand side. If we can show that $|D_t q_1|$ is bounded independently of ε, then that will guarantee that (3.2) is satisfied at least for some amount of time which does not depend on ε. By the Sobolev lemma and Lemma 3.4, we have

$$|D_t q_1|_\infty \ \leq \ c\|D_t q_1\|_1 \ \leq \ c\|D_t z_d^\varepsilon\|_2.$$

This reduces the question to showing that $\|D_t z_d^\varepsilon\|_2$ is bounded independently of ε. We recall the definition $z_d^\varepsilon(\alpha,t) = \int_0^\alpha \frac{L^\varepsilon(t)}{2\pi} e^{i\theta^\varepsilon(\alpha',t)}\,d\alpha'$. Taking the time derivative of this equation, and using the definitions of θ_t^ε and L_t^ε and the bounds (3.5), we see that $\|D_t z_d^\varepsilon\|_2$ is bounded by a constant independent of ε. (It is important that at present we are only attempting to bound a small number of derivatives of z_d^ε.) This proves the corollary. □

We can now prove that the mollified solutions converge in a low norm as the mollification parameter tends to zero. For convenience, we denote by $\mathcal{B}'$ the space $H^1 \times H^{1/2} \times \mathbb{R}$.

THEOREM 5.3. *Solutions $(\theta^\varepsilon, \gamma^\varepsilon, L^\varepsilon)$ of the mollified initial value problem form a Cauchy sequence in $C([0,T];\mathcal{B}')$.*

Proof. Define E_d, the energy function for the difference of two solutions with different values of the mollification parameter, as $E_d^1 + E_d^0 + (L^\varepsilon - L^{\varepsilon'})^2$. Here,

$$E_d^1 = \frac{1}{2}\int (D_\alpha(\theta^\varepsilon - \theta^{\varepsilon'}))^2 + \frac{\pi \mathrm{We}}{L^\varepsilon}(\gamma^\varepsilon - \gamma^{\varepsilon'})\Lambda(\gamma^\varepsilon - \gamma^{\varepsilon'}) + \frac{(\gamma^\varepsilon)^2\pi^2\mathrm{We}^2}{{L^\varepsilon}^2}(\gamma^\varepsilon - \gamma^{\varepsilon'})^2\,d\alpha,$$

$$E_d^0 = \frac{1}{2}\int (\theta^\varepsilon - \theta^{\varepsilon'})\Lambda(\theta^\varepsilon - \theta^{\varepsilon'}) + \frac{\pi \mathrm{We}}{L^\varepsilon}(\gamma^\varepsilon - \gamma^{\varepsilon'})^2 + (\theta^\varepsilon - \theta^{\varepsilon'})^2\,d\alpha.$$

Since both solutions satisfy the same initial conditions, $E_d(0) = 0$. We now wish to estimate how this energy changes over time.

$$\begin{aligned}
\frac{dE_d^1}{dt} &= \int D_\alpha(\theta_t^\varepsilon - \theta_t^{\varepsilon'})D_\alpha(\theta^\varepsilon - \theta^{\varepsilon'}) + \frac{\pi\mathrm{We}}{L^\varepsilon}(\gamma_t^\varepsilon - \gamma_t^{\varepsilon'})\Lambda(\gamma^\varepsilon - \gamma^{\varepsilon'}) \\
&\quad + \left(\frac{\gamma^\varepsilon \pi \mathrm{We}}{L^\varepsilon}\right)^2 (\gamma_t^\varepsilon - \gamma_t^{\varepsilon'})(\gamma^\varepsilon - \gamma^{\varepsilon'}) \\
&\quad + \left[-\frac{\pi\mathrm{We}L_t^\varepsilon}{(L^\varepsilon)^2}(\gamma^\varepsilon - \gamma^{\varepsilon'})\Lambda(\gamma^\varepsilon - \gamma^{\varepsilon'}) + \pi^2\mathrm{We}^2\left(\left(\frac{\gamma^\varepsilon}{L^\varepsilon}\right)^2\right)_t (\gamma^\varepsilon - \gamma^{\varepsilon'})^2\right] \\
&= F_1 + F_2 + F_3 + F_4.
\end{aligned}$$

The estimate of the growth of E_d is very similar to the estimate in Theorem 4.3.

We start with F_1, plugging in for θ_t^ε and $\theta_t^{\varepsilon'}$ from (2.26).

$$F_1 = \int \left(\frac{2\pi^2}{{L^\varepsilon}^2}H(\chi^\varepsilon \gamma_{\alpha\alpha}^\varepsilon) - \frac{2\pi^2}{{L^{\varepsilon'}}^2}H(\chi^{\varepsilon'}\gamma_{\alpha\alpha}^{\varepsilon'})\right)(\theta_\alpha^\varepsilon - \theta_\alpha^{\varepsilon'})\,d\alpha + G_1, \tag{5.4}$$

where G_1 is the remainder. Notice that the integral in (5.4) corresponds to the boxed term in (2.26). We now look at F_2, plugging in for γ_t^ε and $\gamma_t^{\varepsilon'}$ from (2.27).

$$F_2 = \int \frac{\pi \mathrm{We}}{L^\varepsilon}\left(\frac{2\pi}{L^\varepsilon}\frac{\chi^\varepsilon \theta^\varepsilon_{\alpha\alpha}}{\mathrm{We}} - \frac{2\pi}{L^{\varepsilon'}}\frac{\chi^{\varepsilon'}\theta^{\varepsilon'}_{\alpha\alpha}}{\mathrm{We}}\right)\Lambda(\gamma^\varepsilon - \gamma^{\varepsilon'})\, d\alpha \tag{5.5}$$
$$+ \int \frac{\pi \mathrm{We}}{L^\varepsilon}\left(\frac{2\pi^2}{L^{\varepsilon 2}}H(\gamma^{\varepsilon 2}\chi^\varepsilon\theta^\varepsilon_\alpha) - \frac{2\pi^2}{L^{\varepsilon' 2}}H(\gamma^{\varepsilon' 2}\chi^{\varepsilon'}\theta^{\varepsilon'}_\alpha)\right)\Lambda(\gamma^\varepsilon - \gamma^{\varepsilon'})\, d\alpha + G_2,$$

where G_2 is the remainder. Again, the integrals in (5.5) correspond to the boxed terms in (2.27).

In adding F_1 and F_2, we see a cancellation very much like (4.14), the primary cancellation in the proof of Theorem 4.3. To see this, use the name I_1 for the sum of the integral in (5.4) and the first integral in (5.5). We integrate the piece from (5.4) by parts to see

$$I_1 = \int -\left(\frac{2\pi^2}{L^{\varepsilon 2}}\Lambda(\chi^\varepsilon\gamma^\varepsilon) - \frac{2\pi^2}{L^{\varepsilon' 2}}\Lambda(\chi^{\varepsilon'}\gamma^{\varepsilon'})\right)(\theta^\varepsilon_{\alpha\alpha} - \theta^{\varepsilon'}_{\alpha\alpha}) \tag{5.6}$$
$$+ \left(\frac{2\pi^2}{L^{\varepsilon 2}}\chi^\varepsilon\theta^\varepsilon_{\alpha\alpha} - \frac{2\pi^2}{L^\varepsilon L^{\varepsilon'}}\chi^{\varepsilon'}\theta^{\varepsilon'}_{\alpha\alpha}\right)\Lambda(\gamma^\varepsilon - \gamma^{\varepsilon'})\, d\alpha.$$

We add and subtract in order to adjust the factors of L and the operators χ.

$$\begin{aligned} I_1 = &-\int \left(\frac{2\pi^2}{L^{\varepsilon 2}}\Lambda(\chi^\varepsilon\gamma^\varepsilon) - \frac{2\pi^2}{L^{\varepsilon 2}}\Lambda(\chi^\varepsilon\gamma^{\varepsilon'})\right)(\theta^\varepsilon_{\alpha\alpha} - \theta^{\varepsilon'}_{\alpha\alpha})\, d\alpha \\ &-\int \left(\frac{2\pi^2}{L^{\varepsilon 2}}\Lambda(\chi^\varepsilon\gamma^{\varepsilon'}) - \frac{2\pi^2}{L^{\varepsilon' 2}}\Lambda(\chi^\varepsilon\gamma^{\varepsilon'})\right)(\theta^\varepsilon_{\alpha\alpha} - \theta^{\varepsilon'}_{\alpha\alpha})\, d\alpha \\ &-\int \left(\frac{2\pi^2}{L^{\varepsilon' 2}}\Lambda(\chi^\varepsilon\gamma^{\varepsilon'}) - \frac{2\pi^2}{L^{\varepsilon' 2}}\Lambda(\chi^{\varepsilon'}\gamma^{\varepsilon'})\right)(\theta^\varepsilon_{\alpha\alpha} - \theta^{\varepsilon'}_{\alpha\alpha})\, d\alpha \\ &+\int \left(\frac{2\pi^2}{L^{\varepsilon 2}}\chi^\varepsilon\theta^\varepsilon_{\alpha\alpha} - \frac{2\pi^2}{L^{\varepsilon 2}}\chi^\varepsilon\theta^{\varepsilon'}_{\alpha\alpha}\right)\Lambda(\gamma^\varepsilon - \gamma^{\varepsilon'})\, d\alpha \\ &+\int \left(\frac{2\pi^2}{L^{\varepsilon 2}}\chi^\varepsilon\theta^{\varepsilon'}_{\alpha\alpha} - \frac{2\pi^2}{L^\varepsilon L^{\varepsilon'}}\chi^\varepsilon\theta^{\varepsilon'}_{\alpha\alpha}\right)\Lambda(\gamma^\varepsilon - \gamma^{\varepsilon'})\, d\alpha \\ &+\int \left(\frac{2\pi^2}{L^\varepsilon L^{\varepsilon'}}\chi^\varepsilon\theta^{\varepsilon'}_{\alpha\alpha} - \frac{2\pi^2}{L^\varepsilon L^{\varepsilon'}}\chi^{\varepsilon'}\theta^{\varepsilon'}_{\alpha\alpha}\right)\Lambda(\gamma^\varepsilon - \gamma^{\varepsilon'})\, d\alpha. \end{aligned} \tag{5.7}$$

The first and fourth of these integrals cancel exactly because χ^ε is self-adjoint. We look at the second integral; we integrate it by parts, and since r is large enough, we use the uniform bound of Theorem 4.3 to bound $\|\chi^\varepsilon\gamma^{\varepsilon'}_{\alpha\alpha}\|$ by a constant. Thus, the second integral can be bounded by $c|L^\varepsilon - L^{\varepsilon'}|\|\theta^\varepsilon - \theta^{\varepsilon'}\|_1$, which can in turn be bounded by cE_d. Similarly, the fifth integral can be bounded by cE_d.

The third and sixth integrals involve differences of χ^ε and $\chi^{\varepsilon'}$. We integrate the third integral by parts, and using the uniform bound of Theorem 4.3, we see that it is bounded by $c\|\theta^\varepsilon - \theta^{\varepsilon'}\|_1\|\chi^\varepsilon\gamma^{\varepsilon'} - \chi^{\varepsilon'}\gamma^{\varepsilon'}\|_2$. Using Lemma 2.2 together with Theorem 4.3, we can bound this by $c\max\{\varepsilon,\varepsilon'\}E_d^{1/2}$. Similarly, the sixth integral can be bounded by $c\max\{\varepsilon,\varepsilon'\}E_d^{1/2}$.

We leave out the remaining details of the estimate of E_d because they are very similar to what we have just done and to the estimate in the proof of Theorem 4.3.

More details can be found in [Amb02]. The result is

$$\frac{dE_d}{dt} \;\leq\; cE_d + c\max\{\varepsilon,\varepsilon'\}E_d^{1/2}.$$

This can be restated as $\frac{dE_d^{1/2}}{dt} \leq cE_d^{1/2} + c\max\{\varepsilon,\varepsilon'\}$. Since E_d is initially zero, we solve the differential inequality to see that

$$E_d^{1/2} \;\leq\; \max\{\varepsilon,\varepsilon'\}(e^{ct}-1). \tag{5.8}$$

We wish to relate this to the norms of θ, γ, and L. From the definition of E_d, it is apparent that $\|(\theta^\varepsilon - \theta^{\varepsilon'}, \gamma^\varepsilon - \gamma^{\varepsilon'}, L^\varepsilon - L^{\varepsilon'})\|_{\mathcal{B}'} \leq CE_d^{1/2}$. Using (5.8) with this, and taking the supremum in time, we have

$$\sup_{0\leq t\leq T} \|(\theta^\varepsilon - \theta^{\varepsilon'}, \gamma^\varepsilon - \gamma^{\varepsilon'}, L^\varepsilon - L^{\varepsilon'})\|_{\mathcal{B}'} \;\leq\; C\max\{\varepsilon,\varepsilon'\}(e^{cT}-1).$$

Thus, solutions do form a Cauchy sequence in $C([0,T];H^1\times H^{1/2}\times\mathbb{R})$. □

We now know that the solutions of the mollified problem, $(\theta^\varepsilon,\gamma^\varepsilon,L^\varepsilon)$, approach a limit as $\varepsilon\to 0$. Call this limit (θ,γ,L). We prove that (θ,γ,L) is a solution to the nonmollified initial value problem, that this solution is unique, and that the solution is in the space $C([0,T];\mathcal{O})$.

In order to do this, we first prove that the limit of the mollified solutions is weakly continuous in time. That is, we will show that $\theta\in C_W([0,T];H^{r-1})$ and $\gamma\in C_W([0,T];H^{r-3/2})$. For $s\in\mathbb{R}$, the statement $u\in C_W([0,T];H^s)$ means that $u(t)\in H^s$ for all $t\in[0,T]$, and for all $\psi\in H^{-s}$, the duality pairing $\langle u(t),\psi\rangle$ is a continuous function of time.

THEOREM 5.4. *The limit of the solutions to the mollified initial value problem found in the previous theorem, (θ,γ,L), is in $C([0,T];H^m\times H^{m'}\times\mathbb{R})$ for all $1\leq m< r-1$ and $1/2\leq m'<r-3/2$, and it is also in $C_W([0,T];\mathcal{O})$. Furthermore, (θ,γ,L) is a solution to the exact evolution equations for the vortex sheet with surface tension* ((2.22), (2.23), (2.24)).

Proof. We have found a triple $(\theta,\gamma,L)\in C([0,T];H^1\times H^{1/2}\times\mathbb{R})$, which is the limit of the solutions of the mollified equations, $(\theta^\varepsilon,\gamma^\varepsilon,L^\varepsilon)$. Recall that the conclusion of Theorem 4.3 is that

$$\sup_{0\leq t\leq T}\|(\theta^\varepsilon,\gamma^\varepsilon,L^\varepsilon)\|_{\mathcal{B}}\leq C.$$

Since $\mathcal{B}$ is a Hilbert space, its unit ball is weakly compact. Thus, $(\theta^\varepsilon,\gamma^\varepsilon,L^\varepsilon)$ must have a weak limit in $\mathcal{B}$. Since $\mathcal{B}\subset H^1\times H^{1/2}\times\mathbb{R}$, this weak limit must be (θ,γ,L). Thus, $(\theta,\gamma,L)\in\mathcal{B}$.

We can now apply Lemma 3.3 to the differences $\theta^\varepsilon-\theta$ and $\gamma^\varepsilon-\gamma$. The conclusion is that

$$\|\theta^\varepsilon-\theta\|_m \;\leq\; C\|\theta^\varepsilon-\theta\|_0^{1-\frac{m}{r-1}}\|\theta^\varepsilon-\theta\|_{r-1}^{\frac{m}{r-1}},$$

$$\|\gamma^\varepsilon-\gamma\|_{m'} \;\leq\; C\|\gamma^\varepsilon-\gamma\|_0^{1-\frac{m'}{r-3/2}}\|\gamma^\varepsilon-\gamma\|_{r-3/2}^{\frac{m'}{r-3/2}}.$$

Notice that the right-hand sides of these inequalities go to zero uniformly on $[0,T]$. This implies $(\theta,\gamma,L)\in H^m\times H^{m'}\times\mathbb{R}$ for all $1\leq m<r-1$ and all $\frac{1}{2}\leq m'<r-\frac{3}{2}$.

This regularity is enough to conclude that (θ, γ, L) is a solution to the nonmollified system (2.22), (2.23), (2.24).

We now show that $(\theta, \gamma, L) \in C_W([0,T]; \mathcal{B})$. Let $\eta > 0$ be given. Let $\phi \in H^{-(r-1)}$ be given. For any m satisfying $1 \le m < r-1$, let $\tilde{\phi} \in H^{-m}$ be given such that

$$\left\| \phi - \tilde{\phi} \right\|_{-(r-1)} \le \frac{\eta}{3}. \tag{5.9}$$

We know that such a $\tilde{\phi}$ can be found since H^s is dense in $H^{s'}$ whenever $s' < s$. We show that the difference of the duality pairings of θ and θ^ε with ϕ can be made small uniformly in time:

$$\langle \phi, \theta^\varepsilon \rangle - \langle \phi, \theta \rangle = \langle \phi - \tilde{\phi}, \theta^\varepsilon \rangle + \langle \tilde{\phi}, \theta^\varepsilon - \theta \rangle + \langle \tilde{\phi} - \phi, \theta \rangle. \tag{5.10}$$

The two of these involving $\phi - \tilde{\phi}$ can be bounded by $\eta/3$ using (5.9) and the uniform bound on θ and θ^ε in H^{r-1}. For the term involving $\theta - \theta^\varepsilon$, we choose ε small enough so that $\|\theta - \theta^\varepsilon\|_m \le \frac{\eta}{3}$. Thus, (5.10) is bounded by η. Since η was arbitrary and since these bounds are all uniform in time, and since θ^ε is in $C([0,T]; H^{r-1})$, we conclude that $\theta \in C_W([0,T]; H^{r-1})$. A similar argument applies to γ. Thus, we conclude that $(\theta, \gamma, L) \in C_W([0,T]; \mathcal{B})$. □

Before proving the highest regularity theorem for solutions of the nonmollified problem, we first need uniqueness of these solutions.

THEOREM 5.5. *Solutions $(\theta, \gamma, L) \in \mathcal{B}$ of the exact initial value problem for the vortex sheet with surface tension* ((2.22), (2.23), (2.24), (2.25)) *are unique.*

We do not prove this theorem here. The proof is very similar to the proof of Theorem 5.3 because both proofs require that the difference of two solutions be estimated. This theorem is simpler than Theorem 5.3 because of the absence of the χ operators in the nonmollified problem. The proof is contained in [Amb02].

We can now demonstrate our highest regularity theorem: $(\theta, \gamma, L) \in C([0,T]; \mathcal{B})$. Because we already know that $(\theta^\varepsilon, \gamma^\varepsilon, L^\varepsilon)$ converges to (θ, γ, L) in the weak topology, we need to show only that $\|(\theta, \gamma, L)\|_{\mathcal{B}}$ is continuous in time. To do this, we make use of the time-reversibility of our equations.

THEOREM 5.6. *Solutions (θ, γ, L) of the exact initial value problem for the vortex sheet with surface tension* ((2.22), (2.23), (2.24), (2.25)) *are in $C([0,T]; \mathcal{B})$.*

Since $(\theta, \gamma, L) \in C_W([0,T]; \mathcal{B})$, we see that L is continuous and, by Fatou's lemma,

$$\|\theta_0\|_{r-1}^2 \le \liminf_{t \to 0^+} \|\theta(t)\|_{r-1}^2, \tag{5.11}$$

$$\|\gamma_0\|_{r-3/2}^2 \le \liminf_{t \to 0^+} \|\gamma(t)\|_{r-3/2}^2. \tag{5.12}$$

Also, by (4.8), we have the estimate $\limsup_{t \to 0^+} E(t) \le E(0)$. Since $\liminf(a) + \liminf(b) \le \liminf(a+b)$, we can combine (5.11) and (5.12) to get $\liminf_{t \to 0^+} E(t) \ge E(0)$. Thus, the energy is right-continuous at the initial time. To see that $\|(\theta, \gamma, L)\|_{\mathcal{B}}$ is also right-continuous at the initial time, we look at the various parts of the energy. First, we consider E_3^j, where j is at most r. Its definition is

$$E_3^j(t) = \frac{1}{2} \int_0^{2\pi} \frac{\gamma^2(\alpha,t)\pi^2 \mathrm{We}^2}{L^2(t)} (D_\alpha^{j-2} \gamma(\alpha,t))^2 \, d\alpha.$$

Since we already know that $\gamma \in C([0,T]; H^{r-2})$, we can conclude that E_3^j is continuous. Similarly, we know that E_1^j and E_2^j are continuous for j at most $r-1$. Also,

we already know that L and $E^0 = \|\theta\|_0^2 + \|\gamma\|_0^2$ are continuous. This tells us that the difference between E and the parts of the energy which we know to be individually continuous is continuous. In particular, we now know that

$$E_1^r(t) + E_2^r(t) = \frac{1}{2}\int_0^{2\pi} (D_\alpha^{r-1}\theta(\alpha,t))^2 + \frac{\pi \mathrm{We}}{L(t)}(D_\alpha^{r-2}\gamma(\alpha,t))\Lambda(D_\alpha^{r-2}\gamma(\alpha,t))\, d\alpha$$

is continuous. Combining this with the lower-order terms, it means

$$\|\theta\|_{r-1}^2 + \frac{\pi \mathrm{We}}{L}\|\gamma\|_{r-3/2}^2 \tag{5.13}$$

is right-continuous at the initial time. Notice that (5.11) and (5.12) imply that $\|\theta\|_{r-1}$ and $\|\gamma\|_{r-3/2}$ are right lower semicontinuous at the initial time, and we already know that L is right-continuous at the initial time. Since the sum (5.13) is right-continuous at the initial time, this implies that $\|\theta\|_{r-1}$ and $\|\gamma\|_{r-3/2}$ are each right-continuous at the initial time.

Given a time $T' \in (0,T)$, we can interpret T' as a new initial time. We know that $(\theta(T'), \gamma(T'), L(T'))$ satisfies conditions (3.2) and (3.5). Thus, we could repeat our arguments for existence of mollified solutions starting at time T', and we would find that a solution to the nonmollified equations exists on some time interval around T'. By Theorem 5.5, the solution we found starting from $t = 0$ and the solution we found starting from $t = T'$ must be the same. Using the argument by which we showed that solutions are right-continuous at $t = 0$, we see that solutions are right-continuous at $t = T'$. Since all of our analysis works with time reversed, we can also conclude that solutions are left-continuous at $t = T'$. Similarly, we can show that solutions are also left-continuous at the final time, $t = T$. We conclude that $(\theta, \gamma, L) \in C([0,T], \mathcal{B})$. ∎

We also have the following theorem that demonstrates that the solutions of the initial value problem are continuous in a low norm of the initial data.

THEOREM 5.7. *If $(\theta, \gamma, L) \in \mathcal{B}$ and $(\theta', \gamma', L') \in \mathcal{B}$ are both solutions of the exact initial value problem for the vortex sheet with surface tension for the interval of time $[0,T]$, with the corresponding initial data $(\theta_0, \gamma_0, L_0) \in \mathcal{O}$ and $(\theta_0', \gamma_0', L_0') \in \mathcal{O}$, then*

$$\sup_{0\le t\le T} \|(\theta - \theta', \gamma - \gamma', L - L')\|_{\mathcal{B}'} \;\le\; \|(\theta_0 - \theta_0', \gamma_0 - \gamma_0', L_0 - L_0')\|_{\mathcal{B}'} \exp\{cT\}.$$

We do not include the proof of this theorem since it is very similar to those of Theorems 5.3 and 5.5.

6. The case of fluids with different densities. Thus far, we have considered only the case in which the upper and lower fluids both have the same density. Our method also proves well-posedness in the general case of arbitrary densities. We define the Atwood ratio, At, to be

$$\mathrm{At} = \frac{\rho_1 - \rho_2}{\rho_1 + \rho_2},$$

where ρ_1 is the density of the lower fluid and ρ_2 is the density of the upper fluid.

This section of the paper consists of three subsections. In the first, we give the evolution equation for γ in the two-density case and we rewrite it; this is similar to what we have already done for the density-matched case in section 2. This new evolution equation is actually an integral equation for γ_t. In the second subsection, we discuss the solvability of the integral equation and give estimates for quantities related to the integral operator. In the final subsection, we state the existence theorem for the vortex sheet in the two-density case.

6.1. The new γ_t equation. In considering $\mathrm{At} \neq 0$, the evolution equations for θ and L do not change. The new evolution equation for γ is

$$\gamma_t = \frac{2\pi\theta_{\alpha\alpha}}{L\mathrm{We}} + \frac{2\pi}{L} D_\alpha \left((T - \mathbf{W}\cdot\hat{\mathbf{t}})\gamma \right) - 2\mathrm{At} \left(\frac{L}{2\pi}\mathbf{W}_t \cdot \hat{\mathbf{t}} + \frac{\pi^2}{L^2}\gamma\gamma_\alpha - (T - \mathbf{W}\cdot\hat{\mathbf{t}})\mathbf{W}_\alpha \cdot \hat{\mathbf{t}} \right). \tag{6.1}$$

If $\mathrm{At} = 0$, then this is just the same equation as (2.7). This equation for γ_t can be derived by a variation on the argument presented in [BMO82]; this derivation is included in [Amb02]. There is a small error in this form of the equation in [HLS94] and [HLS97], but it appears correctly in [CH96].

We need to rewrite this equation, similarly to the way we rewrote all of the evolution equations in section 2. We focus first on expanding the factor which multiplies the Atwood ratio. We begin by writing out $\frac{L}{2\pi}\mathbf{W}_t \cdot \hat{\mathbf{t}}$, with the goal of separating it into important and routine pieces:

$$\begin{aligned} \frac{L}{2\pi}\mathbf{W}_t \cdot \hat{\mathbf{t}} &= \mathrm{Re}\left(z_\alpha \Phi(\mathbf{W}_t)^*\right) \\ &= \mathrm{Re}\left\{ \frac{z_\alpha(\alpha)}{4\pi i}\mathrm{PV}\int \gamma_t(\alpha')\cot\frac{1}{2}(z(\alpha) - z(\alpha'))\, d\alpha' \right\} \\ &\quad - \mathrm{Re}\left\{ \frac{z_\alpha(\alpha)}{4\pi i}\mathrm{PV}\int \frac{\gamma(\alpha')}{2}(z_t(\alpha) - z_t(\alpha'))\csc^2\frac{1}{2}(z(\alpha) - z(\alpha'))\, d\alpha' \right\}. \end{aligned} \tag{6.2}$$

We make the definition of an integral operator, $\mathcal{J}[z]$, to be

$$\mathcal{J}[z]f(\alpha) = -\mathrm{PV}\int f(\alpha')\mathrm{Re}\left(i z_\alpha(\alpha)\cot\frac{1}{2}(z(\alpha) - z(\alpha')) \right) d\alpha'.$$

Notice that we can write $\mathcal{J}[z]$ as

$$\mathcal{J}[z]f(\alpha) = -2\pi\mathrm{Re}\left(i z_\alpha H\left(\frac{f}{z_\alpha}\right)\right) + 4\pi\mathrm{Re}(z_\alpha \mathcal{K}[z]f(\alpha)), \tag{6.3}$$

with $\mathcal{K} = \mathcal{K}_1 + \mathcal{K}_2$ defined in (2.35) and (2.36). We can now write

$$\frac{L}{2\pi}\mathbf{W}_t \cdot \hat{\mathbf{t}} = \frac{1}{4\pi}\mathcal{J}[z]\gamma_t + R_3,$$

where R_3 is the second term on the right-hand side of (6.2). Estimates for $\mathcal{J}[z]$ are the subject of the next subsection.

We will now expand R_3. As we have done many times in section 2, we multiply and divide the integrand in R_3 by $z_\alpha(\alpha')$ and then recognize that part of the integrand is an α'-derivative. We integrate by parts to get

$$\begin{aligned} R_3 &= -\mathrm{Re}\left\{ \frac{z_\alpha(\alpha)}{4\pi i}\mathrm{PV}\int \frac{\gamma(\alpha')}{z_\alpha(\alpha')} z_{\alpha t}(\alpha')\cot\frac{1}{2}(z(\alpha) - z(\alpha'))\, d\alpha' \right\} \\ &\quad + \mathrm{Re}\left\{ \frac{z_\alpha(\alpha)}{4\pi i}\mathrm{PV}\int D_{\alpha'}\left(\frac{\gamma(\alpha')}{z_\alpha(\alpha')}\right)(z_t(\alpha) - z_t(\alpha'))\cot\left(\frac{1}{2}(z(\alpha) - z(\alpha'))\right) d\alpha' \right\}. \end{aligned} \tag{6.4}$$

We write $R_3 = R_4 + R_5$, with R_4 equal to the first term on the right-hand side of (6.4) and R_5 equal to the second. We can rewrite R_5 as

$$R_5 = \mathrm{Re}\left\{ \frac{z_\alpha}{2i}[H, z_t]\left(\frac{1}{z_\alpha} D_\alpha\left(\frac{\gamma}{z_\alpha}\right)\right)\right\} + \mathrm{Re}\left\{ z_\alpha z_t \mathcal{K}[z]\left(D_\alpha\left(\frac{\gamma}{z_\alpha}\right)\right) - z_\alpha \mathcal{K}[z]\left(z_t D_\alpha\left(\frac{\gamma}{z_\alpha}\right)\right)\right\}. \tag{6.5}$$

R_4 is more important, and we continue to expand it. To begin to do this, we write

$$R_4 = -\mathrm{Re}\left\{ \frac{z_\alpha}{2i} H\left(\frac{\gamma}{z_\alpha^2} z_{\alpha t}\right)\right\} - \mathrm{Re}\left\{ z_\alpha \mathcal{K}[z]\left(\frac{\gamma}{z_\alpha} z_{\alpha t}\right)\right\}.$$

In order to understand this, we look at $z_{\alpha t}$:

$$z_{\alpha t} = \left(\frac{L}{2\pi} e^{i\theta}\right)_t = \frac{L_t}{2\pi} e^{i\theta} + \frac{L}{2\pi} i\theta_t e^{i\theta} = \frac{L_t}{L} z_\alpha + i\theta_t z_\alpha. \tag{6.6}$$

Using (6.6), and substituting from (2.22) for θ_t, we have

$$\begin{aligned} R_4 = &- \mathrm{Re}\left\{ z_\alpha \mathcal{K}[z]\left(\frac{\gamma}{z_\alpha} z_{\alpha t}\right)\right\} - \mathrm{Re}\left\{ \frac{z_\alpha}{2i} H\left(\frac{L_t \gamma}{L z_\alpha}\right)\right\} \\ &- \mathrm{Re}\left\{ \frac{z_\alpha}{2} H\left(\frac{2\pi^2 \gamma H(\gamma_\alpha)}{L^2 z_\alpha}\right)\right\} - \mathrm{Re}\left\{ \frac{z_\alpha}{2} H\left(\frac{2\pi\gamma(T - \mathbf{W}\cdot\hat{\mathbf{t}})\theta_\alpha}{L z_\alpha}\right)\right\} \\ &- \mathrm{Re}\left\{ \frac{z_\alpha}{2} H\left(\frac{2\pi\gamma \mathbf{m}\cdot\hat{\mathbf{n}}}{L z_\alpha}\right)\right\}. \end{aligned} \tag{6.7}$$

We still need to rewrite the third and fourth of these terms so that we can see more clearly the most important part of R_4. We begin by bringing γ/z_α outside of the first Hilbert transform in the third term:

$$-\mathrm{Re}\left\{ \frac{z_\alpha}{2} H\left(\frac{2\pi^2 \gamma H(\gamma_\alpha)}{L^2 z_\alpha}\right)\right\} = \frac{\pi^2}{L^2}\gamma\gamma_\alpha - \mathrm{Re}\left\{ \frac{\pi^2 z_\alpha}{L^2}\left[H, \frac{\gamma}{z_\alpha}\right](H(\gamma_\alpha))\right\}. \tag{6.8}$$

Similarly, in the fourth term, we bring $(T - \mathbf{W}\cdot\hat{\mathbf{t}})/z_\alpha$ outside of the Hilbert transform:

$$-\mathrm{Re}\left\{ \frac{z_\alpha}{2} H\left(\frac{2\pi\gamma(T - \mathbf{W}\cdot\hat{\mathbf{t}})\theta_\alpha}{L z_\alpha}\right)\right\} = \frac{\pi(T - \mathbf{W}\cdot\hat{\mathbf{t}})}{L} H(\gamma\theta_\alpha) - \mathrm{Re}\left\{ \frac{\pi z_\alpha}{L}\left[H, \frac{(T - \mathbf{W}\cdot\hat{\mathbf{t}})}{z_\alpha}\right](\gamma\theta_\alpha)\right\}. \tag{6.9}$$

Finally, this allows us to write

$$\frac{L}{2\pi}\mathbf{W}_t\cdot\hat{\mathbf{t}} = \frac{1}{4\pi}\mathcal{J}[z]\gamma_t - \frac{\pi(T - \mathbf{W}\cdot\hat{\mathbf{t}})}{L} H(\gamma\theta_\alpha) + \frac{\pi^2}{L^2}\gamma\gamma_\alpha + R_5 + R_6, \tag{6.10}$$

where R_6 is a collection of terms from the right-hand sides of (6.7), (6.8), and (6.9). To be specific, R_6 is defined by

$$\begin{aligned} R_6 = &-\mathrm{Re}\left\{ z_\alpha \mathcal{K}[z]\left(\frac{\gamma}{z_\alpha} z_{\alpha t}\right)\right\} - \mathrm{Re}\left\{ \frac{z_\alpha}{2i} H\left(\frac{L_t\gamma}{L z_\alpha}\right)\right\} - \mathrm{Re}\left\{ \frac{z_\alpha}{2} H\left(\frac{2\pi\gamma\mathbf{m}\cdot\hat{\mathbf{n}}}{L z_\alpha}\right)\right\} \\ &- \mathrm{Re}\left\{ \frac{\pi^2 z_\alpha}{L^2}\left[H, \frac{\gamma}{z_\alpha}\right](H(\gamma_\alpha))\right\} - \mathrm{Re}\left\{ \frac{\pi z_\alpha}{L}\left[H, \frac{(T - \mathbf{W}\cdot\hat{\mathbf{t}})}{z_\alpha}\right](\gamma\theta_\alpha)\right\}. \end{aligned} \tag{6.11}$$

Since our present goal is to rewrite equation (6.1), we also would like to expand $(T - \mathbf{W}\cdot\hat{\mathbf{t}})\mathbf{W}_\alpha\cdot\hat{\mathbf{t}}$. We can do this using (2.20):

$$(T - \mathbf{W}\cdot\hat{\mathbf{t}})\mathbf{W}_\alpha\cdot\hat{\mathbf{t}} = -\frac{\pi(T-\mathbf{W}\cdot\hat{\mathbf{t}})}{L}H(\gamma\theta_\alpha) + (T-\mathbf{W}\cdot\hat{\mathbf{t}})m\cdot\hat{\mathbf{t}}. \tag{6.12}$$

We can now use (6.10) and (6.12) to expand the factor which multiplies 2At in (6.1); notice the cancellation which occurs between $H(\gamma\theta_\alpha)$ terms:

$$\begin{aligned}\frac{L}{2\pi}\mathbf{W}_t\cdot\hat{\mathbf{t}} + \frac{\pi^2}{L^2}\gamma\gamma_\alpha - (T-\mathbf{W}\cdot\hat{\mathbf{t}})\mathbf{W}_\alpha\cdot\hat{\mathbf{t}} \\ = \left(\frac{1}{4\pi}\mathcal{J}[z]\gamma_t + -\frac{\pi(T-\mathbf{W}\cdot\hat{\mathbf{t}})}{L}H(\gamma\theta_\alpha) + \frac{\pi^2}{L^2}\gamma\gamma_\alpha + R_5 + R_6\right) + \frac{\pi^2}{L^2}\gamma\gamma_\alpha \\ - (T-\mathbf{W}\cdot\hat{\mathbf{t}})\left(-\frac{\pi}{L}H(\gamma\theta_\alpha) + \mathbf{m}\cdot\hat{\mathbf{t}}\right) \\ = \frac{1}{4\pi}\mathcal{J}[z]\gamma_t + 2\frac{\pi^2}{L^2}\gamma\gamma_\alpha + R_5 + R_6 - (T-\mathbf{W}\cdot\hat{\mathbf{t}})\mathbf{m}\cdot\hat{\mathbf{t}}.\end{aligned} \tag{6.13}$$

Before we give our final rewritten form of (6.1), we recall how we previously rewrote the term $\frac{2\pi}{L}D_\alpha((T-\mathbf{W}\cdot\hat{\mathbf{t}})\gamma)$:

$$\begin{aligned}\frac{2\pi}{L}D_\alpha((T-\mathbf{W}\cdot\hat{\mathbf{t}})\gamma) = \frac{2\pi^2}{L^2}H\left(\gamma^2\theta_\alpha\right) + \frac{2\pi}{L}(T-\mathbf{W}\cdot\hat{\mathbf{t}})\gamma_\alpha \\ + \frac{\gamma(L_t - 2\pi\mathbf{m}\cdot\hat{\mathbf{t}})}{L} - \frac{2\pi^2}{L^2}[H,\gamma](\gamma\theta_\alpha).\end{aligned} \tag{6.14}$$

Combining (6.1) with (6.13) and (6.14), we have

$$\gamma_t = -\frac{\mathrm{At}}{2\pi}\mathcal{J}[z]\gamma_t + \frac{2\pi}{L}\frac{\theta_{\alpha\alpha}}{\mathrm{We}} + \frac{2\pi^2}{L^2}H\left(\gamma^2\theta_\alpha\right) + \left(\frac{2\pi}{L}(T-\mathbf{W}\cdot\hat{\mathbf{t}}) - \frac{4\mathrm{At}\pi^2}{L^2}\right)\gamma_\alpha + m_{\mathrm{At}}, \tag{6.15}$$

where m_{At} is given by

$$m_{\mathrm{At}} = \frac{\gamma(L_t - 2\pi\mathbf{m}\cdot\hat{\mathbf{t}})}{L} - \frac{2\pi^2}{L^2}[H,\gamma](\gamma\theta_\alpha) - 2\mathrm{At}\left[R_5 + R_6 - (T-\mathbf{W}\cdot\hat{\mathbf{t}})\mathbf{m}\cdot\hat{\mathbf{t}}\right]. \tag{6.16}$$

We now introduce mollifiers to the new γ_t equation:

$$\begin{aligned}\gamma_t^\varepsilon = -\frac{\mathrm{At}}{2\pi}\mathcal{J}[z^\varepsilon]\gamma_t^\varepsilon + \frac{2\pi}{L^\varepsilon}\frac{\chi^\varepsilon\theta^\varepsilon_{\alpha\alpha}}{\mathrm{We}} + \frac{2\pi^2}{{L^\varepsilon}^2}H\left((\gamma^\varepsilon)^2\chi^\varepsilon\theta^\varepsilon_\alpha\right) \\ + \chi^\varepsilon\left(\left(\frac{2\pi}{L^\varepsilon}(T^\varepsilon - \mathbf{W}^\varepsilon\cdot\hat{\mathbf{t}}^\varepsilon) - \frac{4\mathrm{At}\pi^2}{{L^\varepsilon}^2}\right)\chi^\varepsilon\gamma^\varepsilon_\alpha\right) + m^\varepsilon_{\mathrm{At}}.\end{aligned} \tag{6.17}$$

The reader may want to compare this equation with (2.27). Notice that the operator $\mathcal{J}$ is not mollified, but it does make use of the mollified curve z^ε. The term $m^\varepsilon_{\mathrm{At}}$ is defined by replacing all of the quantities on the right-hand side of (6.16) with their mollified counterparts.

6.2. Estimates for $\mathcal{J}[z^\varepsilon]$. Now that we have given the new mollified evolution equation, we are ready to begin estimating the new system. In this subsection, we give two lemmas relating to the integral operator $\mathcal{J}[z^\varepsilon]$.

LEMMA 6.1. *The operator* $\left(I + \frac{\mathrm{At}}{2\pi}\mathcal{J}[z^\varepsilon]\right)^{-1}$ *is bounded from* H^0 *to* H^0, *where* I *is the identity operator.*

We do not include a proof here; see [BMO82] or the discussion in the appendix of [BHL96]. In the next lemma, the energy function E appears. The definition of the energy is the same as we used in the density-matched case; see (3.3) and (3.4).

LEMMA 6.2. *If* $r \geq 4$, $\theta^\varepsilon \in H^{r-1}$, *and* $\gamma^\varepsilon \in H^{r-3/2}$, *then* $\mathcal{J}[z^\varepsilon](\gamma_t^\varepsilon) \in H^{r-1}$. *Furthermore,* $\|\mathcal{J}[z^\varepsilon](\gamma_t^\varepsilon)\|_{r-1} \leq C_1 \exp\{C_2 E\}$.

Proof. We begin by making the definition

$$\tau^\varepsilon = \frac{2\pi}{L^\varepsilon}\frac{\chi^\varepsilon\theta^\varepsilon_{\alpha\alpha}}{\mathrm{We}} + \frac{2\pi^2}{L^{\varepsilon 2}}H\left((\gamma^\varepsilon)^2\chi^\varepsilon\theta^\varepsilon_\alpha\right) + \chi^\varepsilon\left(\left(\frac{2\pi}{L^\varepsilon}(T^\varepsilon - \mathbf{W}^\varepsilon\cdot\hat{\mathbf{t}}^\varepsilon) - \frac{4\mathrm{At}\pi^2}{L^{\varepsilon 2}}\right)\chi^\varepsilon\gamma^\varepsilon_\alpha\right) + m^\varepsilon_{\mathrm{At}}. \tag{6.18}$$

This allows us to write (6.17) as

$$\gamma^\varepsilon_t = \left(I + \frac{\mathrm{At}}{2\pi}\mathcal{J}[z^\varepsilon]\right)^{-1}\tau^\varepsilon.$$

Since we can see from (6.18) that the most singular term in τ^ε is proportional to $\theta^\varepsilon_{\alpha\alpha}$ and we have assumed $\theta^\varepsilon \in H^{r-1}$, we conclude $\tau^\varepsilon \in H^{r-3} \subset H^0$. Lemma 6.1 now implies that $\gamma^\varepsilon_t \in H^0$.

Next, we use (6.3) to show that $\mathcal{J}[z^\varepsilon](\gamma^\varepsilon_t)$ is in H^{r-2}. By pulling $1/z_\alpha$ out of the Hilbert transform, we rewrite the first part of (6.3):

$$-2\pi\mathrm{Re}\left(iz^\varepsilon_\alpha H\left(\frac{f}{z^\varepsilon_\alpha}\right)\right) = -2\pi\mathrm{Re}\left(iH(f)\right) - 2\pi\mathrm{Re}\left(iz^\varepsilon_\alpha\left[H,\frac{1}{z^\varepsilon_\alpha}\right]f\right).$$

Since the first part of the right-hand side is zero, we have

$$\mathcal{J}[z^\varepsilon](\gamma^\varepsilon_t) = -2\pi\mathrm{Re}\left(iz^\varepsilon_\alpha\left[H,\frac{1}{z^\varepsilon_\alpha}\right]\gamma^\varepsilon_t\right) + 4\pi\left(\mathrm{Re}(z^\varepsilon_\alpha\mathcal{K}[z^\varepsilon]\gamma^\varepsilon_t(\alpha))\right). \tag{6.19}$$

We use Lemma 3.7 on the first of these terms and Lemma 3.5 on the second; since $z^\varepsilon_\alpha \in H^{r-1}$, this demonstrates that $\mathcal{J}[z^\varepsilon](\gamma_t) \in H^{r-2}$. Since $\gamma^\varepsilon_t = -\frac{\mathrm{At}}{2\pi}\mathcal{J}[z^\varepsilon](\gamma^\varepsilon_t) + \tau^\varepsilon$, we see that $\gamma^\varepsilon_t \in H^{r-3}$.

In estimating (6.19) just now, we used that $\gamma^\varepsilon_t \in H^0$. Now that we know $\gamma^\varepsilon_t \in H^{r-3}$, we can estimate (6.19) again. By again using Lemma 3.5, but this time using Corollary 3.8 in place of Lemma 3.7, we conclude that $\mathcal{J}[z^\varepsilon](\gamma_t) \in H^{r-1}$. ∎

6.3. The new existence theorem. Our goal is to prove that solutions to the evolution equations exist for the new system (i.e., replacing the old γ_t equation with (6.1)). To prove this we need to modify the proofs to the previous theorems, but the definition of the energy function is exactly the same as it was before. In the earlier case, we needed two kinds of estimates: the first kind was a bound on the growth of the energy. In Lemma 6.3 below, we demonstrate that the new term m_{At} is bounded in terms of the energy. The second kind of estimate we needed involved differences between two solutions—for example, in the proof of Theorem 5.5. We will not prove

the second kind for the new γ_t equation, but the proofs are straightforward and very much like what we have already done.

LEMMA 6.3. *The term m_{At} is bounded in terms of the energy:*

$$\|m_{\mathrm{At}}\|_{r-3/2} \leq C_1 \exp\{C_2 E\}.$$

Proof. We can see from (6.16) that many of these terms are familiar. It is clear from Lemma 4.2 and from the definition of the energy that the $H^{r-3/2}$ norm of the first two terms on the right-hand side of (6.16) is bounded in terms of the energy. Similarly, it is clear that the final term in the brackets in (6.16), $(T - \mathbf{W} \cdot \hat{\mathbf{t}})m \cdot \hat{\mathbf{t}}$, is bounded in terms of the energy. We see from (6.11) that the $H^{r-3/2}$ norm of R_6 can be bounded in terms of the energy by straightforward applications of Lemma 3.5, Corollary 3.8, and Lemma 4.2.

To estimate R_5, it is clear from (6.5) that we need to be able to estimate z_t. To do this, we first remark that (2.26) and Lemma 4.2 imply that $\theta_t \in H^{r-5/2}$ when $\gamma \in H^{r-3/2}$ and $\theta \in H^{r-1}$. Next, we notice that (6.6) implies that $z_t \in H^{r-3/2}$. Using Corollary 3.8 and Lemma 3.5, we can now bound the $H^{r-3/2}$ norm of R_5 in terms of the energy. This completes the proof of the lemma. ∎

We now know from Lemmas 6.2 and 6.3 that all of the terms in the evolution equation for γ which appear only when the Atwood ratio is nonzero can be bounded in terms of the energy. Since the energy function is the same as in the density-matched case, this means that the proofs of the new existence, uniqueness, and continuous dependence theorems are only trivially different from the proofs in the previous case. In summary, we have the following.

THEOREM 6.4. *Given the initial condition $(\theta_0, \gamma_0, L_0) \in \mathcal{O}$, there exists a unique solution to the exact initial value problem for the vortex sheet with surface tension in the two-density case given by* (2.22), (6.1), (2.24), *and* (2.25) *which satisfies the non-self-intersection condition* (3.2) *and the boundedness condition* (3.5). *There exists a time $T > 0$ such that the solution, (θ, γ, L), is in $C^1([0,T]; \mathcal{O})$. T depends on r, $\mathcal{O}$, the Weber number, and $\|(\theta_0, \gamma_0, L_0)\|_{\mathcal{B}}$.*

Acknowledgments. This work was done as part of the author's doctoral dissertation. The author would like to thank his advisor, Tom Beale, for all his help and patience, and for reading many drafts of this work. The author would also like to thank the referees for their careful reading of this paper and for their many helpful comments. Finally, the author would like to thank Jonathan Goodman for discussions about revisions to the paper.

REFERENCES

[AF97] M. ABLOWITZ AND A. FOKAS, *Complex Variables: Introduction and Applications*, Cambridge University Press, Cambridge, UK, 1997.

[Amb02] D. AMBROSE, *Well-Posedness of Vortex Sheets with Surface Tension*, Ph.D. thesis, Duke University, Durham, NC, 2002.

[BMO82] G. BAKER, D. MEIRON, AND S. ORSZAG, *Generalized vortex methods for free-surface flow problems*, J. Fluid Mech., 123 (1982), pp. 477–501.

[Bea81] J. T. BEALE, *The initial value problem for the Navier-Stokes equations with a free surface*, Comm. Pure Appl. Math., 34 (1981), pp. 359–392.

[BHL93] J. T. BEALE, T. HOU, AND J. LOWENGRUB, *Growth rates for the linearized motion of fluid interfaces away from equilibrium*, Comm. Pure Appl. Math., 46 (1993), pp. 1269–1301.

[BHL96] J. T. BEALE, T. Y. HOU, AND J. LOWENGRUB, *Convergence of a boundary integral method for water waves*, SIAM J. Numer. Anal., 33 (1996), pp. 1797–1843.

[Bir62] G. Birkhoff, *Helmholtz and Taylor instability*, in Proceedings of Symposia in Applied Mathematics, Vol. 13, G. Birkhoff, R. Bellman, and C. C. Lin, eds., AMS, Providence, RI, 1962, pp. 55–76.

[CH96] H. Ceniceros and T. Hou, *Convergence of a non-stiff boundary integral method for interfacial flows with surface tension*, Math. Comp., 67 (1998), pp. 137–182.

[Fol95] G. Folland, *Introduction to Partial Differential Equations*, 2nd ed., Princeton University Press, Princeton, NJ, 1995.

[HLS94] T. Hou, J. Lowengrub, and M. Shelley, *Removing the stiffness from interfacial flows with surface tension*, J. Comput. Phys., 114 (1994), pp. 312–338.

[HLS97] T. Hou, J. Lowengrub, and M. Shelley, *The long-time motion of vortex sheets with surface tension*, Phys. Fluids, 9 (1997), pp. 1933–1954.

[Igu01] T. Iguchi, *Well-posedness of the initial value problem for capillary-gravity waves*, Funkcial. Ekvac., 44 (2001), pp. 219–241.

[ITT97] T. Iguchi, N. Tanaka, and A. Tani, *On the two-phase free boundary problem for two-dimensional water waves*, Math. Ann., 309 (1997), pp. 199–223.

[MB02] A. Majda and A. Bertozzi, *Vorticity and Incompressible Flow*, Cambridge University Press, Cambridge, UK, 2002.

[Saf95] P. G. Saffman, *Vortex Dynamics*, Cambridge University Press, Cambridge, UK, 1995.

[Sie95] M. Siegel, *A study of singularity formation in the Kelvin–Helmholtz instability with surface tension*, SIAM J. Appl. Math., 55 (1995), pp. 865–891.

[Tay96] M. Taylor, *Partial Differential Equations* III: *Nonlinear Equations*, Appl. Math. Sci. 117, Springer-Verlag, New York, 1996.

[Wu97] S. Wu, *Well-posedness in Sobolev spaces of the full water wave problem in* 2-*D*, Invent. Math., 130 (1997), pp. 39–72.

[Wu99] S. Wu, *Well-posedness in Sobolev spaces of the full water wave problem in* 3-*D*, J. Amer. Math. Soc., 12 (1999), pp. 445–495.

[Zei86] E. Zeidler, *Nonlinear Functional Analysis and Its Applications* I: *Fixed Point Theorems*, Springer-Verlag, New York, 1986.

SIAM J. MATH. ANAL.
Vol. 35, No. 1, pp. 245–267

ANALYSIS OF OIL TRAPPING IN POROUS MEDIA FLOW*

M. BERTSCH†, R. DAL PASSO†, AND C. J. VAN DUIJN‡

Abstract. We analyze a one-dimensional nonlinear convection-diffusion equation describing the flow of water and oil through a porous medium composed of two types of rock with different permeability. We prove existence, uniqueness, and regularity properties, as well as matching conditions between the two rock types.

Key words. degenerate parabolic equation, porous media flow, existence, uniqueness, qualitative properties, matching conditions

AMS subject classifications. Primary, 35K65; Secondary, 35B05, 35K10, 35K55, 76S05

DOI. 10.1137/S0036141002407375

1. Introduction and problem formulation.

It is well known that capillary forces, combined with spatial variations of rock properties, considerably reduce the recovery factor of an oil reservoir. For instance, it is difficult to remove oil from parts of the reservoir with small scale heterogeneities. Sometimes the oil may even remain trapped; see, for instance, [K, W]. This is clearly a difficult problem, mainly due to the complex nature of rock (soil) heterogeneities.

To understand oil trapping in heterogeneous media more quantitatively, [DMN] considered the case of a 2-phase water-oil flow which is perpendicular to an interface, separating two types of rock, across which the permeability changes abruptly. Under simplifying assumptions this leads to a one-dimensional flow problem which allowed them to investigate the role of convection and capillary diffusion in relation to the discontinuous permeability. They used formal asymptotics and numerical techniques. In this paper we will take their formulation as a starting point. The aim is to analyze the structure of the model equations resulting in existence, uniqueness, and regularity properties, as well as matching conditions between the two rock types.

Following [DMN] (further references are given there), the one-dimensional flow of water and oil through a porous medium is described by a nonlinear convection-diffusion equation for the reduced water saturation $S = S(x,t)$, with $0 \le S \le 1$. This equation has the form

$$\Phi \frac{\partial S}{\partial t} + \frac{\partial}{\partial x}\left\{ q f_w(S) + k(x) H(s) \frac{\partial p}{\partial x} \right\} = 0, \tag{1.1}$$

where Φ (porosity) and q (discharge) are positive constants, and where the functions $f_w,\ H : [0,1] \to [0,\infty)$ satisfy $f_w(0) = 0,\ f_w(S) > 0$ for $0 < S \le 1$ (typically convex-concave behavior) and $H(0) = H(1) = 0,\ H(S) > 0$ for $0 < S < 1$. Further $k(x)$ denotes permeability and p capillary pressure. Situating the discontinuity in

*Received by the editors May 9, 2002; accepted for publication December 13, 2002; published electronically July 8, 2003.
http://www.siam.org/journals/sima/35-1/40737.html
†Dipartimento di Matematica, Università di Roma, Tor Vergata, Via della Ricerca Scientifica, I-00133 Roma, Italy (bertsch@mat.uniroma2.it, dalpasso@mat.uniroma2.it).
‡Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands (c.j.v.duijn@tue.nl).

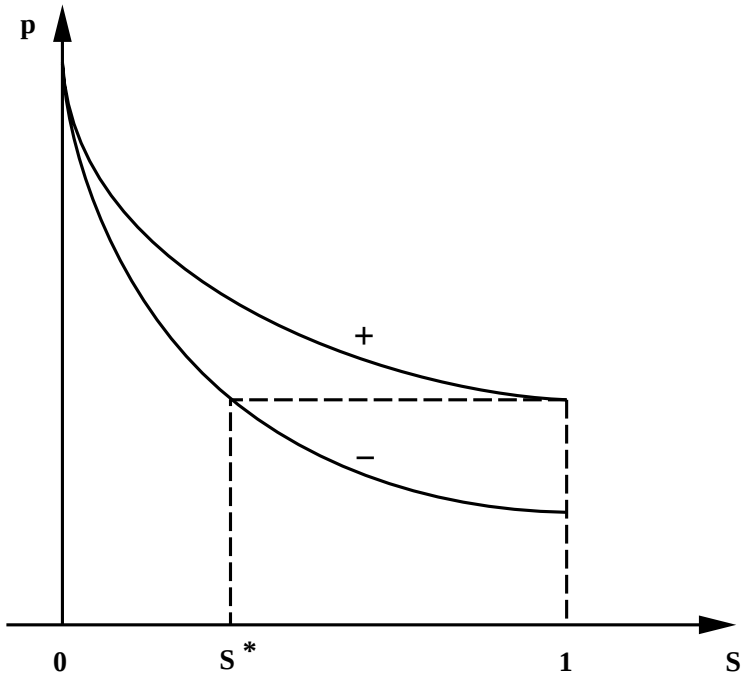


FIG. 1. *Capillary pressure curves for fine* (+) *and coarse* (−) *material. Here* $J(1) > 0$*, so an entry pressure exists.*

permeability at $x = 0$, we have

$$k(x) = \begin{cases} k^- & \text{for} \quad x < 0, \\ k^+ & \text{for} \quad x > 0. \end{cases} \tag{1.2}$$

Without loss of generality we take $0 < k^+ < k^- < \infty$. This means that coarse material occupies $\{x < 0\}$ and fine material $\{x > 0\}$. The flow is in positive x-direction.

For the capillary pressure the Leverett model [L] was used. With $\sigma > 0$ denoting interfacial tension, this means

$$p = p(x, S) = \sigma \frac{J(S)}{\sqrt{k(x)/\Phi}} \qquad \text{for} \quad 0 < S \leq 1, \tag{1.3}$$

where the Leverett function J is strictly decreasing in $(0, 1]$ with $J(1) \geq 0$. The quantity $\sqrt{k/\Phi}$ may be associated with the mean pore diameter, and the J-Leverett function is typical for the lithology of the porous medium. When $J(1) > 0$, the medium has an entry pressure given by $J(1)/\sqrt{k/\phi}$. This is the minimum pressure needed for the oil to enter a medium that is saturated by water. In this paper we assume $J(1) > 0$ and show that the occurrence of an entry pressure causes trapping of oil at the interface when the medium changes from coarse to fine. Figure 1 shows two typical capillary pressure functions, the top curve for fine material ($x > 0$), the bottom curve for coarse material ($x < 0$).

Because k is discontinuous, the capillary pressure may be discontinuous as well. This makes the interpretation of (1.1) across $x = 0$ difficult. To circumvent this problem, [DMN] considered (1.1) for $x < 0$ and $x > 0$, with matching conditions at

$x = 0$. One condition is obvious. Conservation of mass across $x = 0$ requires that the fluxes to the left and right of $x = 0$ be equal:

$$(\widetilde{\mathrm{M}}_1) \qquad \left(qf_w + k^- H \frac{\partial p}{\partial x}\right)_{x=0^-} = \left(qf_w + k^+ H \frac{\partial p}{\partial x}\right)_{x=0^+}$$

for all $t > 0$. A condition related to the pressure was obtained by a formal regularization procedure. Replacing in (1.1) $k(x)$ by C^∞ approximations $k_n(x)$, according to

$$k_n(x) = \begin{cases} k^- & \text{for} \quad x \le -\dfrac{1}{n}, \\ \varphi(nx) & \text{for} \quad -\dfrac{1}{n} < x < \dfrac{1}{n}, \\ k^+ & \text{for} \quad x \ge \dfrac{1}{n}, \end{cases} \tag{1.4}$$

with φ smooth ($\varphi(-1) = k^-$, $\varphi(1) = k^+$, and $\varphi' \le 0$), blowing up the transition region by $x \to nx$ and letting $n \to \infty$, we found the following. Let S^* be defined by the relation

$$\frac{J(S^*)}{\sqrt{k^-}} = \frac{J(1)}{\sqrt{k^+}} > 0, \tag{1.5}$$

and let S^- and S^+ denote, respectively, the left and right limits of S at $x = 0$. Then for all $t > 0$ (see also Figure 1),

$$(\widetilde{\mathrm{M}}_2) \qquad \begin{cases} \dfrac{J(S^-)}{\sqrt{k^-}} = \dfrac{J(S^+)}{\sqrt{k^+}} & \text{if} \quad S^- \le S^* \quad \text{(pressure continuous)}, \\ S^+ = 1 & \text{if} \quad S^- > S^* \quad \text{(positive pressure jump)}. \end{cases}$$

Instead of analyzing (1.1) and conditions $(\widetilde{\mathrm{M}}_{1-2})$ in the form presented above, we shall consider a further simplified model problem, without losing essential characteristic features. We take in (1.1)

$$f(S) = S, \quad H(S) = 1 - S, \quad \text{and} \quad J(S) = 2 - S.$$

After a trivial scaling, the following equations result for the oil saturation $u = 1 - S$:

$$u_t + f_x = 0 \qquad (u \ge 0), \tag{1.6}$$

$$f = u - N_c\, k\, u\, p_x, \tag{1.7}$$

$$p = \frac{1 + u}{\sqrt{k(x)}}, \tag{1.8}$$

where f denotes the flux and N_c the dimensionless capillary number

$$N_c = \frac{\sigma\sqrt{K\phi}}{q\mu_w L}.$$

Here K is a characteristic k-value, L a characteristic length scale, and μ_w the water viscosity. By an additional scaling we may set $N_c = 1$. Further, k is given by (1.2) and the subscripts t and x denote partial differentiation.

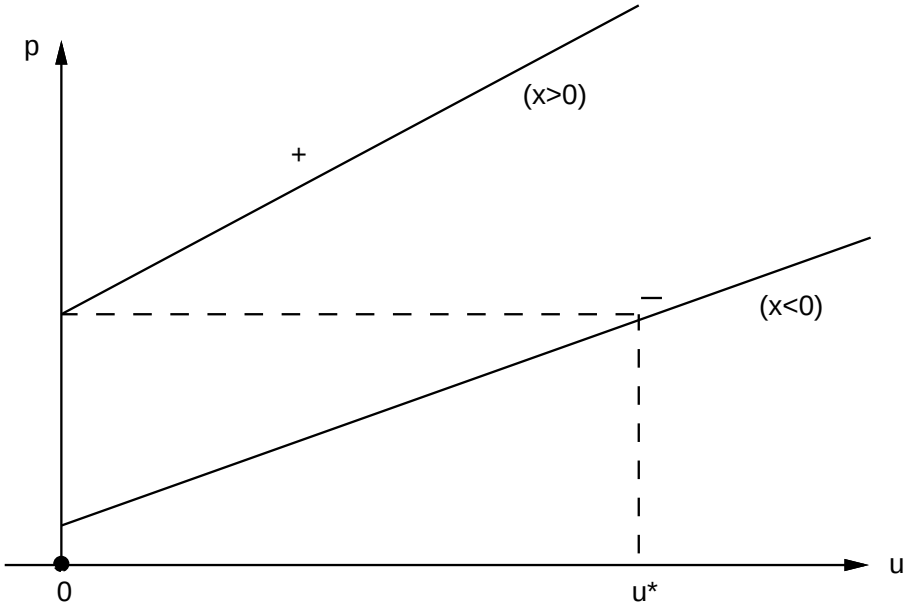


FIG. 2. *Transformed capillary pressures.*

We solve (1.6)–(1.8) in the subdomains

$$Q^{\pm} = \{(x,t) : x \in \mathbb{R}^{\pm},\ t \in (0,\infty)\},$$

with transformed matching conditions at $x = 0$. These are

$$[f] = 0 \qquad \text{in} \quad (0,\infty), \tag{M$_1$}$$

and (see Figure 2)

$$\begin{cases} \dfrac{1+u^-}{\sqrt{k^-}} = \dfrac{1+u^+}{\sqrt{k^+}} & \text{if} \quad u^- \geq u^* \\ u^+ = 0 & \text{if} \quad u^- < u^* \end{cases} \qquad \text{in} \quad (0,\infty),$$

or, equivalently,

$$u^+[p] = 0\ , \quad [p] \geq 0 \quad \text{in} \quad (0,\infty). \tag{M$_2$}$$

Here $u^* = \sqrt{\frac{k^-}{k^+}} - 1$. As before, $u^{\pm} = u^{\pm}(t) = u(0\pm,t)$, $[u] = u^+ - u^-$, and f and p have similar notation.

At $t = 0$ we prescribe

$$u(\cdot,0) = u_0(\cdot) \quad \text{in} \quad \mathbb{R}, \tag{1.9}$$

with u_0 satisfying

$$\text{(H)} \qquad \begin{cases} u_0 : \mathbb{R} \to [0,\infty), \quad \operatorname{supp}(u_0) \subset \mathbb{R} \text{ is bounded;} \\ u_0 \text{ uniformly Lipschitz continuous in } \mathbb{R}\backslash\{0\}; \\ u_0^+[p_0] = 0, \quad f_0 := u_0 - \dfrac{\sqrt{k}}{2}(u_0^2)' \in BV(\mathbb{R}\backslash\{0\}). \end{cases}$$

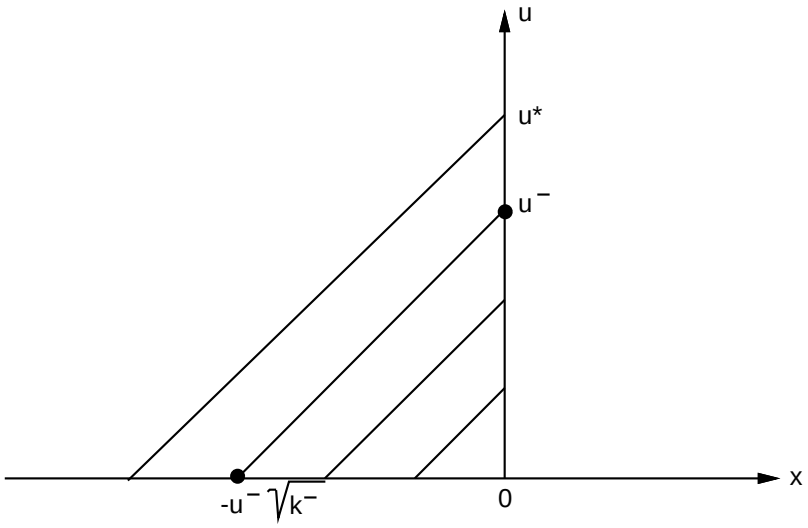


FIG. 3. *Admissible steady state solutions* ($u^- \le u^*$).

The pressure condition at $t = 0$ is needed to construct an approximate sequence $\{u_{0n}\}$ for which the corresponding fluxes $f_{0n} := u_{0n} - k_n u_{0n}(p_{0n})'$ are uniformly bounded in $BV(\mathbb{R})$. This in turn will imply $f \in L^\infty((0,\infty);\ BV(\mathbb{R}))$, which is a crucial point in the existence proof. If the k_n are taken as in (1.4), then $[p_0] \ge 0$ is needed as well. We will return to this in section 2 and in the appendix.

For steady state solutions, the role of (M_2) can be seen explicitly. Assume $u = u(x)$ only, with $u(-\infty) = u(+\infty) = 0$. Then

$$f = u - k\,up' = 0 \qquad \text{in} \quad \mathbb{R}\backslash\{0\}. \tag{1.10}$$

Using $u \ge 0$, we obtain

$$u(x) = 0 \qquad \text{for} \quad x > 0.$$

Hence the first condition in (M_2) is always satisfied. Given any $u^- \ge 0$, we see that

$$u(x) = \left(u^- + \frac{1}{\sqrt{k^-}}\,x\right)_+ \tag{1.11}$$

satisfies (1.10) for $x < 0$. Here $(\ \cdot\ )_+ := \max\{\cdot, 0\}$. However, only for $u^- \in [0, u^*]$ we have $[p] \ge 0$. Thus we have a family of admissible steady state solutions, as shown in Figure 3.

Integrating the maximal steady state gives the maximal amount of oil that can be trapped to the left of the permeability discontinuity. It is given by

$$\bar{M} = \frac{1}{2}(u^*)^2\sqrt{k^-}. \tag{1.12}$$

Next we give the weak formulation of the trapping problem. Because the flux is expected to be continuous across $x = 0$, it will be defined globally in the formulation.

The saturation (and pressure) will be considered in the subdomains Q^- and Q^+ separately. Let

$$Q^0 := Q^- \cup Q^+ \quad \text{and} \quad Q := \mathbb{R} \times (0, \infty).$$

Combining the saturation equations and the matching conditions gives the following.

Problem P. Find $u : Q^0 \to [0, \infty)$, $f : Q \to \mathbb{R}$ such that

(i) u, $(u^2)_x \in L^\infty(Q^0)$; u is uniformly continuous in Q^0;

(ii) $f \in L^\infty((0, \infty); BV(\mathbb{R}))$;

(iii) $f = u - \frac{\sqrt{k}}{2}(u^2)_x$ a.e. in Q^0 and $\int_Q (u\zeta_t + f\zeta_x)dxdt + \int_{\mathbb{R}} u_0(x)\zeta(x, 0)dx = 0$ for all $\zeta \in H^1(Q) \cap C(\bar{Q})$, with compact support in $\bar{Q}$;

(iv) $u^+[p] = 0$ and $[p] \geq 0$ in $(0, \infty)$, where $p := \frac{1+u}{\sqrt{k}}$ in Q^0.

To prove existence we apply a k-regularization as in (1.4). This yields a sequence of approximating problems on Q for which we derive the necessary estimates. This is done in section 2. In section 3 we consider the limit $n \to \infty$ giving existence for Problem P, with u satisfying a porous media equation ($m = 2$) with linear convection in Q^0. Clearly (M_2) is satisfied. The weak equation in (iii) implies $[f] = 0$ a.e. in $(0, \infty)$. The comparison principle, with uniqueness as a consequence, is shown in section 4. In section 5 we give sufficient conditions for oil trapping; i.e., conditions that imply $u(x, t) = 0$ for $x > 0$ and for all $t > 0$. Finally, in section 6, we present some closing remarks about nonuniqueness, waiting times, and optimal regularity.

In a recent paper [DMP] considered oil transport in a multilayered porous medium. This work involves a discontinuous permeability which varies periodically in space. Using homogenization techniques they derived effective (upscaled) transport equations for the case where the periodicity length is small compared to the characteristic length L. In their analysis matching conditions ($\widetilde{M}_1$) and ($\widetilde{M}_2$) play a crucial role. They lead to a macroscopic irreducible oil saturation.

2. The approximate problem. In this section we study the approximate equation in which k is replaced by the smooth function k_n, defined by (1.4). Together with k we also need to approximate the initial value u_0. We construct approximations u_{0n}, so that the corresponding fluxes

$$f_{0n} := u_{0n} - k_n u_{0n} p'_{0n}, \qquad p_{0n} := \frac{1 + u_{0n}}{\sqrt{k_n}} \tag{2.1}$$

have a uniformly bounded total variation. In addition we require that each u_{0n} is strictly positive to eliminate the degeneracy of the equation at points where u vanishes. The existence of such u_{0n} is given in the following lemma. Since the proof is quite technical, it is given in the appendix.

LEMMA 2.1. *Let $n \in \mathbb{N}$ and let k_n be defined by* (1.4). *Suppose u_0 satisfies hypothesis* (H) *and in addition*

$$[p_0] = \frac{1 + u_0^+}{\sqrt{k^+}} - \frac{1 + u_0^-}{\sqrt{k^-}} \geq 0. \tag{2.2}$$

Then there exist $u_{0n} \in W^{1,\infty}(\mathbb{R})$ and $\varepsilon_n \in \mathbb{R}^+$ such that

(i) *$u_{0n} \geq \varepsilon_n > 0$ in $\mathbb{R}$, and $u_{0n}(x) = \varepsilon_n$ for $|x|$ sufficiently large;*

(ii) *u_{0n} is uniformly bounded in $\mathbb{R}$, and f_{0n}, defined by* (2.1), *is uniformly bounded in $BV(\mathbb{R})$;*

(iii) *As* $n \to \infty$,

$$u_{0n} \to u_0 \quad \textit{uniformly in } \mathbb{R}\backslash\{0\}$$

and

$$u_{0n} - \varepsilon_n \to u_0 \quad \textit{in } L^1(\mathbb{R}).$$

For each $n \in \mathbb{N}$ we consider the approximate problem

$$(\mathrm{P}_n) \qquad \begin{cases} u_t + u_x = (k_n u\, p_x)_x, \quad p = \dfrac{1+u}{\sqrt{k_n}} \quad \text{in } Q, \\ u(x,0) = u_{0n}(x) \quad \text{for } x \in \mathbb{R}. \end{cases}$$

In the remainder of this section we prove the following results.

THEOREM 2.2. *Let* u_{0n} *be given by Lemma* 2.1. *Then problem* (P_n) *has a solution* $u_n \in C^\infty(Q) \cap C(\bar{Q})$ *such that*

(i) $0 < u_n \le \mathcal{C}$ *in* Q, *where* $\mathcal{C}$ *does not depend on* n;

(ii) $f_n := u_n - k_n u_n (\frac{1+u_n}{\sqrt{k_n}})_x$ *is uniformly bounded in* $L^\infty([0,\infty); BV(\mathbb{R}))$;

(iii) u_n *is uniformly continuous in* $\{\mathbb{R}\backslash(-\varepsilon,\varepsilon)\} \times [0,\infty)$ *for all* $\varepsilon > 0$.

Proof. Since $u_{0n} \ge \varepsilon_n > 0$ in $\mathbb{R}$, problem (P_n) is nondegenerate at $t = 0$. Hence it has a unique local (with respect to t) classical solution u_n; see, for instance, [LSU] and [F]. This solution can be continued as long as it remains bounded and bounded away from zero. Let $Q_{T_n} := \mathbb{R} \times (0, T_n)$ denote the maximal existence domain for u_n.

A positive lower bound follows from the maximum principle. Indeed, if we set $L_n := \max_{\mathbb{R}} |(\sqrt{k_n})''|$ we observe that the solution of the initial value problem

$$(\mathrm{LB}) \qquad \begin{cases} s' = -L_n s(1+s) \quad \text{for} \quad t > 0, \\ s(0) = \varepsilon_n \end{cases}$$

is a subsolution for problem (P_n). Hence if s_n denotes the solution of (LB), we have

$$u_n(x,t) \ge s_n(t) > 0 \quad \text{for} \quad (x,t) \in Q_{T_n}. \tag{2.3}$$

Before proving a uniform upper bound for u_n, we observe that the flux f_n is uniformly bounded in Q_{T_n}. A straightforward calculation yields for f_n the linear equation

$$f_t = a_n f_{xx} + b_n f_x, \tag{2.4}$$

where

$$a_n(x,t) := u_n \sqrt{k_n}, \qquad b_n(x,t) := -\frac{f_n}{u_n} - \frac{u_n k_n'}{2\sqrt{k_n}}. \tag{2.5}$$

Hence, by the maximum principle

$$||f_n||_{L^\infty(Q_{T_n})} \le ||f_{0n}||_{L^\infty(\mathbb{R})} \le \mathcal{C} \tag{2.6}$$

for all $n \in \mathbb{N}$.

We use this estimate to demonstrate a uniform upper bound for u_n in Q_{T_n}. As a first observation we note that (2.6) implies the differential inequality

$$|u_n - \sqrt{k^-}\, u_n u_{nx}| \le \mathcal{C} \quad \text{in} \quad \left(-\infty, -\frac{1}{n}\right] \times [0, T_n). \tag{2.7}$$

Then the upper bound for u_n in this set is immediate if we can control the decay of u_n as $x \to -\infty$. This decay results from the following argument.

Let $\bar{u}_n$ be a steady state solution satisfying

$$\begin{cases} u - k_n u\, p' = \varepsilon_n, \qquad p = \dfrac{1+u}{\sqrt{k_n}} \quad \text{in } \mathbb{R}, \\ u(\pm\infty) = \varepsilon_n. \end{cases}$$

Clearly, $\bar{u}_n(x) = \varepsilon_n$ for all $x \geq \frac{1}{n}$. The corresponding pressure $\bar{p}_n$ satisfies

$$\begin{cases} k_n(p\sqrt{k_n} - 1)p' = p\sqrt{k_n} - 1 - \varepsilon_n \quad \text{for} \quad x < \dfrac{1}{n}, \\ p\left(\dfrac{1}{n}\right) = \dfrac{1+\varepsilon_n}{\sqrt{k^+}}. \end{cases}$$

At points where $\bar{p}_n' = 0$, we must have $\bar{p}_n > 0$ and $\bar{p}_n'' < 0$. We use this to obtain $\bar{p}_n' > 0$ and $\bar{p}_n > \frac{1+\varepsilon_n}{\sqrt{k^-}}$ on $(-\infty, \frac{1}{n})$, and $\bar{p}_n(x) \to \frac{1+\varepsilon_n}{\sqrt{k^-}}$ as $x \to -\infty$. In particular, $\bar{u}_n(x) \to \varepsilon_n$ exponentially as $x \to -\infty$ and $\bar{u}_n - \varepsilon_n \in L^1(\mathbb{R})$, uniformly in $n \in \mathbb{N}$.

Now using Lemma 2.1(iii) and an argument as in the proof of Theorem 4.1, one finds for $t > 0$ the L^1-contraction

$$\int_{\mathbb{R}} |u_n(x,t) - \bar{u}_n(x)|dx \leq \int_{\mathbb{R}} |u_{0n}(x) - \bar{u}_n(x)|dx.$$

This inequality controls the behavior of u_n as $|x| \to \infty$. Combined with (2.7) it gives the upper bound in $(-\infty, -\frac{1}{n}] \times (0, T_n)$. Arguing similarly for $x > \frac{1}{n}$, we conclude that for all $n \in \mathbb{Z}^+$

$$u_n(x,t) \leq \mathcal{C} \quad \text{for} \quad |x| \geq \frac{1}{n}, \quad 0 \leq t < T_n. \tag{2.8}$$

To obtain the upper bound in the remaining strip $[-\frac{1}{n}, \frac{1}{n}] \times [0, T_n)$ we express (2.6) in terms of the pressure p_n:

$$|p_n\sqrt{k_n} - 1 - k_n(p_n\sqrt{k_n} - 1)p_{nx}| \leq \mathcal{C}. \tag{2.9}$$

By (2.8), $p_n(\pm\frac{1}{n}, t)$ is uniformly bounded. Then (2.9) implies that p_n, and thus u_n, is uniformly bounded as well.

The uniform upper bound, together with lower bound (2.3), guarantees existence for all $t > 0$. Hence, $T_n = \infty$ for each $n \in \mathbb{N}$. This completes the proof of (i).

The proof of (ii) is a direct consequence of Lemma 2.1(ii) and the total variation estimate for the flux in Lemma 2.4 below.

We conclude by proving (iii). The boundedness of u_n and the flux estimate (2.7) imply that u_n is uniformly Hölder continuous (exponent $\frac{1}{2}$) with respect to x in $\{(x,t) : x < -\frac{1}{n},\ t > 0\}$. The same result holds in $\{(x,t) : x > \frac{1}{n},\ t > 0\}$. The smoothness and boundedness of the coefficients in the u_n-equation allow us to apply [G1], yielding that u_n is uniformly Hölder continuous (exponent $\frac{1}{4}$) with respect to t in $\{(x,t) : |x| > \frac{1}{n},\ t > 0\}$. Since, for fixed $\varepsilon > 0$, $\frac{1}{n} < \varepsilon$ for n large enough, this proves (iii) and completes the proof of Theorem 2.2. □

Remark 2.3. It is not difficult to show that the steady states $\bar{u}_n$, corresponding to $k = k_n$ and $\bar{u}_n(\pm\infty) = \varepsilon_n$, approximate the maximal steady state in Figure 3. In essence this follows from $\bar{u}_n(x) = \varepsilon_n$ for all $x \geq \frac{1}{n}$ and, using the pressure equation,

$$0 < \bar{p}_n\left(\frac{1}{n}\right) - \bar{p}_n\left(-\frac{1}{n}\right) = \int_{-\frac{1}{n}}^{+\frac{1}{n}} \frac{1}{k_n} \frac{\bar{p}_n(x)\sqrt{k_n} - 1 - \varepsilon_n}{\bar{p}_n(x)\sqrt{k_n} - 1}\, dx \to 0$$

as $n \to \infty$.

It remains to prove the following lemma used in the proof of Theorem 2.2.

LEMMA 2.4. *Let u_{0n} be given by Lemma 2.1 and let u_n be the corresponding solution of problem* (P_n). *Then*

$$TV_{\mathbb{R}}(f_n(t)) \le TV_{\mathbb{R}}(f_{0n}) \qquad \textit{for all} \quad t > 0.$$

Proof. Each flux f_n satisfies the linear problem

$$\begin{cases} f_t = a_n f_{xx} + b_n f_x & \text{in } Q, \\ f(x,0) = f_{0n}(x) & \text{for} \quad x \in \mathbb{R}, \end{cases}$$

where a_n and b_n, defined in (2.5), are bounded functions and where f_{0n} has uniformly bounded variation. First we proceed formally. Let us fix $\varepsilon > 0$ and calculate (dropping the subscript n)

$$\begin{aligned} \frac{d}{dt}\int_{\mathbb{R}} \left\{ \sqrt{f_x^2 + \varepsilon} - \sqrt{\varepsilon} \right\} &= \int_{\mathbb{R}} \frac{f_x}{\sqrt{f_x^2+\varepsilon}} (a f_{xx} + b f_x)_x \\ &= -\varepsilon \int_{\mathbb{R}} \frac{f_{xx}(a f_{xx} + b f_x)}{(f_x^2+\varepsilon)^{3/2}}. \end{aligned}$$

Integrating in time gives, for any $t > 0$,

$$\begin{aligned} \int_{\mathbb{R}} \left\{ \sqrt{f_x^2(t) + \varepsilon} - \sqrt{\varepsilon} \right\} - \int_{\mathbb{R}} \left\{ \sqrt{{f'_{0n}}^2 + \varepsilon} - \sqrt{\varepsilon} \right\} &= -\varepsilon \int_{\mathbb{R}\times(0,t)} \frac{a f_{xx}^2 + b f_x f_{xx}}{(f_x^2+\varepsilon)^{3/2}} \\ &\le -\int_{\mathbb{R}\times(0,t)} \frac{\varepsilon f_x}{(f_x^2+\varepsilon)^{3/2}}\, b f_{xx}. \end{aligned}$$

Since

$$\left| \frac{\varepsilon f_x}{(f_x^2+\varepsilon)^{3/2}} \right| \le 1$$

and

$$\frac{\varepsilon f_x}{(f_x^2+\varepsilon)^{3/2}} \to 0, \quad \text{pointwise in } Q \text{ as } \varepsilon \to 0,$$

the boundedness of b and Lebesgue's dominated convergence theorem imply

$$\int_{\mathbb{R}} |f_x(t)| \le \int_{\mathbb{R}} |f'_{0n}|,$$

provided $f_{xx} \in L^1(\mathbb{R}\times(0,t))$. To complete the proof of the lemma we need to make this argument rigorous.

It is enough to apply a mollifier to the initial function f_{0n} of the linear flux problem. This ensures the smoothness up to $t = 0$ necessary to carry out the above calculations. ∎

3. Existence for Problem P. Let u_n be the solution of problem (P_n) as stated in Theorem 2.2. By a standard argument there exist a subsequence of $\{u_n\}$, denoted again by $\{u_n\}$, and $u \in L^\infty(Q) \cap C((\mathbb{R}^- \cup \mathbb{R}^+) \times [0,\infty))$ such that

$$u_n \to u \quad \text{in } C_{\mathrm{loc}}((\mathbb{R}^- \cup \mathbb{R}^+) \times [0,\infty))$$

as $n \to \infty$. We show the following theorem.

THEOREM 3.1. *u is a solution of Problem* P.

Proof. Clearly u is a (weak) solution of the equation

$$u_t + u_x = \frac{1}{2}\sqrt{k^\pm}(u^2)_{xx} \quad \text{in } Q^\pm$$

and

$$f = u - \frac{1}{2}\sqrt{k^\pm}\,(u^2)_x \in L^\infty([0,\infty);\ BV(\mathbb{R}^\pm)).$$

The boundedness of u and f implies that u^2 is uniformly Lipschitz continuous with respect to x in Q^0. Hence the following quantities are well defined for each $t > 0$:

$$u^\pm(t), \qquad f^\pm(t), \qquad \text{and} \qquad p^\pm(t) = \frac{1 + u^\pm(t)}{\sqrt{k^\pm}}.$$

Using the equation

$$u_t + f_x = 0 \qquad \text{a.e. in } Q^\pm$$

and again the boundedness of f, we obtain as in [DP] that the functions

$$t \to u^\pm(t)$$

are continuous in $[0,\infty)$.

Next we claim

$$f^+(t) = f^-(t) \quad \text{for almost all } t > 0. \tag{3.1}$$

Indeed, using the asymptotic behavior of $u_n(x,t)$ as $|x| \to \infty$, we find, for $n \to \infty$,

$$\int_{\mathbb{R}} (u_n(x,t) - \varepsilon_n)dx = \int_{\mathbb{R}} (u_{0n}(x) - \varepsilon_n)dx \to \int_{\mathbb{R}} u_0(x)dx$$

and hence

$$\int_{\mathbb{R}} u(x,t)dx = \int_{\mathbb{R}} u_0(x)dx \quad \text{for all} \quad t > 0,$$

which expresses conservation of mass. This identity implies

$$\begin{aligned} 0 &= \lim_{\delta \to 0^+} \left(\int_{-\infty}^{-\delta} u(x,t)dx + \int_{\delta}^{\infty} u(x,t)dx - \int_{-\infty}^{-\delta} u_0(x)dx - \int_{\delta}^{\infty} u_0(x)dx \right) \\ &= \int_0^t (f^+(s) - f^-(s))ds \quad \text{for all} \quad t > 0. \end{aligned}$$

Together with the equations in $Q^\pm$, equality (3.1) implies the weak form (iii) of Problem P.

It remains to prove

$$u^+[p] = 0 \quad \text{and} \quad [p] \geq 0 \quad \text{for all} \quad t > 0. \tag{3.2}$$

For this purpose we study u_n and p_n in the interval $(-\frac{1}{n}, \frac{1}{n})$. Since k_n changes rapidly there, we make the blow-up

$$y = nx \qquad \text{for} \quad -\frac{1}{n} < x < \frac{1}{n}.$$

Knowing that the fluxes f_n are uniformly bounded, we obtain

$$|u_n - nk_n u_n (p_n)_y| \leq \mathcal{C}.$$

Thus for appropriate $\mathcal{C} > 0$ we have

$$|u_n (p_n)_y| \leq \frac{\mathcal{C}}{n},$$

or

$$\left| (u_n^2)_y - \frac{\varphi'}{\varphi} u_n (1 + u_n) \right| \leq \frac{\mathcal{C}}{n} \tag{3.3}$$

for all $-1 < y < 1$ and $t > 0$.

Hence u_n^2 are Lipschitz continuous in $[-1, 1]$ uniformly with respect to n and t. Up to a subsequence, $u_n \to u$ uniformly in $[-1, 1]$, as $n \to \infty$, for all $t > 0$; in particular $u(-1, t) = u^-(t)$ and $u(1, t) = u^+(t)$. In addition, it follows easily from (3.3) that u satisfies

$$(u^2)_y = \frac{\varphi'}{\varphi} u(1 + u) \tag{3.4}$$

in $\{(y, t) : -1 < y < 1,\ t > 0\}$. Since φ' is nonpositive, u is decreasing. Thus, if $u^+(t) > 0$, we have $u(y, t) > 0$ in $[-1, 1]$ and (3.4) reduces to

$$u_y = \frac{\varphi'}{2\varphi}(1 + u). \tag{3.5}$$

A straightforward calculation gives $[p] = 0$.

Next suppose $u^+(t) = 0$. We have to show $[p] \geq 0$. If $u^-(t) = 0$, we get

$$[p] = \frac{1}{\sqrt{k^+}} - \frac{1}{\sqrt{k^-}} > 0.$$

If $u^-(t) > 0$, define $\tilde{y} := \sup\{y \in [-1, 1] : u(y, t) > 0\}$ and solve (3.5) in $[-1, \tilde{y}]$. This gives

$$\frac{1 + u^-}{\sqrt{k^-}} = \frac{1}{\sqrt{\varphi(\tilde{y})}} \leq \frac{1}{\sqrt{k^+}},$$

which implies $[p] \geq 0$ and $u^-(t) \leq u^*$. □

4. The comparison principle. We start with some preliminary observations for solutions (u, f) of Problem P. Choosing test functions with support in $Q^{\pm}$, we obtain

$$\int_{Q^{\pm}} u\zeta_t + \int_{Q^{\pm}} \left(u - \frac{\sqrt{k^{\pm}}}{2}(u^2)_x \right)\zeta_x = 0.$$

Thus away from $x = 0$ we have two weak equations of "porous media" type $(m = 2)$ with linear convection, implying

$$u_t + \left(u - \frac{\sqrt{k^{\pm}}}{2}(u^2)_x \right)_x = 0 \quad \text{a.e. in } Q^{\pm} \tag{4.1}$$

and

$$\operatorname{supp}(u(t)) \text{ is bounded in } \mathbb{R} \tag{4.2}$$

for all $t \in [0, \infty)$. Further, using hypothesis (H), we can apply the Bernstein argument of [A] in the truncated domain

$$Q^{\delta} := \mathbb{R}\backslash(-\delta, \delta) \times (0, \infty) \qquad (\text{for } \delta > 0, \text{ fixed})$$

to obtain

$$||u_x||_{L^{\infty}(Q^{\delta})} \leq C(\delta). \tag{4.3}$$

We use this to derive an estimate on u_t in Q^{δ}. Let u be a smooth solution of (4.1) in the sense of the usual "porous media" approximations, and let $\xi : \mathbb{R} \to [0, 1]$ be an even C^1 cut-off function satisfying

$$\xi(x) = \begin{cases} 0 & \text{for} \quad 0 \leq x \leq \delta/2, \\ 1 & \text{for} \quad \delta \leq x \leq L, \\ 0 & \text{for} \quad x \geq L + 1 \end{cases}$$

for any $L > \delta$. Multiplying (4.1) by $\xi^2 u_t$ gives

$$\int_{Q_\tau} \xi^2 u_t^2 = -\int_{Q_\tau} \xi^2 u_t u_x - \int_{Q_\tau} \xi\xi' \sqrt{k} u_t (u^2)_x - \int_{Q_\tau} \xi^2 \frac{\sqrt{k}}{2}\, u_{xt}\, (u^2)_x,$$

where $Q_\tau = \mathbb{R} \times (0, \tau)$ with $\tau > 0$ arbitrarily chosen. Using $u_{xt}(u^2)_x = u(u_x^2)_t$, the last integral becomes

$$\int_{\mathbb{R}} \xi^2 \frac{\sqrt{k}}{2}\, u u_x^2 \Big|_0^{\tau} - \int_{Q_\tau} \xi^2 \frac{\sqrt{k}}{2}\, u_t u_x^2.$$

Then (i) of Problem P and (4.3) in $Q^{\delta/2}$ give

$$\int_{Q_\tau} \xi^2 u_t^2 \leq C(\delta, \tau),$$

implying

$$u_t \in L^2_{\text{loc}}(\bar{Q^{\delta}}). \tag{4.4}$$

We are now in a position to prove the following theorem.

THEOREM 4.1. *Let (u_1, f_1) and (u_2, f_2) be weak solutions of Problem* P *corresponding to initial values u_{01} and u_{02}, respectively. Then $u_{01} \le u_{02}$ in $\mathbb{R}$ implies $u_1 \le u_2$ in Q^0.*

Proof. Let $\tau > 0$ be arbitrary. In the weak equation for the difference

$$\int_Q \{(u_1 - u_2)\zeta_t + (f_1 - f_2)\zeta_x\} + \int_{\mathbb{R}} \zeta(u_{01} - u_{02}) = 0,$$

we take the test function

$$\zeta = \xi\psi S_\varepsilon(u_1^2 - u_2^2),$$

where the following hold:

(i) ξ is an even C^1 cut-off function near $x = 0$,

$$\xi(x) = \begin{cases} 0 & \text{for} \quad 0 \le x \le \delta/2, \\ 1 & \text{for} \quad x \ge \delta, \end{cases} \qquad \xi'(x) \ge 0 \text{ for } \delta/2 < x < \delta.$$

(ii) ψ is a C^1 cut-off function near $t = \tau$,

$$\psi(t) = \begin{cases} 1 & \text{for} \quad 0 \le t \le \tau - \mu, \\ 0 & \text{for} \quad \tau \le t, \end{cases} \qquad \psi'(t) \le 0 \text{ for } \tau - \mu < t < \tau.$$

(iii) $S_\varepsilon : \mathbb{R} \to [0, 1]$ is given by

$$S_\varepsilon(r) = \begin{cases} 0, & r \le 0, \\ \dfrac{r}{\sqrt{r^2 + \varepsilon^2}}, & r > 0. \end{cases}$$

Here δ, μ, and ε are small positive parameters. Note that for $\varepsilon \searrow 0$

$$\left.\begin{array}{l} S_\varepsilon(r) \to \chi_{\{r>0\}} := \left\{ \begin{array}{ll} 1, & r > 0, \\ 0, & r \le 0, \end{array} \right. \\ r S_\varepsilon'(r) \to 0 \end{array}\right\} \quad \text{pointwise in } \mathbb{R}. \tag{4.5}$$

Integrating the first term by parts gives

$$\begin{aligned} &\int_{Q_\tau} (u_1 - u_2)_t\, \xi\psi S_\varepsilon(u_1^2 - u_2^2) \\ &\quad = \int_{Q_\tau} (f_1 - f_2)\psi \Big\{ \xi' S_\varepsilon(u_1^2 - u_2^2) + \xi S_\varepsilon'(u_1^2 - u_2^2)(u_1^2 - u_2^2)_x \Big\} \\ &\quad \le \int_{Q_\tau} (f_1 - f_2)\, \psi\xi' S_\varepsilon(u_1^2 - u_2^2) + \int_{Q_\tau} (u_1 - u_2)\, \psi\xi S_\varepsilon'(u_1^2 - u_2^2)(u_1^2 - u_2^2)_x. \end{aligned}$$

For fixed $\mu, \delta > 0$, we first let $\varepsilon \searrow 0$. Using (4.5), we have

$$(u_1 - u_2)\psi\xi S_\varepsilon'(u_1^2 - u_2^2) \to 0 \quad \text{pointwise in } Q_\tau.$$

Hence by (4.4) we obtain

$$\int_{Q_\tau} \xi\psi((u_1 - u_2)_+)_t \le \int_{Q_\tau} (f_1 - f_2)\psi\xi' \chi_{\{u_1 > u_2\}}.$$

Next we let $\mu \searrow 0$. This gives

$$\begin{aligned}\int_{\mathbb{R}} & \xi(u_1-u_2)_+(\tau)\\ &\le \int_0^\tau \left\{ \int_{-\delta}^{-\delta/2} (f_1-f_2)\xi' \chi_{\{u_1>u_2\}} + \int_{\delta/2}^{\delta} (f_1-f_2)\xi' \chi_{\{u_1>u_2\}} \right\} \\ &=: \int_0^\tau \{I_\delta^- + I_\delta^+\}.\end{aligned} \tag{4.6}$$

Let $t \in (0,\tau)$ be chosen such that f^-, f^+ exist. Consider the possibilities:

(i) $u_1^+ \neq u_2^+$, say $u_1^+ > u_2^+$. Then $u_1 > u_2$ in a right neighborhood of $x=0$ and $\chi_{\{u_1>u_2\}} = 1$ in $(\delta/2,\delta)$ for δ sufficiently small. The pressure conditions (M_2) give $u_1^- > u_2^-$: if $u_2^+ > 0$, then $[p_1]=[p_2]=0$ implies $u_1^- > u_2^-$; if $u_2^+ = 0$, then $u_2^- \le u^*$, while $u_1^- > u^*$. Therefore also $\chi_{\{u_1>u_2\}} = 1$ in $(-\delta,-\delta/2)$. As a consequence

$$\lim_{\delta\searrow 0}(I_\delta^- + I_\delta^+) = [f_1]-[f_2] = 0.$$

(ii) $u_1^+ = u_2^+$. Now we need to compare the corresponding fluxes. Suppose that $f_1^+ = f_2^+$. Then

$$\sup_{(\delta/2,\delta)} (f_1-f_2)\chi_{\{u_1>u_2\}} \to 0 \quad \text{as} \quad \delta \searrow 0,$$

and the same applies in $(-\delta,-\delta/2)$. Thus again

$$\lim_{\delta\searrow 0}(I_\delta^- + I_\delta^+) = 0.$$

If $f_1^+ > f_2^+$, then $(u_1^2)_x < (u_2^2)_x$ and therefore $u_1 < u_2$ in $(\delta/2,\delta)$. Thus

$$I_\delta^- + I_\delta^+ = I_\delta^- \le 0 \qquad \text{for } \delta > 0 \text{ sufficiently small.}$$

Finally, if $f_1^+ < f_2^+$, then $(u_1^2)_x > (u_2^2)_x$ and $u_1 > u_2$ in $(\delta/2,\delta)$. Thus $\lim_{\delta\searrow 0} I_\delta^+ = f_1^+ - f_2^+$. Furthermore, since

$$(f_1-f_2)\xi'\chi_{\{u_1>u_2\}} \le (f_1-f_2)\xi' \quad \text{in } (-\delta,-\delta/2),$$

$$\limsup_{\delta\searrow 0}(I_\delta^- + I_\delta^+) \le 0.$$

Combining these results, we obtain from (4.6)

$$u_1(\cdot,\tau) - u_2(\cdot,\tau) \le 0 \qquad \text{in} \quad \mathbb{R}\backslash\{0\},$$

which proves the theorem. □

As an immediate consequence we have the following.

COROLLARY 4.2. *Problem* P *has at most one solution* (u,f).

5. Oil trapping. The steady state solutions shown in Figure 3 suggest that oil may be trapped at the interface between coarse and fine material. Indeed, if $u_0(x) = 0$ for $x > 0$ and if for some $u^- \in (0, u^*]$

$$u_0(x) \le \left(u^- + \frac{1}{\sqrt{k^-}}\, x \right)_+ \qquad \text{for} \quad x < 0,$$

then the comparison principle guarantees

$$u(x,t) \le \left(u^- + \frac{1}{\sqrt{k^-}}\, x \right)_+ \qquad \text{for all} \quad (x,t) \in Q^-$$

and

$$u = 0 \quad \text{in} \quad Q^+.$$

The following theorem explains trapping in terms of the oil mass. For convenience, let

$$\bar{u}(x) := \begin{cases} \left(u^* + \frac{1}{\sqrt{k^-}}\, x \right)_+ & \text{for} \quad x < 0, \\ 0 & \text{for} \quad x > 0 \end{cases}$$

denote the maximal admissible steady state having $\bar{M}$, given by (1.12), as corresponding mass.

THEOREM 5.1. *Let u_0 satisfy hypothesis* (H) *and let*

$$\int_{-\infty}^{x} u_0(s)ds \ge \int_{-\infty}^{x} \bar{u}(s)ds \qquad \text{for} \quad x < 0.$$

Then the solution of Problem P *satisfies*

$$\int_{-\infty}^{0} u(s,t)ds \ge \bar{M} \qquad \text{for all} \quad t > 0.$$

Proof. Fix any $\delta > 0$ and set

$$V_\delta(x,t) = \int_{-\infty}^{x} u(s,t)ds + \delta \qquad \text{for} \quad (x,t) \in \bar{Q}.$$

Then $V_\delta \in C(\bar{Q}), V(\cdot,t) \in C^1((-\infty,0]) \cup C^1([0,\infty))$ for all $t > 0$, and

$$V_\delta = \delta \ \text{ to the left of the support of } u \text{ in } Q^-,$$

$$V_\delta = \int_{\mathbb{R}} u_0(s)ds + \delta \ \text{ to the right of the support of } u \text{ in } Q^+.$$

As a consequence $V_\delta \ge \bar{M}$ in Q^+, and it satisfies

$$V_t + V_x - \sqrt{k^-} V_x V_{xx} = 0 \qquad \text{a.e. in} \quad Q^-. \tag{5.1}$$

Setting

$$\bar{v}(x) := \int_{-\infty}^{x} \bar{u}(s)ds \qquad \text{for} \quad x \in (-\infty, 0],$$

we have

$$V_\delta > \bar{v} \qquad \text{in} \quad Q_t^- := (-\infty, 0] \times (0, t)$$

for t sufficiently small. Let

$$t_0 = \sup\{t > 0 : V_\delta > \bar{v} \quad \text{in} \quad Q_t^-\}.$$

Below we show $t_0 = \infty$. Suppose $t_0 < \infty$. Then there exists $(x_0, t_0) \in \bar{Q^-}$ such that

$$V_\delta > \bar{v} \qquad \text{in} \quad Q_{t_0}^- \tag{5.2}$$

and

$$V_\delta(x, t_0) \geq \bar{v}(x) \qquad \text{for all} \quad x \in (-\infty, 0] \quad \text{with} \quad V_\delta(x_0, t_0) = \bar{v}(x_0). \tag{5.3}$$

We first rule out $x_0 = 0$.

If $x_0 = 0$, we distinguish the three following cases:

(i) $u(0^-, t_0) > u^*$. Then we have

$$\frac{\partial V_\delta}{\partial x}(0^-, t_0) = u(0^-, t_0) > u^* \ = \frac{d\bar{v}}{\partial x}(0^-).$$

This contradicts (5.3).

(ii) $u(0^-, t_0) < u^*$. By continuity there exists $\varepsilon > 0$ such that $u(0^-, t) < u^*$ and $u(0^+, t) = 0$ for $t_0 - \varepsilon < t < t_0$. Since $f^-(t) = f^+(t) \leq 0$ for almost all $t \in (t_0 - \varepsilon, t_0)$ (see also section 6), we find from integrating the u-equation in $(-\infty, 0) \times (t_0 - \varepsilon, t_0)$

$$\int_{-\infty}^{0} u(s, t_0) ds - \int_{-\infty}^{0} u(s, t_0 - \varepsilon) ds = -\int_{t_0 - \varepsilon}^{t_0} f^-(t) dt \geq 0.$$

Hence

$$V_\delta(0, t_0 - \varepsilon) \leq V_\delta(0, t_0) = \bar{v}(0),$$

which contradicts (5.2).

(iii) $u(0^-, t_0) = u^*$. Then $V_\delta(0^-, t_0) = \bar{v}(0)$ as well as

$$\frac{\partial V_\delta}{\partial x}(0^-, t_0) = \frac{d\bar{v}}{dx}(0^-) = u^*.$$

Using (5.1) locally in Q^- and the strong maximum principle, we again obtain a contradiction.

Hence $x_0 \neq 0$ and $V_\delta(0, \cdot) > \bar{v}(0)$ in $[0, t_0]$. We then apply the comparison principle to (5.1) in $Q_{t_0}^-$ to find $V_\delta > \bar{v}$ in $(-\infty, 0] \times [0, t_0]$. This shows that $t_0 = \infty$. As a consequence $V_\delta > \bar{v}$ in $\bar{Q^-}$ for any $\delta > 0$, which implies the assertion of the theorem. □

Similarly we show the following.

THEOREM 5.2. *Let u_0 satisfy hypothesis* (H) *and let*

$$\int_x^\infty u_0(s) ds \leq \int_x^\infty \bar{u}(s) ds \qquad \textit{for } x \in \mathbb{R}.$$

Then

$$u = 0 \qquad \textit{in } \bar{Q}^+.$$

6. Closing remarks. In this section we briefly discuss some qualitative properties of solutions of Problem P.

6.1. Nonuniqueness. In the proof of the comparison principle, implying uniqueness, we have used the pressure condition

$$[p] \geq 0. \tag{6.1}$$

By means of a counterexample we show here that uniqueness fails if we drop condition (6.1). Let u_0 satisfy the structural properties

$$(\tilde{\mathrm{H}}) \qquad \begin{cases} u_0(x) = 0 \quad \text{if} \quad x > 0, \quad u_0 \not\equiv \bar{u} \quad \text{in} \ \mathbb{R}, \\ \bar{u}(x) \leq u_0(x) \leq (u^* + \delta x)_+ \quad \text{if} \ \ x < 0 \ \ \text{for some} \ \ 0 < \delta < \dfrac{1}{\sqrt{k^-}}. \end{cases}$$

Based on the results of section 5, we expect that the corresponding solution u of Problem P will have a nontrivial component in Q^+; i.e., $u \not\equiv 0$ in Q^+. We will construct a second solution $\tilde{u}$ which solves Problem P, except condition (6.1), and which satisfies $\tilde{u} \equiv 0$ in Q^+. This construction is based on a modification of k. Instead of (1.2) we consider

$$\tilde{k}_n(x) = \begin{cases} k^- & \text{for} \quad x < 0, \\ \kappa & \text{for} \quad 0 < x < \dfrac{1}{n}, \\ k^+ & \text{for} \quad x > \dfrac{1}{n}, \end{cases} \tag{6.2}$$

where $0 < \kappa < k^+ < k^-$, and we let $n \to \infty$.

THEOREM 6.1. *Let u_0 satisfy hypotheses* (H) *and* ($\tilde{\mathrm{H}}$) *and let u denote the unique solution of Problem* P. *Then*

(i) *$u \not\equiv 0$ in Q^+;*

(ii) *there exists a second solution $\tilde{u}$ of Problem* P, *except* (6.1), *which satisfies $\tilde{u} \equiv 0$ in Q^+.*

Proof. We first show that $u \not\equiv 0$ in Q^+. Arguing by contradiction, we assume

$$u(0^+, t) = 0 \qquad \text{for all} \quad t > 0.$$

Using $[p] \geq 0$ and $u \geq \bar{u}$ in Q, we conclude

$$u(0^-, t) = u^* \qquad \text{for all} \quad t > 0.$$

Hence u solves in Q^- the problem

$$(\mathrm{P}^-) \qquad \begin{cases} u_t + (u - \sqrt{k^-} u u_x)_x = 0 & \text{in} \quad Q^-, \\ u(0,t) = u^* & \text{for} \quad t > 0, \\ u(x,0) = u_0(x) & \text{for} \quad x < 0. \end{cases}$$

Now observe that $\bar{z} := (u^* + \delta x)_+$ is a supersolution for problem (P^-). Hence the solution $z(x,t)$ of problem (P^-) with initial data $z(\cdot, 0) = \bar{z}(x)$ is decreasing with respect to time and converges to a steady state solution $s(x)$. By comparison $s \geq \bar{u}$ in $\mathbb{R}^-$, but since $\bar{u}$ is maximal we have

$$s = \bar{u} \qquad \text{in} \ \ \mathbb{R}^-.$$

Using

$$\bar{u}(x) \le u(x,t) \le z(x,t) \qquad \text{for all} \quad (x,t) \in Q^-,$$

we obtain

$$\lim_{t\to\infty} u(x,t) = \bar{u}(x) \qquad \text{uniformly in} \ \ x < 0.$$

Combining this result with $u \equiv 0$ in Q^+, we find

$$\lim_{t\to\infty} \int_{-\infty}^{+\infty} u(x,t)ds \to \int_{-\infty}^{+\infty} \bar{u}(x)dx < \int_{-\infty}^{+\infty} u_0(x)dx,$$

which contradicts mass conservation for u.

Next we use (6.2) to explain the construction of $\tilde{u}$. As a first observation we note that the class of steady state solutions of the equation

$$\left(u - \tilde{k}_n \left(\frac{1+u}{\sqrt{\tilde{k}_n}} \right)' \right)' = 0 \qquad \text{in} \quad \mathbb{R},$$

having compact support and satisfying (M_1) and (M_2), has the same structure as the one shown in Figure 3, but with $u^* = \sqrt{\frac{k^-}{k^+}} - 1$ replaced by $\tilde{u}^* = \sqrt{\frac{k^-}{\kappa}} - 1$. In particular this class does not depend on n. For κ sufficiently small we find, for $\bar{\tilde{u}}$, the maximal steady state

$$u_0 \le \bar{\tilde{u}} \qquad \text{in} \quad \mathbb{R}.$$

As a consequence, the solution $\tilde{u}_n$ of the problem

$$\begin{cases} u_t + \left(u - \tilde{k}_n \left(\dfrac{1+u}{\sqrt{\tilde{k}_n}} \right)_x \right)_x & \text{in} \quad Q, \\ u(x,0) = u_0(x) \qquad \text{for} \quad x \in R \end{cases}$$

satisfies

$$\tilde{u}_n(x,t) \le \bar{\tilde{u}}(x) \qquad \text{for all} \quad (x,t) \in Q.$$

In particular

$$\tilde{u}_n \equiv 0 \qquad \text{in} \quad Q^+$$

for all $n \in \mathbb{Z}^+$. Finally, letting $n \to \infty$, $\tilde{u}_n$ converges along subsequences to a function $\tilde{u} = \tilde{u}(x,t)$ which satisfies all properties required for Problem P except (6.1). □

6.2. Waiting times and optimal regularity. Numerical simulations reported in [DMN] show that the right free boundary of u has a "waiting time" when it reaches the permeability discontinuity. The free boundary becomes stagnant there, while the oil saturation increases. It continues whenever the pressure exceeds the entry pressure of the low permeable region.

The following makes this precise.

THEOREM 6.2. *Let u_0 satisfy hypothesis* (H) *and let* $\mathrm{supp}(u_0) \subset \mathbb{R}^-$. *Further, let the solution u of Problem* P *satisfy $u \not\equiv 0$ in Q^+. Set*

$$t_1 := \lim_{\varepsilon\to 0} \sup\{\tau > 0 : u \equiv 0 \ \ \textit{in} \ \ (-\varepsilon,\infty) \times (0,\tau)\}$$

and

$$t_2 := \sup\{\tau > 0 : u \equiv 0 \quad in \quad \mathbb{R}^+ \times (0,\tau)\}.$$

Then

$$0 < t_1 < t_2 < \infty \qquad (t_2 - t_1 \text{ is the waiting time})$$

and

$$u(0^-, t_1) = 0, \qquad u(0^-, t_2) = u^*.$$

Proof. Clearly t_1 and t_2 are well defined. Continuity of $u^{\pm}(t)$ and (M_2) imply directly $t_2 > t_1$ and $u(0^-, t_1) = 0$.

Suppose $u(0^-, t_2) < u^*$. By continuity, there exists $\delta > 0$ such that $u(0^-, t) < u^*$, and thus $u(0^+, t) = 0$, for $t_2 \le t < t_2 + \delta$. Thus $u \equiv 0$ in $\mathbb{R}^+ \times (0, t_2 + \delta)$, contradicting the definition of t_2. □

Next we consider the case where the oil initially is positioned in the fine material ($x > 0$). If the initial position is sufficiently close to the interface at $x = 0$, diffusion may drive the oil towards $x = 0$, i.e., against the flow, where it will penetrate the coarse material. This follows from the transformation $y = x - t,\ t = t$ and by considering an appropriate subsolution for the resulting porous media equation; see [G2].

Supposing the oil reaches $x = 0$, we have the following result.

THEOREM 6.3. *Let u_0 satisfy hypothesis* (H) *and let* $\operatorname{supp}(u_0) \subset \mathbb{R}^+$. *Further, let the solution u of Problem* P *satisfy $u \not\equiv 0$ in Q^-. Set*

$$t_1 := \sup\{\tau > 0 : u \equiv 0 \quad in \quad \mathbb{R}^- \times (0,\tau)\}.$$
$$t_2 := \sup\{\tau > 0 : u(0^+, t) = 0 \quad for \quad 0 < t < \tau\}.$$

Then

$$0 < t_1 < t_2 \le \infty.$$

In addition, there exists $t \in (t_1, t_2)$ such that for some $A > 0$

$$u(x,t) = A\sqrt{x}(1 + o(1)) \qquad as \quad x \to 0^+.$$

Proof. By the finite speed of propagation we have $t_1 > 0$. Continuity of $u(0^-, \cdot)$ implies $u(0^-, t_1) = 0$ and $u(0^-, t) \le u^*$ and hence $u(0^+, t) = 0$ for all t in an upper neighborhood of t_1. Hence $t_2 > t_1$. If $u(0^-, t) \le u^*$ for all $t > 0$, we have $t_2 = \infty$. Since $u \not\equiv 0$ in $\mathbb{R}^- \times (t_1, t_2)$ and $u(0^+, \cdot) = 0$ in (t_1, t_2), there exists $t \in (t_1, t_2)$ such that

$$f(t) = f^-(t) = f^+(t) < 0.$$

Hence, for this t fixed, setting $f(t) = -\mathcal{C}(\mathcal{C} > 0)$,

$$u - \sqrt{k^+} u u_x = -\mathcal{C}(1 + o(1)) \qquad \text{as} \quad x \to 0^+,$$

giving

$$\frac{1}{2}u^2(x,t) = \frac{\mathcal{C}}{\sqrt{k^+}} x(1 + o(1)) \qquad \text{as} \quad x \to 0^+. \qquad \square$$

Appendix. Proof of Lemma 2.1. Let $\varepsilon_n > 0$ be such that

$$\varepsilon_n = o\left(\frac{1}{n}\right) \qquad \text{as} \quad n \to \infty,$$

and set

$$u_{0n}(x) = \begin{cases} \sqrt{u_0^2(x - \frac{1}{n}) + \varepsilon_n^2} & \text{if} \quad x > \dfrac{1}{n}, \\ \sqrt{(u_0^+)^2 + \varepsilon_n^2} & \text{if} \quad x = \dfrac{1}{n}, \end{cases}$$

where $u_0^+ = \lim_{x \searrow 0} u_0(x)$. Since $|u_{0n}'(x)| \le |u_0'(x - \frac{1}{n})|$ for $x > \frac{1}{n}$, the uniform Lipschitz continuity of u_0 in $\mathbb{R}^+$ implies

$$u_{0n} \text{ is uniform Lipschitz continuous in } [\tfrac{1}{n}, \infty].$$

Since

$$f_0 = u_0 - \frac{1}{2}\sqrt{k^+}(u_0^2)' \qquad \text{in} \quad \mathbb{R}^+,$$

$$f_{0n} = u_{0n} - \frac{1}{2}\sqrt{k^+}(u_{0n}^2)' \qquad \text{in} \quad \left[\frac{1}{n}, \infty\right),$$

the total variation of $(u_0^2)'$ in $\mathbb{R}^+$, $TV_{\mathbb{R}^+}((u_0^2)')$, is bounded, and since $(u_{0n}^2)'(x) = (u_0^2)'(x - \frac{1}{n})$,

$$TV_{(\frac{1}{n}, \infty)}(f_{0n}) \to TV_{\mathbb{R}^+}(f_0) \qquad \text{as} \quad n \to \infty. \tag{A.1}$$

In order to extend u_{0n} to the interval $[-\frac{1}{n}, \infty]$ we distinguish two different cases: $u_0^+ > 0$ and $u_0^+ = 0$. At this point we remind the reader that the constant u^* is defined by

$$\frac{1 + u^*}{\sqrt{k^-}} = \frac{1}{\sqrt{k^+}}\,, \qquad \text{i.e.,} \quad u^* = \sqrt{\frac{k^-}{k^+}} - 1.$$

(i) *Case* $u_0^+ > 0$. We define u_{0n} in $[-\frac{1}{n}, \frac{1}{n})$ by the relation $p_{0n} = p_{0n}(\frac{1}{n})$ in $[-\frac{1}{n}, \frac{1}{n})$, i.e.,

$$u_{0n}(x) = -1 + \sqrt{\frac{k_n(x)}{k^+}}\left(1 + \sqrt{(u_0^+)^2 + \varepsilon_n^2}\right).$$

In particular, as $n \to \infty$,

$$\begin{aligned} u_{0n}(-\tfrac{1}{n}) &= -1 + \sqrt{\frac{k^-}{k^+}}\left(1 + \sqrt{(u_0^+)^2 + \varepsilon_n^2}\right) \\ &\to -1 + \sqrt{\frac{k^-}{k^+}}(1 + u_0^+) = u_0^-, \end{aligned} \tag{A.2}$$

where we have used, by hypothesis (H), $[p_0] = 0$ if $u_0^+ > 0$. Since $u_{0n}(-\frac{1}{n}) > u_0^-$, there exist $\delta_n > 0$ such that

$$u_{0n}\left(-\frac{1}{n}\right) = \sqrt{(u_0^-)^2 + \delta_n^2}. \tag{A.3}$$

It follows directly from the construction of u_{0n} that

$$TV_{(-\frac{1}{n},\frac{1}{n})}(f_{0n}) = -u_{0n}\left(\frac{1}{n}\right) + u_{0n}\left(-\frac{1}{n}\right) \to -[u_0] \quad \text{as} \quad n \to \infty \tag{A.4}$$

and

$$f_{0n}\left(\frac{1}{n}+\right) - f_{0n}\left(\frac{1}{n}-\right) = -\frac{1}{2}\sqrt{k^+}(u_{0n}^2)'\left(\frac{1}{n}+\right) = -\frac{1}{2}\sqrt{k^+}(u_0^2)'(0^+). \tag{A.5}$$

(ii) *Case* $u_0^+ = 0$. Since $[p_0] \geq 0$, $u_0^+ = 0$ implies that

$$0 \leq u_0^- \leq -1 + \sqrt{\frac{k^-}{k^+}} = u^*.$$

Hence

$$(1+\varepsilon_n)\sqrt{k^-} > \sqrt{k^-} \geq \sqrt{k^+}(1+u_0^-),$$

and there exist $\delta_n > 0$ such that

$$\begin{cases} \delta_n \to 0 \quad \text{as} \quad n \to \infty, \\ (1+\varepsilon_n)\sqrt{k^-} > \sqrt{k^+}\left(1+\sqrt{(u_0^-)^2+\delta_n^2}\right), \\ \sqrt{(u_0^-)^2+\delta_n^2} > \varepsilon_n. \end{cases}$$

These two inequalities imply that for some $\kappa_n \in (k^+, k^-)$

$$(1+\varepsilon_n)\sqrt{k^-} = \sqrt{\kappa_n}\left(1+\sqrt{(u_0^-)^2+\delta_n^2}\right).$$

Then there exists $x_n \in (-\frac{1}{n}, \frac{1}{n})$ such that

$$k_n(x_n) = \kappa_n,$$

and we define u_{0n} in $[-\frac{1}{n}, \frac{1}{n})$ by the relations

$$u_{0n}(x) \equiv u_{0n}\left(\frac{1}{n}\right) \quad (=\varepsilon_n) \qquad \text{if} \quad x_n \leq x < \frac{1}{n}$$

and

$$p_{0n}(x) \equiv p_{0n}(x_n) \quad \left(= \frac{1+\varepsilon_n}{\sqrt{\kappa_n}}\right) \qquad \text{if} \quad -\frac{1}{n} < x < x_n.$$

By the definition of κ_n and p_{0n}, the latter relation can be written as

$$u_{0n}(x) = -1 + \sqrt{\frac{k_n(x)}{k^-}}\left(1+\sqrt{(u_0^-)^2+\delta_n^2}\right) \qquad \text{if} \quad -\frac{1}{n} \leq x < x_n.$$

In particular we have

$$u_{0n}' \leq 0 \text{ in } \left(-\frac{1}{n}, x_n\right) \text{ and } u_{0n}\left(-\frac{1}{n}\right) = \sqrt{(u_0^-)^2+\delta_n^2} \to u_0^- \qquad \text{as } n \to \infty, \tag{A.6}$$

and

$$TV_{(-\frac{1}{n},x_n)}(f_{0n}) = TV_{(-\frac{1}{n},x_n)}(u_{0n}) \to -[u_0] \qquad \text{as } n \to \infty. \tag{A.7}$$

Since $|k_n^-| \le \frac{\mathcal{C}}{n}$ and $\varepsilon_n = o(\frac{1}{n})$ as $n \to \infty$, and since

$$f_{0n}(x) = \varepsilon_n + \frac{1}{2}\varepsilon_n(1+\varepsilon_n)\frac{k_n'(x)}{\sqrt{k_n(x)}} \qquad \text{if} \quad x_n < x < \frac{1}{n},$$

it follows that

$$TV_{(x_n,\frac{1}{n})}(f_{0n}) \to 0 \qquad \text{as} \quad n \to \infty. \tag{A.8}$$

In addition, as $n \to \infty$,

$$f_{0n}\left(\frac{1}{n}+\right) - f_{0n}\left(\frac{1}{n}-\right) \to -\frac{1}{2}\sqrt{k^+}(u_0^2)'(0^+) \tag{A.9}$$

and

$$f_{0n}(x_n^+) - f_{0n}(x_n^-) = \frac{1}{2}\varepsilon_n(1+\varepsilon_n)\frac{k_n'(x_n)}{\sqrt{\kappa_n}} \to 0. \tag{A.10}$$

Combining (A.6)–(A.10) gives

$$TV_{(-\frac{1}{n},\frac{1}{n})}(f_{0n}) \to -[u_0] \qquad \text{as} \quad n \to \infty. \tag{A.11}$$

Finally we have to define $u_{0n}(x)$ for $x < -\frac{1}{n}$. In view of (A.3) and (A.6) it seems natural to set

$$u_{0n}(x) = \sqrt{u_0^2\left(x+\frac{1}{n}\right) + \delta_n^2} \qquad \text{if} \quad x < -\frac{1}{n}. \tag{A.12}$$

Arguing as in the interval $(\frac{1}{n},\infty)$, we obtain as $n \to \infty$

$$TV_{(-\infty,-\frac{1}{n})}(f_{0n}) \to TV_{\mathbb{R}^-}(f_0) \tag{A.13}$$

and

$$f_{0n}\left(\left(-\frac{1}{n}\right)+\right) - f_{0n}\left(\left(-\frac{1}{n}\right)-\right) \to \frac{1}{2}\sqrt{k^-}(u_0^2)'(0^-). \tag{A.14}$$

Combining (A.1), (A.13), and (A.14) with, respectively, (A.4), (A.5) if $u_0^+ > 0$ and (A.9), (A.11) if $u_0^+ = 0$, we find

$$TV_{\mathbb{R}}(f_{0n}) \to TV_{\mathbb{R}}(f_0) \quad \text{as} \quad n \to \infty.$$

Now, if $\delta_n = \varepsilon_n$, u_{0n} satisfies all properties of Lemma 2.1. In general, however, $\delta_n \neq \varepsilon_n$ and we have to correct the construction of u_{0n} in $(-\infty,-\frac{1}{n})$. Since $u_{0n}(-\frac{1}{n}) > u_{0n}(\frac{1}{n}) \ge \varepsilon_n$, we can still use definition (A.12) in a neighborhood of $x = -\frac{1}{n}$. Since k_n is constant in $(-\infty,-\frac{1}{n})$, the expression for the flux is simply

$$f_{0n} = u_{0n} - \frac{1}{2}\sqrt{k^-}(u_{0n}^2)' \qquad \text{in} \quad \left(-\infty,-\frac{1}{n}\right).$$

Therefore it is not difficult to change slightly the definition of u_{0n} such that $u_{0n} \ge \varepsilon_n$ in $\mathbb{R}$ and $u_{0n}(x) = \varepsilon_n$ for $-x$ sufficiently large. We leave the details to the reader.

Acknowledgment. The authors gratefully acknowledge support of EU by the TMR program *Nonlinear Parabolic Partial Differential Equations: Methods and Applications*, FMRX-CT98-0201.

REFERENCES

[A] D.G. Aronson, *Regularity properties of flows through porous media*, SIAM J. Appl. Math., 17 (1969), pp. 461–467.

[DMP] C.J. van Duijn, A. Mikelic, and I.S. Pop, *Effective equations for two-phase flow with trapping on the micro scale*, SIAM J. Appl. Math., 62 (2002), pp. 1531–1568.

[DMN] C.J. van Duijn, J. Molenaar, and M.J. de Neef, *The effect of capillary forces on immiscible two-phase flow in heterogeneous porous media*, Transport in Porous Media, 21 (1995), pp. 71–93.

[DP] C.J. van Duijn and L.A. Peletier, *Nonstationary filtration in partially saturated media*, Arch. Ration. Mech. Anal., 78 (1982), pp. 173–198.

[F] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964.

[G1] B.H. Gilding, *Hölder continuity of solutions of parabolic equations*, J. London Math. Soc. (2), 13 (1976), pp. 103–106.

[G2] B.H. Gilding, *The occurrence of interfaces in nonlinear diffusion-advection processes*, Arch. Ration. Mech. Anal., 100 (1988), pp. 243–263.

[K] T.F.M. Kortekaas, *Water/oil displacement characteristics in crossbedded reservoir zones*, Soc. Petroleum Eng. J., 25 (1985), pp. 917–926.

[LSU] O.A. Ladyzhenskaya, V.A. Solonnikov, and N.N. Ural'ceva, *Linear and Quasi-Linear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.

[L] M.C. Leverett, *Capillary behavior in porous solids*, Trans. AIME Petr. Eng. Div., 142 (1941), pp. 152–169.

[W] K.J. Weber, *How heterogeneity affects oil recovery*, in Reservoir Characterization, L.W. Lake and H.B. Carrol, eds., Academic Press, Orlando, FL, 1986, pp. 487–544.

SIAM J. MATH. ANAL.
Vol. 35, No. 1, pp. 268–278

AN EXAMPLE OF CATASTROPHIC SELF-FOCUSING IN NONLINEAR OPTICS?*

ÉRIC DUMAS†

Abstract. As the wavelength ε goes to zero, the slowly varying envelope approximation allows one to replace the fields (solutions to Maxwell equations) with profile solutions to a nonlinear Schrödinger equation (NLS). Depending on the model, this equation may be critical and focusing, and then admits explosive solutions. In this case, the approximation breaks down, and, for ε fixed, the fields may be globally defined in time, and smooth. This happens in the case of Maxwell–Bloch equations [P. Donnat and J. Rauch, *Arch. Ration. Mech. Anal.*, 136 (1996), pp. 291–303], [E. Dumas, *Existence globale pour les systèmes de Maxwell-Bloch*, in Séminaire École Polytechnique, 2002–2003, Ecole Polytechnique, Palaiseau, France], of the anharmonic oscillator with saturated nonlinearity [J. L. Joly, G. Métivier, and J. Rauch, *SIAM J. Math. Anal.*, 227 (1996), pp. 903–913], and of propagation in a ferromagnetic medium [J. L. Joly, G. Métivier, and J. Rauch, *Ann. Henri Poincaré*, 1 (2000), pp. 307–340], [H. Haddar, *Modèles asymptotiques en ferromagnétisme: Couches minces et homogénéisation*, Ph.D. thesis, thèse INRIA-École Nationale des Ponts et Chaussées, 2000].

We analyze the question of self-focusing for a wave equation in space dimension 2; the same techniques apply to usual models in greater dimensions. We give a new representation of the fields in terms of oscillating profiles, ruled by focusing rays. Furthermore, we prove that the approximation by an explosive solution of NLS is valid up to a time of the order of a negative power of $\ln(1/\varepsilon)$ before explosion; this exhibits an amplification of the fields by a positive power of $\ln(1/\varepsilon)$ between $t = 0$ and that time.

Key words. nonlinear diffractive optics, nonlinear critical Schrödinger equation, self-focusing

AMS subject classifications. 35L, 35Q60, 78A60

DOI. 10.1137/S0036141002414482

1. Introduction.

A standard model for describing the propagation of an electromagnetic wave through a field responsive (or Kerr) medium is the coupling of Maxwell equations with an anharmonic oscillator (cf. [3], [2]):

$$\begin{cases} \partial_t E = -\text{curl } B - \partial_t P, \\ \partial_t B = \text{curl } E, \\ \varepsilon^2 \partial_t^2 P + \nabla_P V(P) = \gamma E. \end{cases} \tag{1}$$

Here, (E, B) is the electromagnetic field, and P is the polarization of the medium. The physically relevant fields also satisfy $\text{div}\,(E + P) = \text{div}\, B = 0$, which is true for all times as soon as it is at one given time. The response of matter is given by a nonlinear spring force, with the same frequency $1/\varepsilon$ as the wave. Many other models are available, including magnetic effects (Landau–Lifshitz model; see [14]) or quantum descriptions (Bloch equations; see [18]); but all these systems can be approximated by a single equation, in the regime we describe now.

For a small amplitude wave, slowly modulated in space and time, the so-called "slowly varying envelope approximation" (see [16, pp. 32–38] and [17]) leads to a *nonlinear Schrödinger equation* (NLS). This approximation is also used in other physical

*Received by the editors September 13, 2002; accepted for publication (in revised form) January 31, 2003; published electronically July 8, 2003.
http://www.siam.org/journals/sima/35-1/41448.html

†Laboratoire de Mathématiques et Physique Théorique, Parc de Grandmont, 37200 Tours, France. Present address: Université Grenoble 1, BP 74, 100, rue des Mathématiques, Domaine Universitaire, 38402 Saint Martin d'Hères, France (edumas@ujf-grenoble.fr, http://www-fourier.ujf-grenoble.fr/∼edumas/).

contexts, such as water waves and plasma waves, showing that NLS is a canonical equation, capturing the essential features of nonlinear wave dynamics (see [23] for a survey). In the case of Maxwell equations with an anharmonic oscillator, the potential V is replaced by its Taylor expansion at the origin,

$$V(P) \simeq \alpha|P|^2 - \beta|P|^4,$$

and the vector $u = (E, B, P, \varepsilon\partial_t P)$ is approximated by $u_{app} = \varepsilon\mathcal{U}(\varepsilon t, y_1, y_2)e^{i\frac{y_3+t}{\varepsilon}}$, $\mathcal{U} = (\mathcal{E}, \mathcal{B}, \mathcal{P}, \mathcal{Q})$. The fields must be polarized,

$$\mathcal{E} = \begin{pmatrix} K \\ L \\ 0 \end{pmatrix}, \ \mathcal{B} = c_1 \begin{pmatrix} L \\ -K \\ 0 \end{pmatrix}, \ \mathcal{P} = c_2 \begin{pmatrix} K \\ L \\ 0 \end{pmatrix}, \ \mathcal{Q} = ic_3 \begin{pmatrix} K \\ L \\ 0 \end{pmatrix}, \tag{2}$$

and the amplitudes $K(T, y_1, y_2), L(T, y_1, y_2)$ are solutions to

$$i\partial_t \begin{pmatrix} K \\ L \end{pmatrix} - \kappa_1 \Delta_{y_1,y_2} \begin{pmatrix} K \\ L \end{pmatrix} - \kappa_2(|K|^2 + |L|^2) \begin{pmatrix} K \\ L \end{pmatrix} = 0. \tag{3}$$

This is a critical NLS equation in space dimension 2, possibly with *focusing* nonlinearity ($\kappa_1\kappa_2 > 0$), depending on the details of the model (i.e., on the coefficients α and β).

This approximation can be rigorously justified in the case of a general hyperbolic system of nonlinear partial differential equations (see [5]), for ε small enough, on any time interval $[0, t_1/\varepsilon]$ such that the solution of (3) remains smooth on $[0, t_1]$. (Similar results hold for water waves in the limit of long waves, leading to the Korteveg–de Vries equation or Boussinesq equation; see [19], [20], [21].) However, a time $t_\star$ of explosion for a solution to (3) is usually thought of as an indication of *self-focusing*: a variation of the refractive index of the medium induces curved light rays, which concentrate in the region of maximal refractive index. But, to our knowledge, there is no rigorous result linking the explosion of the profile (solution to NLS) and the behavior of the exact solution u.

First, the slowly varying envelope assumption (from which (3) is derived) is violated in this region, where fields become too large. Second, the electromagnetic field may be globally defined in time (and smooth), even if the profile $\mathcal{U}$ explodes in finite time: in [6], Donnat and Rauch consider two-level Maxwell–Bloch systems. This result is extended in [8] to general Maxwell–Bloch systems, following ideas of Joly, Métivier, and Rauch [13] and Haddar [11] concerning the Landau–Lifshitz model. In [12], Joly, Métivier, and Rauch deal with (1), when the potential V is *saturated*: if the second and third derivatives of V are bounded, $H^2(\mathbb{R}^3)$ initial data generate global solutions to (1).

In this paper, we investigate further the mechanism of self-focusing, evaluating more precisely how long the Schrödinger approximation is valid. We present our method for the simplest (i.e., with the lowest space dimension) hyperbolic equation allowing transverse diffraction, a scalar wave equation in space dimension 2:

$$\Box u + iF(\partial_t u) = 0, \text{ with } F(z) = |z|^4 z. \tag{4}$$

THEOREM 1.1. *Fix $t_\star > 0$, and define*

$$a_0(t, Y) := (t_\star - t)^{-1/2} e^{i\frac{Y^2-1}{2(t_\star-t)}} R\left(\frac{Y}{t_\star - t}\right), \text{ with } R(Y) = \frac{3^{1/4}}{\sqrt{\text{ch }(2Y)}}. \tag{5}$$

There are $\varepsilon_0, C > 0$ such that for all $\varepsilon \in]0, \varepsilon_0]$, the initial value problem associated with (4)*, for initial data*

$$\begin{cases} u^\varepsilon_{|t=0} = \varepsilon a_0\left(0, y_2/\sqrt{\varepsilon}\right) e^{iy_1/\varepsilon}, \\ \partial_t u^\varepsilon_{|t=0} = \dfrac{i}{\varepsilon} u^\varepsilon_{|t=0} + \mathcal{O}(\varepsilon) \text{ in } \mathcal{S}(\mathbb{T}_\varepsilon \times \mathbb{R}), \text{ with } \mathbb{T}_\varepsilon \text{ the torus } \mathbb{R}/(2\pi\varepsilon), \end{cases}$$

admits a unique smooth solution $u^\varepsilon \in \mathcal{C}^1(\mathcal{S}(\mathbb{T}_\varepsilon \times \mathbb{R}))$ for $t \in [0, t_\star - C(\ln 1/\varepsilon)^{-1/3}]$. Furthermore, as $\varepsilon \to 0$, we have the approximation

$$\left\| \partial_t u^\varepsilon - i a_0\left(t, y_2/\sqrt{\varepsilon}\right) e^{i\frac{y_1+t}{\varepsilon}} \right\|_{L^\infty_y} = o\left(\|a_0(t)\|_{L^\infty}\right).$$

The link with the physical context above will be clearer after a few remarks.

Remark 1.1. From the profile $a_0(t, Y)e^{i\theta}$, in which we substitute $Y = y_2/\sqrt{\varepsilon}$, $\theta = (y_1 + t)/\varepsilon$, we recover the long-time setting leading to (3) after rescaling: replace (t, y) by $\sqrt{\varepsilon}(t, y)$, and $\sqrt{\varepsilon}$ by ε. We have chosen these scales so as to give an alternative description of u^ε in terms of curved phases and focusing rays. See section 4 and [7]. Once rescaled, Theorem 1.1 shows that the slowly varying envelope approximation for (4) can be justified up to times $\frac{1}{\varepsilon}(t_\star - C(\ln 1/\varepsilon)^{-1/3})$—compare with the previous t_1/ε, when $t_1 < t_\star$ is fixed.

Remark 1.2. The same method can be applied to a general hyperbolic system of PDEs, also in higher dimension (see Remark 2.1), provided that the NLS equation for the envelope, analogous to (3), is critical and focusing. For example, in the case of system (1), with space dimension 3 and cubic nonlinearity ($\nabla_P V(P) = 2\alpha P - 4\beta|P|^2 P$ for suitable α and β), the exact solution $u = (E, B, P, \varepsilon\partial_t P)$ is approximated in $L^\infty([t_\star - C(\ln 1/\varepsilon)^{-1/3}] \times \mathbb{R}^3)$ via $K_{app} = (t_\star - \varepsilon t)^{-1} e^{i\frac{|y_1, y_2|^2 - 1}{2(t_\star - \varepsilon t)}} R_2(\frac{y_1}{t_\star - \varepsilon t}, \frac{y_2}{t_\star - \varepsilon t}) e^{i\frac{y_3 + t}{\varepsilon}}$ from (2), with $R_2 \in \mathcal{S}(\mathbb{R}^2)$ a positive solution to $-\Delta R + R - R^3 = 0$.

Remark 1.3. Equation (4) preserves only $\|\partial_{t,y} u\|_{L^2}$ (taking the real part of (4) times $\partial_t \overline{u}$). This is not sufficient to guarantee global existence of u: when the maximal existence time $t^\star$ of a smooth solution u is finite, the quantity $\|\partial_t u(t)\|_{L^\infty}$ explodes as $t \to t^\star$. Thus, the standard proofs of global existence, such as for small data (see [22]), consist in controlling $\|\partial_t u(t)\|_{L^\infty}$ by t and the initial data. Here, $\|\partial_{t,y} u\|_{L^2}$ does not allow this control. The theorem doesn't prove that u^ε (with ε fixed) explodes at $t = t_\star$ but shows an amplification of $\|\partial_t u\|_{L^\infty}$ by a factor $(\ln 1/\varepsilon)^{-1/6}$ between $t = 0$ and $t = t_\star - C(\ln 1/\varepsilon)^{-1/3}$.

Remark 1.4. Global existence is achieved in the other limit: ε fixed, $t_\star \to 0$, which corresponds to small initial data. Saturation, replacing the nonlinearity $F(z)$ by $G^\varepsilon(z) = \frac{|z|^4 z}{1 + \varepsilon z^3}$, also ensures global existence of u^ε, as in [12]. Thus, the mechanism of catastrophic self-focusing seems to be as follows: first, a concentration due to linear focusing of rays ("self-focusing"); second, activation of nonlinear effects by this amplification. Blow-up ("catastrophic" self-focusing) then depends on the strength of the nonlinearity: saturation stops the development of the singularity, but without saturation, or special geometric properties (see [13], [11] about the Landau–Lifshitz model, and [6], [8] about Maxwell–Bloch systems), one is inclined to think that blow-up occurs.

Remark 1.5. In fact, the slowly varying envelope approximation used here necessarily produces a Schrödinger equation with a restrictive kind of nonlinearity: namely, the first nonvanishing term in the Taylor expansion of the original nonlinearity. Since

the structure of the nonlinear terms is crucial for blow-up, there are also attempts for understanding the role played by various perturbations of NLS: see [9] and [4].

The paper is organized as follows:

Section 2: (formal) definition of the explosive profile a_0 via the conformal invariance of NLS.

Section 3: Wentzel–Kramers–Brillouin (WKB) asymptotics. A corrector a_c is defined to get a better approximation.

Section 4: thanks to nonuniqueness of the profile representation, we give an alternative description of u^ε based on focusing nonplanar phases.

Section 5: proof of Theorem 1.1. We first change scales (section 5.1) and look at $U(x) = u(t, y_1/\varepsilon, y_2/\sqrt{\varepsilon})$, 2π-periodic in y_1. In section 5.2, we write down energy estimates for the residual $\partial_t V = \partial_t U - \partial_t U_{app}$ ($\|\partial_t V\|_{L^\infty}$ is then controlled by Sobolev's inequality). This is the notable difference between this work and [15], where the authors need global existence of the (small) approximate solution, whereas we "follow" the explosive approximate solution up to some boundary layer before $t_\star$. This boundary layer appears when requiring the corrector a_c to remain small compared to a_0 (section 5.3) and in the bootstrap argument showing $\|\partial_t V\|_{L^\infty} \ll \|\partial_t U_{app}\|_{L^\infty}$ (section 5.4).

2. From the wave equation to explosive solutions of NLS. A classical technique for constructing explosive solutions to NLS comes from the pseudoconformal invariance of this equation (see [10]): if $b(t, Y)$ is a solution to

$$2i\partial_t u - \partial_Y^2 u - |u|^4 u = 0, \tag{6}$$

then, for all $t_\star \in \mathbb{R}$, we define another solution by

$$a(t, Y) := \left(t^{1/2} e^{iY^2/2t} b\right)\left(\frac{1}{t_\star - t}, \frac{Y}{t_\star - t}\right) = (t_\star - t)^{-1/2} e^{iY^2/2(t_\star - t)} b\left(\frac{1}{t_\star - t}, \frac{Y}{t_\star - t}\right).$$

When seeking a solution u to (4) in the "slowly varying envelope" form $u^\varepsilon(x) = \varepsilon \mathcal{U}^\varepsilon(t, y_2/\sqrt{\varepsilon}, (y_1 + t)/\varepsilon)$ with a profile $\mathcal{U}^\varepsilon(t, Y, \theta) \in \mathcal{C}^2([0, t_\star[\times \mathbb{R} \times \mathbb{T})$ (periodic w.r.t. the last variable, θ), the chain rule leads to the following equation:

$$[2\partial_t \partial_\theta - \partial_Y^2]\mathcal{U}^\varepsilon + \varepsilon \partial_t^2 \mathcal{U}^\varepsilon + i|\partial_\theta \mathcal{U}^\varepsilon + \varepsilon \partial_t \mathcal{U}^\varepsilon|^4 (\partial_\theta \mathcal{U}^\varepsilon + \varepsilon \partial_t \mathcal{U}^\varepsilon) = 0. \tag{7}$$

When $\mathcal{U}^\varepsilon$ satisfies (7), u^ε above is a solution to (6).

In order to let these quantities vanish at first order (see the WKB expansions in section 3), it is then natural to look for a profile $u_0(t, Y, \theta)$ such that

$$[2\partial_t \partial_\theta - \partial_Y^2]u_0 + i|\partial_\theta u_0|^4 \partial_\theta u_0 = 0. \tag{8}$$

There are explicit solutions to this equation. When $u_0 = b(t, Y)e^{i\theta}$, it is equivalent to require that b satisfies (6), and $b(t, Y) = e^{-it/2} R(Y)$ is a solution, with $R(Y) = 3^{1/4}(\mathrm{ch}\,(2Y))^{-1/2}$ (the unique positive solution of $R'' - R + R^5 = 0$, up to translation).

Now, use the pseudoconformal invariance of (6) to get a solution $u_0(t, Y, \theta) = a_0(t, Y)e^{i\theta}$ to (8):

$$a_0(t, Y) = (t_\star - t)^{-1/2} e^{i(Y^2 - 1)/2(t_\star - t)} R\left(\frac{Y}{t_\star - t}\right). \tag{9}$$

For all $t \in [0, t_\star[$, $a_0(t) \in \mathcal{S}(\mathbb{R})$, and a_0 explodes at $t = t_\star$ ($\|a_0(t)\|_{L^\infty} = 3^{1/4}(t_\star - t)^{-1/2}$).

Remark 2.1. The same construction is possible in higher dimension N: the pseudoconformal transform is $u(t,x) \mapsto t^{-N/2}e^{-i|x|^2/2t}\overline{u}(1/t, x/t)$. Then $-\Delta R + R - R^{1+4/N} = 0$ also has a solution in $\mathcal{S}$, and the same is valid for the equation $-\Delta R + mR + g(R) = 0$, $m = cst$, under suitable behavior of g at the origin and at infinity (see [1]).

3. WKB expansions and initial data for u^ε. If we want to deduce from an approximate solution $\varepsilon\mathcal{U}^\varepsilon(t, y_2/\sqrt{\varepsilon}, (y_1+t)/\varepsilon)$ the existence of an exact solution to (4), we must construct a solution to (7) to higher order than $\varepsilon u_0(t, y_2/\sqrt{\varepsilon}, (y_1+t)/\varepsilon)$. That's why we need a corrector for the first profile u_0. The general form for $\mathcal{U}^\varepsilon_{app}$ (from [5]) has two such correctors: $\mathcal{U}^\varepsilon_{app} = \varepsilon(u_0 + \sqrt{\varepsilon}u_1 + \varepsilon u_2)$. Here, when u_1 vanishes at $t = 0$, we can let it vanish for all times. Thus, we set

$$u^\varepsilon_{app}(x) = \varepsilon\mathcal{U}^\varepsilon_{app}(t, y_2/\sqrt{\varepsilon}, (y_1+t)/\varepsilon), \quad \mathcal{U}^\varepsilon_{app} = u_0 + \varepsilon u_c = (a_0 + \varepsilon a_c)(t, Y)e^{i\theta}.$$

Proposition 3.1. *We have the (formal) WKB expansion*

$$\Box u^\varepsilon_{app} + i|\partial_t u^\varepsilon_{app}|\partial_t u^\varepsilon_{app} = (\mathcal{E}_0 + \varepsilon\mathcal{E}_1 + \mathcal{R}^\varepsilon)(t, y_2)e^{i\frac{y_1+t}{\varepsilon}}, \tag{10}$$

with

$$\begin{aligned}
\mathcal{E}_0(t,Y) &= 2i\partial_t a_0 - \partial_Y^2 a_0 - |a_0|^4 a_0,\\
\mathcal{E}_1(t,Y) &= 2i\partial_t a_c - \partial_Y^2 a_c + \partial_t^2 a_0 + G(a_0, a_c),\\
G(a_0, a_c) &= -|a_0|^4(\partial_t a_0 + i a_c) + 4i a_0|a_0|^2 Re(a_0(\partial_t\overline{a_0} + i\overline{a_c})),\\
\mathcal{R}^\varepsilon(t,Y) &= \varepsilon^2\partial_t^2 a_c + iF((i+\varepsilon\partial_t)(a_0 + \varepsilon a_c)) + F(a_0) - \varepsilon G(a_0, a_c).
\end{aligned}$$

We can construct a_0 and a_c such that $\mathcal{E}_0 = \mathcal{E}_1 = 0$: such an $a_0 \in \mathcal{C}^\infty([0, t_\star[\times\mathbb{R})$ is given by (9), and $\mathcal{E}_1 = 0$ is a linear Schrödinger equation, which has a unique solution for any $a_{c|_{t=0}} \in L^2(\mathbb{R})$.

Our goal is to show the existence of an exact solution u^ε to (4) close to $\varepsilon u_0(t, y_2/\sqrt{\varepsilon}, (y_1+t)/\varepsilon)$. Towards this end, we choose

$$a_{c|_{t=0}} = 0, \tag{11}$$

which provides us with a (unique) corrector $a_c \in \mathcal{C}^\infty([0, t_\star[\times\mathbb{R})$. Next, we take the simplest initial data for u^ε, in view of evaluating $u^\varepsilon - u^\varepsilon_{app}$:

$$\begin{cases} u^\varepsilon{}_{|_{t=0}} = u^\varepsilon_{app}{}_{|_{t=0}},\\ \partial_t u^\varepsilon{}_{|_{t=0}} = \partial_t u^\varepsilon_{app}{}_{|_{t=0}}. \end{cases} \tag{12}$$

Remark 3.1. (i) We can compute $\partial_t u^\varepsilon_{app}{}_{|_{t=0}}$ in terms of the function R, since a_0 is known explicitly, and $\partial_t a_c$ is given by the equation $\mathcal{E}_1 = 0$, so that

$$\partial_t a_{c|_{t=0}} = \frac{i}{2}\left[\partial_t^2 a_0 - |a_0|^4\partial_t a_0 + 2ia_0|a_0|^2(a_0\partial_t\overline{a_0} + \overline{a_0}\partial_t a_0)\right]_{|_{t=0}}. \tag{13}$$

(ii) Since the data are $2\pi\varepsilon$-periodic in y_1, the standard uniqueness argument shows that so is $u^\varepsilon(t)$ for each time.

4. Linear focusing. We can give an alternative profile description of the data in (12): $u^\varepsilon{}_{|t=0}$ also has a representation via $\tilde{u}_0^0 \in \mathcal{S}(\mathbb{R}\times\mathbb{T})$,

$$u^\varepsilon{}_{|t=0} = \varepsilon \tilde{u}_0^0\left(\frac{y_2}{\sqrt{\varepsilon}}, \frac{y_1+y_2^2/2t_\star}{\varepsilon}\right), \text{ where } \tilde{u}_0^0(Y,\theta) = t_\star^{-1/2} e^{-i/2t_\star} R\left(\frac{Y}{t_\star}\right) e^{i\theta}.$$

Similarly,

$$\partial_t u^\varepsilon{}_{|t=0} = \Bigg(i t_\star^{-1/2} R\left(\frac{Y}{t_\star}\right) + \varepsilon \Bigg[\frac{i}{2} t_\star^{-5/2} Y^2 R\left(\frac{Y}{t_\star}\right) + t_\star^{-5/2} Y R'\left(\frac{Y}{t_\star}\right) + \frac{1}{2} t_\star^{-3/2}(1 - i t_\star^{-1}) R\left(\frac{Y}{t_\star}\right)\Bigg] e^{-i/2t_\star} e^{i\theta}\Bigg)_{|_{Y=y_2/\sqrt{\varepsilon}, \theta=(y_1+y_2^2/2t_\star)/\varepsilon}} + \varepsilon^2 \partial_t a_{c|_{t=0, Y=y_2/\sqrt{\varepsilon}}} e^{i y_1/\varepsilon},$$

and $e^{i\frac{y_1+y_2^2/2t_\star}{\varepsilon}}$ also factors in the last term (see (13) in Remark 3.1).

Thus, we have a new profile representation of the initial data, with oscillations involving the *curved* phase $y_1 + y_2^2/2t_\star$. Now, defining v^ε such that

$$\begin{cases} \Box v^\varepsilon + iF(\partial_t v^\varepsilon) = 0, \\ v^\varepsilon{}_{|t=0} = u^\varepsilon{}_{|t=0}, \\ \partial_t v^\varepsilon{}_{|t=0} = i\sqrt{1+(y_2/t_\star)^2}\, v^\varepsilon{}_{|t=0}, \end{cases}$$

we have two different ways of analyzing v^ε:

1. Plane phases:

$$\begin{aligned} \partial_t v^\varepsilon{}_{|t=0} &= i\left(\sqrt{1+\varepsilon(Y/t_\star)^2}\, u_{0|t=0}\right)(y_2/\sqrt{\varepsilon}, y_1/\varepsilon) \\ &= i\left(u_{0|t=0} + \mathcal{O}_{\mathcal{S}(\mathbb{R}\times\mathbb{T})}(\varepsilon)\right)(y_2/\sqrt{\varepsilon}, y_1/\varepsilon), \end{aligned}$$

and from [5], for each $\underline{t} < t_\star$, when ε is small enough,

$$u^\varepsilon = \varepsilon \mathcal{U}^\varepsilon\left(t, \frac{y_2}{\sqrt{\varepsilon}}, \frac{y_1+t}{\varepsilon}\right), \quad v^\varepsilon = \varepsilon \mathcal{V}^\varepsilon\left(t, \frac{y_2}{\sqrt{\varepsilon}}, \frac{y_1+t}{\varepsilon}\right), \quad \text{and } \|\mathcal{U}^\varepsilon - \mathcal{V}^\varepsilon\|_{\cap H^s} \underset{\varepsilon\to 0}{\longrightarrow} 0,$$

so that $\|\partial_t u^\varepsilon - \partial_t v^\varepsilon\|_{L^\infty} \underset{\varepsilon\to 0}{\longrightarrow} 0$ uniformly on $[0, \underline{t}]$.

2. Curved phases: since

$$\begin{cases} v^\varepsilon{}_{|t=0} = \varepsilon \tilde{u}_0^0\left(y_2/\sqrt{\varepsilon}, \phi_0/\varepsilon\right), \\ \partial_t v^\varepsilon{}_{|t=0} = i|\partial_y \phi_0| \tilde{u}_0^0\left(y_2/\sqrt{\varepsilon}, \phi_0/\varepsilon\right) \end{cases}$$

with $\phi_0 = y_1 + y_2^2/2t_\star$, on each time interval $[0, \underline{t}]$, [7] ensures the representation $v^\varepsilon = \varepsilon \tilde{\mathcal{V}}^\varepsilon(t, \frac{y_2}{\sqrt{\varepsilon}}, \frac{\phi}{\varepsilon})$, where ϕ is characteristic for the d'Alembertian operator:

$$\partial_t \phi = |\partial_y \phi| \quad \text{and} \quad \phi_{|t=0} = y_1 + y_2^2/2t_\star.$$

This phase is implicitly determined by the "ray method": $\phi(t,y) = \phi_0(z)$, where z is the origin (at $t=0$) of the ray through (t,y) (which here is a straight line):

$$y - z + t\frac{\nabla\phi_0(z)}{|\nabla\phi_0(z)|} = 0.$$

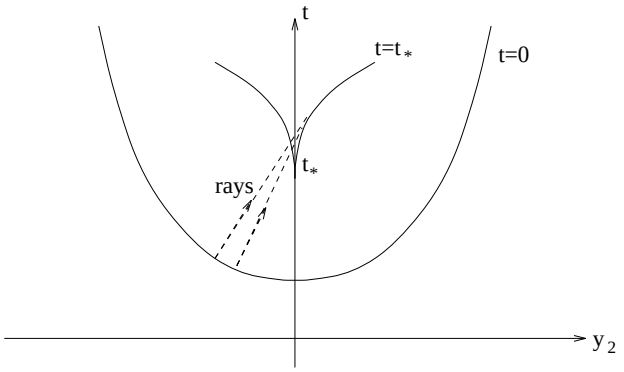


FIG. 4.1. *The graph of* $\phi(t, y_1, .)$ *at* $t = 0$ *and* $t = t_\star$.

Direct computations show that the rays focus exactly at time $t = t_\star$ (generating a caustic, where ϕ is no longer smooth). This is illustrated by Figure 4.1.

Remark 4.1. Here, ϕ can be determined explicitly (before $t = t_\star$) by solving analytically a fourth degree algebraic equation. However, the region of interest is the set of the "first" rays to focus, corresponding to $y_2 = 0$. On these rays, one easily computes that the gradient (w.r.t. t, y) of ϕ is the same as that of the linear phase $y_1 + t$. This indicates that planar and nonplanar phase representations correspond to similar oscillations and give two different ways of understanding the amplitude's behavior.

5. Existence and approximation of u^ε. We prove a slightly stronger approximation than in Theorem 1.1.

THEOREM 5.1. *When* $C > 0$ *is sufficiently large, and for all* $\alpha > 0$, *as* $\varepsilon \to 0$,

$$\left\|\partial_t u^\varepsilon - \partial_t u^\varepsilon_{app}\right\| = o(\varepsilon^{2-\alpha}) \text{ in } L^\infty((0, t_\star - C(\ln 1/\varepsilon)^{-1/3}) \times \mathbb{R}^2).$$

5.1. Rescaling. So as to evaluate the lifespan of u^ε, we look for the times during which $\|\partial_t u^\varepsilon\|_{L^\infty}$ is finite. With the idea of giving an approximation of u^ε, we define $v^\varepsilon := u^\varepsilon - u^\varepsilon_{app}$ and try to verify $\|\partial_t v^\varepsilon\|_{L^\infty} \ll \left\|\partial_t u^\varepsilon_{app}\right\|_{L^\infty}$.

We make use of the wave equation satisfied by v^ε. It provides us with energy estimates for $\partial_{t,y} v^\varepsilon$, which bound $\|\partial_t v^\varepsilon\|_{L^\infty}$, thanks to the Sobolev inequality. However, these direct computations are too crude, because of Remark 3.1(ii): u^ε and u^ε_{app} are $2\pi\varepsilon$-periodic in y_1, and so is v^ε. Thus, estimating $\|\partial_t v^\varepsilon\|_{L^\infty}$ by $\|\partial_t v^\varepsilon\|_{H^s}$, we lose a factor ε^s. That's why we finally try to control $\|\partial_t V^\varepsilon\|_{L^\infty}$, where

$$V^\varepsilon(t, y) := v^\varepsilon(t, \varepsilon y_1, \sqrt{\varepsilon} y_2). \tag{14}$$

In the same way, set $(U^\varepsilon, U^\varepsilon_{app}, R^\varepsilon)(x) := (u^\varepsilon, u^e_{app} p, r^\varepsilon)(t, \varepsilon y_1, \sqrt{\varepsilon} y_2)$ (where $r^\varepsilon(x) = \mathcal{R}^\varepsilon(t, y_2/\sqrt{\varepsilon}) e^{i\frac{y_1+t}{\varepsilon}}$ from section 3).

Notation 5.1. We write $a \preceq b$ when there is a constant C such that $a \leq Cb$.

5.2. Energy estimates for the error $\partial_t V^\varepsilon$. From the relation (14), subtracting (4) and (10) and using Taylor's formula for $F(z) = |z|^4 z$ (as a differentiable function on $\mathbb{R}^2$), we get

$$(\partial_t^2 - \varepsilon^{-2}\partial_{y_1}^2 - \varepsilon^{-1}\partial_{y_2}^2) V^\varepsilon = -i \left(\int_0^1 dF \left(\partial_t U^\varepsilon_{app} + r \partial_t V^\varepsilon \right) dr \right) . \partial_t V^\varepsilon - R^\varepsilon. \tag{15}$$

Consider $\int_{(-\pi\times\pi)\times\mathbb{R}} 2\mathrm{Re}((15)\times\partial_t\overline{V^\varepsilon})dy$. This gives

$$\frac{d}{dt}\Big(\|\partial_t V^\varepsilon\|_{L^2}^2+\varepsilon^{-2}\|\partial_{y_1}V^\varepsilon\|_{L^2}^2+\varepsilon^{-1}\|\partial_{y_2}V^\varepsilon\|_{L^2}^2\Big)$$
$$\preceq\Big(\big\|\partial_t U_{app}^\varepsilon\big\|_{L^\infty}^4+\|\partial_t V^\varepsilon\|_{L^\infty}^4\Big)\|\partial_t V^\varepsilon\|_{L^2}^2+\|R^\varepsilon\|_{L^2}\|\partial_t V^\varepsilon\|_{L^2},$$

so that, writing $N(V^\varepsilon):=(\|\partial_t V^\varepsilon\|_{L^2}^2+\varepsilon^{-2}\|\partial_{y_1}V^\varepsilon\|_{L^2}^2+\varepsilon^{-1}\|\partial_{y_2}V^\varepsilon\|_{L^2}^2)^{1/2}$ and $I_k:=\big\|\partial_t U_{app}^\varepsilon\big\|_{L^\infty}^k+\|\partial_t V^\varepsilon\|_{L^\infty}^k$,

$$\frac{d}{dt}N(V^\varepsilon)\preceq I_4N(V^\varepsilon)+\|R^\varepsilon\|_{L^2}. \tag{16}$$

Differentiating (15), in the same manner we get

$$\frac{d}{dt}N(\partial_yV^\varepsilon)\preceq I_4N(\partial_yV^\varepsilon)+I_3\big\|\partial_t\partial_yU_{app}^\varepsilon\big\|_{L^2}\|\partial_tV^\varepsilon\|_{L^\infty}+\|\partial_yR^\varepsilon\|_{L^2}, \tag{17}$$

and for any second order derivative ∂_y^2,

$$\frac{d}{dt}N(\partial_y^2V^\varepsilon)\preceq I_3\Big(\big\|\partial_t\partial_y^2U_{app}^\varepsilon\big\|_{L^2}\|\partial_tV^\varepsilon\|_{L^\infty}+\big\|\partial_t\partial_yU_{app}^\varepsilon\big\|_{L^2}\|\partial_t\partial_yV^\varepsilon\|_{L^\infty}\Big)$$
$$+\big\|\partial_y^2R^\varepsilon\big\|_{L^2}+I_4N(\partial_y^2V^\varepsilon). \tag{18}$$

Adding (16)–(18), using Sobolev's inequality and Gronwall's lemma, since $V^\varepsilon_{|_{t=0}}=\partial_tV^\varepsilon_{|_{t=0}}=0$, we have

$$\|\partial_tV^\varepsilon\|_{L^\infty}\preceq N(V^\varepsilon,\partial_yV^\varepsilon,\partial_y^2V^\varepsilon)\preceq e^{CJ(t)}\int_0^t\|R^\varepsilon\|_{H^2}dt', \tag{19}$$

where $J(t)=\int_0^t[I_4+I_3(\|\partial_t\partial_yU_{app}^\varepsilon\|+\|\partial_t\partial_y^2U_{app}^\varepsilon\|+\|\partial_t\partial_yU_{app}^\varepsilon\|)]dt'$.

5.3. Defining the boundary layer for the corrector. We first define an interval $[0,\underline{t}(\varepsilon)]$ where $U_{app}^\varepsilon\simeq U_0^\varepsilon:=\varepsilon u_0(t,y_2,y_1+\frac{t}{\varepsilon})$, i.e., where $U_c^\varepsilon:=\varepsilon^2u_c(t,y_2,y_1+\frac{t}{\varepsilon})$ is a corrector to this quantity.

PROPOSITION 5.2. *When $t_\star-t>C(\ln 1/\varepsilon)^{-1/3}$ for some C ($\gg1$),*

$$\|\partial_tU_c^\varepsilon\|_{W^{1,\infty}}\ll\|\partial_tU_{app}^\varepsilon\|_{W^{1,\infty}}\quad\text{and}\quad\|\partial_tU_c^\varepsilon\|_{H^2}\ll\|\partial_tU_{app}^\varepsilon\|_{H^2}.$$

Proof. We use the equation $\mathcal{E}_1=0$ from section 3 to obtain energy estimates. Since $\partial_t[a_c(t,y_2)e^{i(y_1+t/\varepsilon)}]=(\partial_ta_c+\frac{i}{\varepsilon}a_c)e^{i(y_1+t/\varepsilon)}$, we have to estimate $\|a_c\|_{H^s_Y}$ and $\|\partial_ta_c\|_{H^s_Y}$, $s=1,2$:

$$\frac{d}{dt}\|a_c\|_{H^1}\preceq\|\partial_t^2a_0\|_{H^1}+\||a_0|^4\partial_ta_0\|_{H^1}$$
$$+\Big(\|a_0\|_{L^\infty}^4+\|a_0\|_{L^\infty}^3\|\partial_Ya_0\|_{L^\infty}\Big)\|a_c\|_{H^1}, \tag{20}$$

$$\frac{d}{dt}\|\partial_ta_c\|_{H^1}\preceq\|\partial_t^3a_0\|_{H^1}+\||a_0|^3|\partial_ta_0|^2\|_{H^1}+\||a_0|^4\partial_t^2a_0\|_{H^1}$$
$$+\Big(\|a_0\|_{L^\infty}^3\|\partial_ta_0\|_{W^{1,\infty}}+\|a_0\|_{L^\infty}^2\|\partial_ta_0\|_{L^\infty}\|\partial_Ya_0\|_{L^\infty}\Big)\|a_c\|_{H^1}$$
$$+\Big(\|a_0\|_{L^\infty}^4+\|a_0\|_{L^\infty}^3\|\partial_Ya_0\|_{L^\infty}\Big)\|\partial_ta_c\|_{H^1}, \tag{21}$$

$$\frac{d}{dt}\|a_c\|_{H^2} \preceq \|\partial_t^2 a_0\|_{H^2} + \||a_0|^4\partial_t a_0\|_{H^2} + \Big(\|a_0\|_{L^\infty}^4 + \|a_0\|_{L^\infty}^3\|\partial_Y a_0\|_{L^\infty} + \|a_0\|_{L^\infty}^2\|\partial_Y a_0\|_{L^\infty}^2 + \|a_0\|_{L^\infty}^3\big\|\partial_Y^2 a_0\big\|_{L^\infty}\Big)\|a_c\|_{H^2}, \tag{22}$$

(23)
$$\begin{aligned}\frac{d}{dt}\|\partial_t a_c\|_{H^2} \preceq{}& \|\partial_t^3 a_0\|_{H^2} + \||a_0|^3|\partial_t a_0|^2\|_{H^2} + \||a_0|^4\partial_t^2 a_0\|_{H^2}\\ &+ \Big(\|a_0\|_{L^\infty}^3\|\partial_t a_0\|_{W^{1,\infty}} + \|a_0\|_{L^\infty}^2\|\partial_t a_0\|_{L^\infty}\|\partial_Y a_0\|_{L^\infty}\\ &\quad + \|a_0\|_{L^\infty}\|\partial_t a_0\|_{W^{1,\infty}}\|\partial_Y a_0\|_{L^\infty}^2 + \|a_0\|_{L^\infty}^2\|\partial_t a_0\|_{L^\infty}\big\|\partial_Y^2 a_0\big\|_{L^\infty}\\ &\quad + \|a_0\|_{L^\infty}^3\big\|\partial_t\partial_Y^2 a_0\big\|_{L^\infty}\Big)\|a_c\|_{H^2}\\ &+ \Big(\|a_0\|_{L^\infty}^4 + \|a_0\|_{L^\infty}^3\|\partial_Y a_0\|_{L^\infty} + \|a_0\|_{L^\infty}^2\|\partial_Y a_0\|_{L^\infty}^2\\ &\quad + \|a_0\|_{L^\infty}^3\big\|\partial_Y^2 a_0\big\|_{L^\infty}\Big)\|\partial_t a_c\|_{H^2}.\end{aligned}$$

Note that we can compute the exact value of the norm of a_0 from the formula (5):

$$\forall\alpha,\ \ \|(\partial_{t,Y})^\alpha a_0\|_{L^\infty} = C(t_\star - t)^{-1/2-|\alpha|},\ \ \|(\partial_{t,Y})^\alpha a_0\|_{L^2} = C(t_\star - t)^{-|\alpha|}. \tag{24}$$

Thus, from (20)–(24), Gronwall's lemma implies that there are $C > 0$, $\mu \in \mathbb{R}$ such that

$$\|a_c, \partial_t a_c\|_{H^2_Y} \preceq (t_\star - t)^\mu e^{C(t_\star - t)^{-3}}. \tag{25}$$

Now, simply check that out of the boundary layer $t_\star - t \le C(\ln 1/\varepsilon)^{-1/3}$ (with C big enough), we have

$$\begin{aligned}&\|\partial_t U_c^\varepsilon\|_{L^\infty} \preceq \varepsilon^2\|\partial_t a_c\|_{H^1} + \varepsilon\|a_c\|_{H^1} \ll \|\partial_t U_0^\varepsilon\|_{L^\infty} \sim (t_\star - t)^{-1/2},\\ &\|\partial_t\partial_y U_c^\varepsilon\|_{L^\infty} \preceq \varepsilon^2\|\partial_t a_c\|_{H^2} + \varepsilon\|a_c\|_{H^2} \ll \|\partial_t\partial_y U_0^\varepsilon\|_{L^\infty} \sim (t_\star - t)^{-3/2},\\ &\|\partial_t\partial_y U_c^\varepsilon\|_{L^2} \preceq \varepsilon^2\|\partial_t a_c\|_{H^1} + \varepsilon\|a_c\|_{H^1} \ll \|\partial_t\partial_y U_0^\varepsilon\|_{L^2} \sim (t_\star - t)^{-1}.\end{aligned}$$ □

5.4. Endgame: Proof of Theorem 5.1. We now take advantage of (19): in view of Proposition 5.2, when $t_\star - t \le C(\ln 1/\varepsilon)^{-1/3}$,

$$\|\partial_t V^\varepsilon\|_{L^\infty} \preceq e^{C\tilde J(t)}\int_0^t \|R^\varepsilon\|_{H^2}dt', \tag{26}$$

where $R^\varepsilon(x) = \mathcal{R}^\varepsilon(t, y_2)e^{i(y_1 + t/\varepsilon)}$ is given by section 3, and

$$\begin{aligned}\tilde J(t) &= \int_0^t \Big[\tilde I_4 + \tilde I_3\left(\|\partial_t\partial_y U_0^\varepsilon\|_{L^2} + \|\partial_t\partial_y U_0^\varepsilon\|_{L^\infty}\right)\Big]dt',\\ \tilde I_k &= \big\|\partial_t U_{app}^\varepsilon\big\|_{L^\infty}^k + \|\partial_t V^\varepsilon\|_{L^\infty}^k.\end{aligned}$$

Hence, (26) is a relation of the form $\varphi \le \psi e^{\varphi^4}$ for $\varphi(t) := C\|\partial_t V^\varepsilon\|_{L^\infty((0,t)\times\mathbb{R}^2)}$. Even if ψ is "small," this does not imply that φ is: on the contrary, it could be very large. But since here $\varphi_{|t=0} = 0$, continuity w.r.t. t forces that *as long as* (26) is

valid, φ has to be "small." Thus, for each $\varepsilon \in]0,1]$, we look for the maximal time $\underline{t}(\varepsilon) \in]0, t_\star - C(\ln 1/\varepsilon)^{-1/3}]$, until which

$$\|\partial_t V^\varepsilon\|_{L^\infty} \leq \|\partial_t U_0^\varepsilon\|_{L^\infty} \sim (t_\star - t)^{-1/2}, \tag{27}$$

and we replace $\tilde{I}_k$ by $\hat{I}_k := 2\,\|\partial_t U_0^\varepsilon\|_{L^\infty}^k$.

Since (from (24)) $\hat{I}_k(t) \sim (t_\star - t)^{-2}$,

$$\|\partial_t V^\varepsilon\|_{L^\infty} \leq \varepsilon^2 (t_\star - t)^\mu e^{C(t_\star - t)^{-2}}, \tag{28}$$

and the right-hand side is much smaller than $(t_\star - t)^{-1/2}$ as soon as $t_\star - t > C'(\ln 1/\varepsilon)^{-1/2}$, with C' sufficiently large. As ε goes to zero, $(\ln 1/\varepsilon)^{-1/2} \ll (\ln 1/\varepsilon)^{-1/3}$, so that the condition $t_\star - t > C(\ln 1/\varepsilon)^{-1/3}$ is the relevant one.

Furthermore, for each $\alpha > 0$, possibly increasing C, (28) shows that (27) is improved to $\|\partial_t V^\varepsilon\|_{L^\infty((0,t_\star - C(\ln 1/\varepsilon)^{-1/3})\times\mathbb{R}^2)} = o(\varepsilon^{2-\alpha})$.

REFERENCES

[1] H. Berestycki and P.-L. Lions, *Existence d'ondes solitaires dans des problèmes non-linéaires du type Klein-Gordon*, C. R. Acad. Sci. Paris Sér. A, 287 (1978), pp. 503–506.

[2] N. Bloembergen, *Nonlinear Optics*, W.A. Benjamin, New York, 1965.

[3] R. Boyd, *Nonlinear Optics*, Academic Press, New York, 1992.

[4] A. de Bouard, A. Debussche, and L. Di Menza, *Some theoretical and numerical results on nonlinear Schrödinger equations perturbed by noise*, Journées Équations aux Dérivées Partielles, Plestin-les-grèves, France, 2001.

[5] P. Donnat, J.L. Joly, G. Métivier, and J. Rauch, *Diffractive nonlinear geometric optics*, in Séminaire sur les Équations aux Dérivées Partielles, 1995–1996, Exp. 17, Sémin. Équ. Dériv. Partielles, Ecole Polytechnique, Palaiseau, France, 1996.

[6] P. Donnat and J. Rauch, *Global solvability of the Maxwell-Bloch equations from nonlinear optics*, Arch. Ration. Mech. Anal., 136 (1996), pp. 291–303.

[7] E. Dumas, *Nonlinear diffractive optics with curved phases: Beam dispersion and transitions between light and shadow*, to appear.

[8] E. Dumas, *Existence globale pour les systèmes de Maxwell-Bloch*, in Séminaire École Polytechnique, 2002–2003, Ecole Polytechnique, Palaiseau, France.

[9] G. Fibich and G. Papanicolaou, *Self-focusing in the perturbed and unperturbed nonlinear Schrödinger equation in critical dimension*, SIAM J. Appl. Math., 60 (1999), pp. 183–240.

[10] J. Ginibre and G. Velo, *On a class of nonlinear Schrödinger equations.* I: *The Cauchy problem*, J. Funct. Anal., 32 (1979), pp. 1–32.

[11] H. Haddar, *Modèles asymptotiques en ferromagnétisme: Couches minces et homogénéisation*, Ph.D. thesis, thèse INRIA-École Nationale des Ponts et Chaussées, 2000.

[12] J.L. Joly, G. Métivier, and J. Rauch, *Global solvability of the anharmonic oscillator model from nonlinear optics*, SIAM J. Math. Anal., 27 (1996), pp. 905–913.

[13] J.L. Joly, G. Métivier, and J. Rauch, *Global solutions to Maxwell equations in a ferromagnetic medium*, Ann. Henri Poincaré, 1 (2000), pp. 307–340.

[14] L. Landau and E. Lifshitz, *Électrodynamique des milieux continus, cours de physique théorique*, t. 8., éditions Mir, Moscou, 1969.

[15] D. Lannes and J. Rauch, *Validity of nonlinear geometric optics with times growing logarithmically*, Proc. Amer. Math. Soc., 129 (2001), pp. 1087–1096.

[16] A.C. Newell, *Solitons in Mathematics and Physics*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 48, SIAM, Philadelphia, 1985.

[17] A.C. Newell and J.V. Moloney, *Nonlinear Optics*, Addison-Wesley, Redwood City, CA, 1992.

[18] R. Pantell and H. Puthoff, *Fundamentals of Quantum Electronics*, John Wiley, New York, 1969.

[19] G. Schneider, *The long wave limit for a Boussinesq equation*, SIAM J. Appl. Math., 58 (1998), pp. 1237–1245.

[20] G. Schneider and C.E. Wayne, *The long-wave limit for the water wave problem.* I. *The case of zero surface tension*, Comm. Pure Appl. Math., 53 (2000), pp. 1475–1535.
[21] G. Schneider and C.E. Wayne, *The rigorous approximation of long-wavelength capillary-gravity waves*, Arch. Ration. Mech. Anal., 162 (2002), pp. 247–285.
[22] W. Strauss, *Nonlinear Wave Equations*, AMS, Providence, RI, 1989.
[23] C. Sulem and P.L. Sulem, *The Nonlinear Schrödinger Equation*, Appl. Math. Sci. 139, Springer-Verlag, New York, 1999.

SIAM J. MATH. ANAL.
Vol. 35, No. 2, pp. 279–303

SPARSE EVALUATION OF COMPOSITIONS OF FUNCTIONS USING MULTISCALE EXPANSIONS*

ALBERT COHEN†, WOLFGANG DAHMEN‡, AND RONALD DEVORE§

Abstract. This paper is concerned with the estimation and evaluation of wavelet coefficients of the *composition* $\mathcal{F}\circ u$ of two functions $\mathcal{F}$ and u from the wavelet coefficients of u. Our main objective is to show that certain sequence spaces that can be used to measure the sparsity of the arrays of wavelet coefficients are stable under a class of nonlinear mappings $\mathcal{F}$ that occur naturally, e.g., in nonlinear PDEs. We indicate how these results can be used to facilitate the sparse evaluation of arrays of wavelet coefficients of compositions at asymptotically optimal computational cost. Furthermore, the basic requirements are verified for several concrete choices of nonlinear mappings. These results are generalized to compositions by a multivariate map $\mathcal{F}$ of several functions $u_1,\dots,u_n$ and their derivatives, i.e., $\mathcal{F}(D^{\alpha_1}u_1,\dots,D^{\alpha_n}u_n)$.

Key words. nonlinear mappings, thresholding, tree structures, adaptive evaluation of nonlinear operators

AMS subject classifications. 35A15, 35A22, 41A60, 46B15, 46A45, 47H30

DOI. 10.1137/S0036141002412070

1. Introduction.

This paper is concerned with the estimation and evaluation of the wavelet coefficients of a *composition* of two functions $\mathcal{F}$ and u, where u is given in terms of a wavelet expansion. Our interest in this subject stems from recent developments of *adaptive wavelet schemes* for the numerical solution of several types of initial or boundary value problems for PDEs. Such schemes typically rely on the sparsity of the wavelet representation of the solution allowing for data compression, as well as the ability to perform accurate numerical computations in the compressed representation. For initial value problems, *dynamically adaptive* schemes introduced in [20] require a reliable prediction of significant wavelet coefficients from the current state when progressing to the next time level. In the case of hyperbolic conservation laws, this question was first addressed in [19] and further discussed in [12]. Here one has to estimate the action of the nonlinear terms defining the convective fluxes on the current approximation in its multiscale representation. Another related example is the wavelet analysis of turbulent incompressible flows where such estimates are related to the energy transfer between different scales; see, e.g., [18] and [17]. For boundary value problems, adaptive wavelet schemes also require the tracking of the significant coefficients as the iterative solution process progresses; see, e.g., [1], [5], [9], and [10].

In all these examples, we are interested in the following general question: does composition with $\mathcal{F}$ preserve the *sparsity* of the wavelet coefficients of the function u? By the sparsity, we mean that only a quantifiable relatively small set of these

*Received by the editors July 26, 2002; accepted for publication (in revised form) November 25, 2002; published electronically July 18, 2003. This work was supported in part by Office of Naval Research contract N0014-91-J1343, Army Research Office contract DAAD 19-02-1-0028, the TMR network "Wavelets in Numerical Simulation," the SFB 401, funded by the German Research Foundation and by the Alexander von Humboldt Foundation.

http://www.siam.org/journals/sima/35-2/41207.html

†Laboratoire d'Analyse Numerique, Universite Pierre et Marie Curie, 175 Rue du Chevaleret, 75013 Paris, France (cohen@ann.jussieu.fr, http://www.ann.jussieu.fr/∼cohen/).

‡Institut für Geometrie und Praktische Mathematik, RWTH Aachen, 52056 Aachen, Germany (dahmen@igpm.rwth-aachen.de, http://www.igpm.rwth-aachen.de/∼dahmen/).

§Department of Mathematics, University of South Carolina, Columbia, SC 29208 (devore@math.sc.edu, http://www.math.sc.edu/∼devore/).

coefficients is needed to recover the underlying function (with accuracy measured in a given norm) to within some target accuracy. It is well known that sparsity of wavelet coefficients in this sense is closely related (in fact equivalent) to the *regularity* of the function with respect to certain scales of Besov spaces; see, e.g., [16]. Hence the above issue is closely connected with the question, how is the regularity of a given function u affected by the composition with some nonlinear function $\mathcal{F}$, or, more generally, given some regularity spaces $\mathcal{R}_i$, $i = 1, \ldots, m$, what is the image of $\prod_{i=1}^m \mathcal{R}_i$ under the mapping

$$(u_1(\cdot), \ldots, u_m(\cdot)) \to \mathcal{F}(\cdot, u_1(\cdot), \ldots, u_m(\cdot))?$$

This mapping is often referred to as a Nemytskij operator. The mapping properties of Nemytskij operators between Besov spaces were treated by several authors, and the reader is referred, e.g., to [3], [4], [22], and, for a detailed treatment, to the book by Runst and Sickel [21]. Sharp results are indeed available on the amount of smoothness which can be expected for $\mathcal{F}(u)$ given the smoothness of u, under fairly general assumptions on $\mathcal{F}$. Thus, in principle, in all cases covered by these results the sparsity of the wavelet coefficients of compositions can be predicted fairly well. However, these results tell us neither *which* coefficients of compositions $\mathcal{F}(u)$ are significant, based on knowledge about u, nor how to calculate them efficiently once they have been identified, which is a crucial issue in the perspective of numerical computations. The objective of the present paper is therefore also to develop concepts and tools for treating this latter problem.

Our paper is organized as follows. We present the problem formulation in section 2, which involves the wavelet discretization $\mathbf{F}$ of the mapping $\mathcal{F}$ as well as a notion of *tree structure* in the organization of wavelet coefficients. We prove in section 3 that this mapping preserves sparsity, under some general assumptions describing the stability and local action of $\mathbf{F}$ in the space-scale domain. We also present specific algorithms that construct sparse approximants with a prescribed accuracy ε at asymptotically optimal cost. This type of scheme is needed for the adaptive solution process of nonlinear operator equations; see [11]. We shall prove in section 4 the validity of the required assumptions for general local nonlinear mappings of subcritical type. Finally, the generalization of these results to compositions of the form $\mathcal{F}(D^{\alpha_1}u_1, \ldots, D^{\alpha_n}u_n)$ between a multivariate map $\mathcal{F}$ and the derivatives of several functions $u_1, \ldots, u_n$ is discussed in section 5.

2. Problem formulation.

2.1. Background and wavelet prerequisites. To explain the relevant features of the problem it suffices to describe the following (simple) example in a little more detail. Consider the nonlinear boundary value problem of the form

$$-\Delta u + \mathcal{F}(u) = f \ \text{ in } \ \Omega, \ \ u = 0 \ \text{ on } \ \partial\Omega, \tag{2.1}$$

where $\Omega \subset \mathbb{R}^d$ is some open bounded domain. The variational formulation of (2.1) in the space $H = H_0^1(\Omega)$ reads as follows: find $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v + \int_\Omega \mathcal{F}(u)v = \int_\Omega fv \tag{2.2}$$

for all $v \in H_0^1(\Omega)$. Here $H_0^1(\Omega)$ is the usual Sobolev space of distributions with first order weak derivatives in $L_2(\Omega)$ vanishing on the boundary $\partial\Omega$ in the sense of

traces. (Of course, other boundary conditions may also be considered.) For (2.2) to be meaningful $\mathcal{F}$ should map $H_0^1(\Omega)$ into its dual $H^{-1}(\Omega)$. This is perhaps the simplest instance of a variational problem inducing a bijective mapping from a Hilbert space H onto its dual H'.

For more general problems, H is a product of closed subspaces H^t of Sobolev spaces determined, e.g., by homogeneous boundary conditions on part of the domain boundary; see, e.g., [10] for examples. For simplicity we will confine the subsequent discussion to the case of a single model space $H = H^t$ for some $t > 0$.

2.2. Wavelet discretization. As already explained, we are motivated by adaptive numerical methods based on discretizing the variational formulation (2.2) in a wavelet basis $\Psi = \{\psi_\lambda : \lambda \in \mathcal{J}\}$. The indices λ encode scale, spatial location, and the type of the wavelet ψ_λ. We will denote by $|\lambda|$ the *scale* associated with ψ_λ. We shall consider only compactly supported wavelets, i.e., the supports of the wavelets scale, as follows:

$$S_\lambda := \operatorname{supp} \psi_\lambda, \quad c_0 2^{-|\lambda|} \le \operatorname{diam} S_\lambda \le C_0 2^{-|\lambda|}, \tag{2.3}$$

with $c_0, C_0 > 0$ absolute constants. The index set $\mathcal{J}$ has the following structure $\mathcal{J} = \mathcal{J}_\phi \cup \mathcal{J}_\psi$, where $\mathcal{J}_\phi$ is finite and indexes the scaling functions on a fixed coarsest level j_0. $\mathcal{J}_\psi$ indexes the "true wavelets" ψ_λ with $|\lambda| > j_0$. From compactness of the supports we know that at each level, the set $\mathcal{J}_j := \{\lambda \in \mathcal{J} : |\lambda| = j\}$ is finite. In fact, one has $\#\mathcal{J}_j \sim 2^{jd}$ with constants depending on the underlying bounded domain.

As already explained in the introduction, our evaluation algorithms will rely on a tree structure associated to the set of wavelet indices. In the simplest case of a one-dimensional basis $\psi_\lambda = \psi_{j,k} = 2^{j/2}\psi(2^j \cdot -k)$, this structure is obvious: each index (j,k) has two children $(j+1, 2k)$ and $(j+1, 2k+1)$. A similar tree structure can be associated to all available constructions of wavelet bases on a multidimensional domain: each index λ then has $m(\lambda) \ge 2$ children μ such that $|\mu| = |\lambda| + 1$, where $m(\lambda)$ might vary from one index to another but is uniformly bounded by some fixed K. We shall use the notation $\mu \prec \lambda$ in order to express that μ is a descendent of λ in the tree. Moreover, $\mu \preceq \lambda$ means that μ either is a descendent of λ or equals λ. We also have the property

$$\mu \prec \lambda \Rightarrow S_\mu \subset S_\lambda. \tag{2.4}$$

One key feature is that Ψ is a *Riesz basis* of the relevant space $H = H^t$. This means that every $v \in H$ has a unique expansion $v = \sum v_\lambda \psi_\lambda$ and that there exist some constants c, C independent of v such that

$$c\|(v_\lambda)_{\lambda\in\mathcal{J}}\| \le \left\| \sum_{\lambda\in\mathcal{J}} v_\lambda \psi_\lambda \right\|_H \le C\|(v_\lambda)_{\lambda\in\mathcal{J}}\|, \tag{2.5}$$

where $\|(v_\lambda)_{\lambda\in\mathcal{J}}\|^2 = \sum_{\lambda\in\mathcal{J}} |v_\lambda|^2$ denotes the $\ell_2(\mathcal{J})$-norm. In particular, the wavelets will always be assumed to be normalized in H, i.e., $\|\psi_\lambda\|_H = 1$. We abbreviate by

$$\mathbf{v} = (v_\lambda)_{\lambda\in\mathcal{J}}$$

the corresponding sequence of wavelet coefficients. Details on the construction of wavelet bases for Sobolev spaces of general domains can be found in [6], [7], [14].

Note that, by duality, (2.5) is equivalent to

$$C^{-1}\|(\langle w, \psi_\lambda\rangle)_{\lambda\in\mathcal{J}}\| \le \|w\|_{H'} \le c^{-1}\|(\langle w, \psi_\lambda\rangle)_{\lambda\in\mathcal{J}}\|, \quad w \in H', \tag{2.6}$$

where $\langle\cdot,\cdot\rangle$ denotes the duality pairing between H and H'. Clearly the quantities $\langle w,\psi_\lambda\rangle$ are the coordinates of $w\in H'$ with respect to the *dual* Riesz basis $\tilde\Psi$ to Ψ.

Since, as pointed out above, the nonlinearity $\mathcal{F}$ is supposed to map H into H' we shall therefore describe $w=\mathcal{F}(u)$ by its inner product sequence $\mathbf{w}=(w_\lambda)_{\lambda\in\mathcal{J}}$ with

$$w_\lambda=\langle w,\psi_\lambda\rangle,\quad \lambda\in\mathcal{J}. \tag{2.7}$$

We shall denote by $\mathbf{F}$ the corresponding *discrete* nonlinear map

$$\mathbf{u}\mapsto\mathbf{w}=\mathbf{F}(\mathbf{u})=(\langle\mathcal{F}(u),\psi_\lambda\rangle)_{\lambda\in\mathcal{J}}. \tag{2.8}$$

A key issue in the applications mentioned above can roughly be described as follows. Suppose that $u\in H$ can be approximated in the energy norm $\|\cdot\|_H$ within a tolerance ε by a linear combination of $N(\varepsilon,u)$ wavelets ψ_λ. What is the number $N(\varepsilon,\mathcal{F}(u))$ of dual wavelets needed to recover $\mathcal{F}(u)$ within tolerance ε? Note that, due to the *norm equivalences* (2.5) and (2.6), this can be restated as follows: Supposing that the wavelet coefficients $\mathbf{u}$ of $u\in H$ can be approximated in $\ell_2(\mathcal{J})$ with accuracy ε by a finitely supported vector involving only $N(\varepsilon,u)$ nonzero terms, how many entries of the sequence $\mathbf{F}(\mathbf{u})$ are needed to approximate $\mathbf{F}(\mathbf{u})$ in $\ell_2(\mathcal{J})$? Thus in the wavelet coordinate domain *all* approximations take place in $\ell_2(\mathcal{J})$. In brief, when does sparse approximability of $\mathbf{u}$ imply sparse approximability of $\mathbf{F}(\mathbf{u})$?

Questions of the above type are by now well understood for *linear* operators and their wavelet representations, as we shall now describe. In this context, the level of sparsity of $\mathbf{u}$ is measured by the *smallest* $\tau\le 2$ such that $\mathbf{u}\in\ell_\tau^w(\mathcal{J})$. Here $\ell_\tau^w(\mathcal{J})$ is the collection of all $\mathbf{u}\in\ell_2(\mathcal{J})$ which satisfy

$$\#\{\lambda\in\mathcal{J}:|u_\lambda|>\eta\}\le C\eta^{-\tau},\quad \eta>0. \tag{2.9}$$

In fact, $\ell_\tau^w(\mathcal{J})$ is a (quasi-)normed linear space endowed with the norm

$$\|\mathbf{u}\|_{\ell_\tau^w(\mathcal{J})}:=\sup_{\eta>0}\eta[\#\{\lambda\in\mathcal{J}:|u_\lambda|>\eta\}]^{1/\tau}. \tag{2.10}$$

An equivalent norm is given by the quantity

$$\sup_{n>0}n^{1/\tau}u_n^*, \tag{2.11}$$

where $(u_n^*)_{n>0}$ is a nonincreasing rearrangement of $(|u_\lambda|)_{\lambda\in\mathcal{J}}$. Note that if $\tau<2$, we have[1]

$$\|\mathbf{u}\|\ \lesssim\ \|\mathbf{u}\|_{\ell_\tau^w(\mathcal{J})}. \tag{2.12}$$

Moreover, defining the error of *best N-term approximation* in $\ell_2(\mathcal{J})$

$$\sigma_N(\mathbf{u}):=\inf_{\#\operatorname{supp}\mathbf{v}\le N}\|\mathbf{u}-\mathbf{v}\|=\left(\sum_{n>N}|u_n^*|^2\right)^{1/2}, \tag{2.13}$$

one has the following characterization [9].

[1]Here and later we use the notation $a\ \lesssim\ b$ if $a\le Cb$ with an absolute constant C independent of all parameters on which a,b depend.

PROPOSITION 2.1. *For* $\mathbf{u} \in \ell_2(\mathcal{J})$ *and* $s > 0$, *one has* $\sigma_N(\mathbf{u}) \lesssim N^{-s}$ *if and only if* $\mathbf{u} \in \ell^w_\tau(\mathcal{J})$ *with*

$$\frac{1}{\tau} = s + \frac{1}{2}. \tag{2.14}$$

Moreover,

$$\sigma_N(\mathbf{u}) \lesssim N^{-s}\|\mathbf{u}\|_{\ell^w_\tau(\mathcal{J})}. \tag{2.15}$$

Thus the smaller τ is, the fewer terms are needed to achieve a desired target accuracy for $\mathbf{u} \in \ell^w_\tau(\mathcal{J})$. In the case where $\mathbf{F}(\mathbf{u}) = \mathbf{A}\mathbf{u}$ is a linear operator bounded in $\ell_2(\mathcal{J})$, it is shown in [9] that this operator maps $\ell^w_\tau(\mathcal{J})$ into itself provided that it can be approximated by sparse matrices $\mathbf{A}_N$ with N entries per rows and columns at the rate $\|\mathbf{A} - \mathbf{A}_N\|_{\ell_2(\mathcal{J})} \lesssim N^{-r}$ for some $r > \frac{1}{\tau} - \frac{1}{2}$. Moreover, it is also shown how to practically build N-term approximations $\mathbf{w}_N$ of $\mathbf{w} = \mathbf{A}\mathbf{u}$, which fulfill the optimal rate $\|\mathbf{w}_N - \mathbf{w}\|_{\ell_2(\mathcal{J})} \lesssim N^{-s}$, from similar approximations of $\mathbf{u}$ at $\mathcal{O}(N)$ computational cost.

2.3. Tree structures and weak spaces. When dealing with nonlinear mappings, the following slight modification of these notions turns out to be appropriate. The approximants will be constrained by imposing a *tree structure* to the set of indices identifying the active coefficients. We shall say that a set $\mathcal{T} \subset \mathcal{J}$ is a *tree* if $\lambda \in \mathcal{T}$ implies $\mu \in \mathcal{T}$ whenever $\lambda \prec \mu$.

If the tree $\mathcal{T} \subset \mathcal{J}$ is finite, we define the set $\mathcal{L} = \mathcal{L}(\mathcal{T})$ of *outer leaves* as the set of those indices outside the tree such that their parent belongs to the tree

$$\mathcal{L} := \{\lambda \in \mathcal{J} : \lambda \notin \mathcal{T},\ \lambda \prec \mu \Longrightarrow\ \mu \in \mathcal{T}\}. \tag{2.16}$$

We shall make use of the following easily verifiable equivalence:

$$\#\mathcal{T} \sim \#\mathcal{L}, \tag{2.17}$$

where the constants depend only on K. Defining

$$\Gamma_\lambda := \{\mu \in \mathcal{J} : \mu \preceq \lambda\}, \tag{2.18}$$

the tree with root node λ, one easily verifies that

$$\mathcal{J} \setminus \mathcal{T} = \bigcup_{\lambda \in \mathcal{L}(\mathcal{T})} \Gamma_\lambda. \tag{2.19}$$

We are now interested in the approximation of $\mathbf{u}$ by an N-term approximation $\mathbf{v}$, where the support of $\mathbf{v}$ is assumed in addition to have a tree structure. A natural counterpart to classical best N-term approximation error, discussed in the previous section, is therefore given by redefining σ_N according to

$$\sigma_N(\mathbf{u}) := \inf\{\|u - v\| \ :\ \#(\mathrm{supp}(\mathbf{v})) \le N \ \text{ and } \ \mathrm{supp}(\mathbf{v}) \text{ is a tree}\}. \tag{2.20}$$

We define $\mathcal{A}^s$ as the class of vectors $\mathbf{u}$ such that

$$\sigma_N(\mathbf{u}) \lesssim N^{-s} \tag{2.21}$$

and the corresponding quasi norm

$$\|\mathbf{u}\|_{\mathcal{A}^s} := \sup_{N>0} N^s \sigma_N(\mathbf{u}). \tag{2.22}$$

In contrast to best N-term approximation, the practical determination of a best N-term tree approximant is not a simple task. In particular, when $\mathbf{u}$ is a finite vector, the main difficulty is to build such an approximation without searching through all possible subtrees, which would result in exponential complexity in N. In [2], two algorithms have been proposed which construct near best trees in linear time, based on the evaluation of the local residuals

$$\tilde{u}_\lambda := \left(\sum_{\mu \in \Gamma_\lambda} |u_\mu|^2 \right)^{1/2}. \tag{2.23}$$

Note that

$$\|\mathbf{u} - \mathbf{u}|_{\mathcal{T}}\|^2 = \sum_{\lambda \in \mathcal{L}(\mathcal{T})} \tilde{u}_\lambda^2. \tag{2.24}$$

More precisely, given a tolerance ε, the algorithms proposed in [2] allow us to build a tree $\mathcal{T} = \mathcal{T}(\varepsilon, \mathbf{u})$ such that

$$\|\mathbf{u} - \mathbf{u}|_{\mathcal{T}}\| \le \varepsilon \tag{2.25}$$

with the following property: whenever a tree $\tilde{\mathcal{T}}$ satisfies $\|\mathbf{u} - \mathbf{u}|_{\tilde{\mathcal{T}}}\| \le c\varepsilon$, then $\#(\mathcal{T}) \le C\#(\tilde{\mathcal{T}})$, where c, C are fixed constants independent of $\mathbf{u}$ and ε.

A simpler alternative to building tree approximants is to perform thresholding on the residual sequence $\tilde{u}_\lambda$. Indeed, one readily verifies that $\mu \preceq \lambda$ implies $\tilde{u}_\lambda \ge \tilde{u}_\mu$, i.e., for any $\eta > 0$ the set

$$\mathcal{T}_\eta = \mathcal{T}_\eta(\mathbf{u}) := \{\lambda \,:\, |\tilde{u}_\lambda| > \eta\} \tag{2.26}$$

has tree structure. Thus, thresholding with respect to the modified sequences $\tilde{\mathbf{u}}$ creates trees. This motivates us to define

$${}_t\ell^w_\tau(\mathcal{J}) := \{\mathbf{u} \in \ell_2(\mathcal{J}) : \tilde{\mathbf{u}} \in \ell^w_\tau(\mathcal{J})\}, \quad \|\mathbf{u}\|_{{}_t\ell^w_\tau(\mathcal{J})} := \|\tilde{\mathbf{u}}\|_{\ell^w_\tau(\mathcal{J})}. \tag{2.27}$$

Clearly, we have $\|\mathbf{u}\|_{\ell^w_\tau(\mathcal{J})} \le \|\mathbf{u}\|_{{}_t\ell^w_\tau(\mathcal{J})}$ and

$$\#\mathcal{T}_\eta(\mathbf{u}) \le \eta^{-\tau} \|\mathbf{u}\|^\tau_{{}_t\ell^w_\tau(\mathcal{J})}. \tag{2.28}$$

Therefore, the spaces ${}_t\ell^w_\tau(\mathcal{J})$ can also be used to quantify the sparseness of sequences subject to the tree structure constraint. In fact, one has the following counterpart to Proposition 2.1.

PROPOSITION 2.2. *Let* $\mathbf{u}_\eta := \mathbf{u}|_{\mathcal{T}_\eta}$. *Then* $\mathbf{u} \in {}_t\ell^w_\tau(\mathcal{J})$ *implies the error estimate*

$$\|\mathbf{u} - \mathbf{u}_\eta\| \lesssim \eta^{1-\tau/2} \|\mathbf{u}\|^{\tau/2}_{{}_t\ell^w_\tau(\mathcal{J})} \lesssim [\#(\mathcal{T}_\eta)]^{-s} \|\mathbf{u}\|_{{}_t\ell^w_\tau(\mathcal{J})}, \tag{2.29}$$

with $s = 1/\tau - 1/2$. *Therefore* ${}_t\ell^w_\tau(\mathcal{J})$ *is contained in* $\mathcal{A}^s$.

Proof. Let $\mathcal{L}_\eta := \mathcal{L}(\mathcal{T}_\eta)$ denote the set of outer leaves of the tree $\mathcal{T}_\eta$. By (2.19), (2.24) and using (2.28), one has

$$\begin{aligned} \|\mathbf{u} - \mathbf{u}_\eta\|^2 &= \sum_{\lambda \notin \mathcal{T}_\eta} |u_\lambda|^2 = \sum_{\lambda \in \mathcal{L}_\eta} \tilde{u}_\lambda^2 \le \#\mathcal{L}_\eta \eta^2 \\ &\lesssim \#\mathcal{T}_\eta \eta^2 \le \|\mathbf{u}\|^\tau_{{}_t\ell^w_\tau(\mathcal{J})} \eta^{2-\tau}, \end{aligned} \tag{2.30}$$

where we have used (2.17). This confirms the first estimate in (2.29). Since again by definitions (2.27) and (2.28), $\eta \le \|\mathbf{u}\|_{{}_t\ell^w_\tau(\mathcal{J})}(\#\mathcal{T}_\eta)^{-1/\tau}$, the second estimate follows from (2.30). ∎

Note, however, that in the above proposition, we do not have a converse result which would state that the decay property $\|\mathbf{u}-\mathbf{u}_\eta\| \lesssim [\#(\mathcal{T}_\eta)]^{-s}$ implies that $\mathbf{u}$ is in ${}_t\ell^w_\tau(\mathcal{J})$. In particular ${}_t\ell^w_\tau(\mathcal{J})$ is strictly contained in $\mathcal{A}^s$.

Of course, the question arises which property of u implies that the array of wavelet coefficients $\mathbf{u}$ belongs to ${}_t\ell^w_\tau(\mathcal{J})$ and in turn to $\mathcal{A}^s$.

Remark 2.3. Let $H = H^t$. Then $u \in B^{t+sd}_q(L_{\tau'})$ implies $\mathbf{u} \in {}_t\ell^w_\tau(\mathcal{J})$ whenever $\frac{1}{\tau'} < \frac{1}{\tau} = s + \frac{1}{2}$ and $0 < q \le \infty$.

Sketch of proof. It is enough to prove this for $B^{t+sd}_\infty(L_{\tau'})$ and $\tau < \tau' \le 2$, because the remaining cases follow by embeddings. The condition $u \in B^{t+sd}_\infty(L_{\tau'})$ says that the H^t-normalized wavelet coefficients u_λ of u satisfy

$$\left(\sum_{|\lambda|=j} |u_\lambda|^{\tau'}\right)^{1/\tau'} \lesssim 2^{-jd\delta},$$

where $\delta := s + \frac{1}{2} - \frac{1}{\tau'} > 0$ is the discrepancy measuring the "distance" of $B^{t+sd}_\infty(L_{\tau'})$ from the critical embedding line. From this one derives also that $(\sum_{|\lambda|=j} |\tilde{u}_\lambda|^{\tau'})^{1/\tau'} \lesssim 2^{-jd\delta}$, $j \in \mathbb{N}$. This, in turn, implies that the function $\tilde{u}$ with wavelet coefficients $\tilde{\mathbf{u}}$ belongs to $B^{t+sd}_\infty(L_{\tau'})$. By Corollary 4.2 in [8], the best N-term approximation of $\tilde{u}$ in H^t has order N^{-s}. Therefore, by Proposition 2.1, $\tilde{\mathbf{u}} \in \ell^w_\tau(\mathcal{J})$, which, by (2.27) means that $\mathbf{u} \in {}_t\ell^w_\tau(\mathcal{J})$ as claimed. ∎

We therefore have at our disposal two distinct notions of tree approximation rates expressed by the spaces $\mathcal{A}^s$ and ${}_t\ell^w_\tau(\mathcal{J})$. We can now restate the above questions in the following way:

- Does $\mathbf{F}$ map a sequence $\mathbf{u} \in \mathcal{X}$ into a sequence $\mathbf{w} = \mathbf{F}(\mathbf{u}) \in \mathcal{X}$ for $\mathcal{X} = \mathcal{A}^s$ or ${}_t\ell^w_\tau(\mathcal{J})$?
- Can we compute asymptotically optimal sparse approximations of $\mathbf{w} = \mathbf{F}(\mathbf{u})$ from asymptotically optimal sparse approximations of $\mathbf{u}$ in one of the two senses above?

Note that a positive answer to the second question gives a positive answer to the first question in a constructive way. Our next section will give precise answers to these questions for both $\mathcal{A}^s$ and ${}_t\ell^w_\tau(\mathcal{J})$.

3. Sparsity preserving discrete operators.

3.1. General assumptions. We shall use two general assumptions on the function $\mathbf{F}$. The first assumption expresses the fact that $\mathcal{F}$ is a *stable transformation* from H to H'.

Assumption 1. $\mathbf{F}$ is a Lipschitz map from ℓ_2 into itself. More precisely, we assume that we have

$$\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| \le C\|\mathbf{u}-\mathbf{v}\|, \quad \text{with } C = C(\sup\{\|\mathbf{u}\|, \|\mathbf{v}\|\}), \tag{3.1}$$

where $x \mapsto C(x)$ is a positive nondecreasing function.

The fact that the constant C might grow with the norm of $\mathbf{u}$ and $\mathbf{v}$ accounts for the nonlinearity of the transformation. In the context of solving operator equations of the type (2.1), the norms of the arguments of $\mathbf{F}$ will remain bounded (by the $\|\cdot\|_H$-norm of the solution up to the achieved precision) so we can think of C as a constant.

We shall actually use a local version of this stability assumption which will be a direct consequence of (3.1) whenever the nonlinear function $\mathcal{F}$ is local in the physical space: if D is a subdomain of Ω, we have

$$\|(\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v}))|_{\{\lambda : S_\lambda \subset D\}}\| \leq C\|(\mathbf{u} - \mathbf{v})|_{\{\lambda : S_\lambda \cap D \neq \emptyset\}}\|, \tag{3.2}$$

with C depending on $\|\mathbf{u}\|$ and $\|\mathbf{v}\|$ as for the global estimate.

The second assumption describes the *local action of* $\mathbf{F}$ in the space-scale domain of wavelet coefficients.

Assumption 2. If $\mathbf{w} = \mathbf{F}(\mathbf{u})$ for a finitely supported $\mathbf{u}$, we have the estimate

$$|w_\lambda| \leq C \sup_{\mu\,:\,S_\mu \cap S_\lambda \neq \emptyset} |u_\mu| 2^{-\gamma(|\lambda| - |\mu|)} \ \text{ with } \ C = C(\|\mathbf{u}\|) \tag{3.3}$$

for all $\lambda \in \mathcal{J}_\psi$, where $\gamma > d/2$ and $x \mapsto C(x)$ is a positive nondecreasing function.

A typical value of γ is

$$\gamma := r + t + d/2, \tag{3.4}$$

where r reflects the smoothness and order of vanishing moments of the wavelets, i.e., $\psi_\lambda \in C^r$ and $\int_\Omega x^m \psi_\lambda(x) dx = 0$ for $|m| = m_1 + \cdots + m_d < r$. We shall see in the next section that all these assumptions are fulfilled for a fairly general class of local composition operators.

3.2. Tree expansions. Given a tree $\mathcal{T}$, we shall make use of the following expansion process. Given any $\lambda \in \mathcal{J}$, we define $\Phi_0(\lambda) = \{\lambda\}$. If $\Phi_{k-1}(\lambda)$ has already been defined, then we define $\Phi_k(\lambda)$ as the set of all μ, $|\mu| = |\lambda| - k$ such that $S_\mu \cap S_{\mu'} \neq \emptyset$ for some $\mu' \in \Phi_{k-1}(\lambda)$. We define $\Phi(\lambda) := \cup_{k=0}^{|\lambda|} \Phi_k(\lambda)$. We then define the expansion $\tilde{\mathcal{T}}$ as

$$\tilde{\mathcal{T}} := \cup_{\lambda \in \mathcal{T}} \Phi(\lambda). \tag{3.5}$$

Let us note that by construction $\tilde{\mathcal{T}}$ has the following property.

Expansion Property. *If $\mu \in \tilde{\mathcal{T}}$ and $\mu' \in \mathcal{J}$, then*

$$\left.\begin{array}{c} |\mu'| < |\mu| \\ S_{\mu'} \cap S_\mu \neq \emptyset \end{array}\right\} \implies \mu' \in \tilde{\mathcal{T}}. \tag{3.6}$$

The following lemma (see, e.g., [13], [15]) shows that $\tilde{\mathcal{T}}$ has size comparable to $\mathcal{T}$ and that the supports S_λ associated to the outer leaves $\mathcal{L}(\tilde{\mathcal{T}})$ do not overlap too much, a property that we shall use when dealing with the space ${}_t\ell_\tau^w(\mathcal{J})$.

Lemma 3.1. *There exist constants C_1 and C_2 such that for any finite tree $\mathcal{T}$, we have the following.*

(i) $\#(\tilde{\mathcal{T}}) \leq C_1 \#(\mathcal{T})$.

(ii) *For all $\lambda \in \mathcal{L}(\tilde{\mathcal{T}})$ there exist at most C_2 indices $\mu \in \mathcal{L}(\tilde{\mathcal{T}})$ such that $S_\mu \cap S_\lambda \neq \emptyset$.*

Proof. We show first the existence of the constant C_1. To this end, it suffices to show that for each $\mu \in \tilde{\mathcal{T}}$ there exists a *reference element* $\lambda \in \mathcal{T}$ such that $|\lambda| = |\mu|$ and $\operatorname{dist}(S_\lambda, S_\mu) \leq C_0 2^{-|\mu|}$, with C_0 the constant of (2.3). Now any $\mu \in \tilde{\mathcal{T}}$ is in $\Phi_k(\lambda')$ for some $\lambda' \in \mathcal{T}$. We prove by induction on k that there is such a reference

element. For $k = 0$, $\mu = \lambda'$ so we can take $\lambda = \lambda'$. Suppose that we have proven the existence of such a reference element for all $\mu' \in \Phi_{k-1}(\lambda')$ and let μ be an index that has been added in the construction of $\Phi_k(\lambda')$. By the definition of $\Phi_k(\lambda')$ there is a $\mu' \in \Phi_{k-1}(\lambda')$ such that $S_\mu \cap S_{\mu'} \neq \emptyset$. By our induction assumption, there is a reference element $\bar{\lambda} \in \mathcal{T}$, with $|\bar{\lambda}| = |\mu'|$, such that $\operatorname{dist}(S_{\mu'}, S_{\bar{\lambda}}) \leq C_0 2^{-|\mu'|}$. It follows that

$$\operatorname{dist}(S_\mu, S_{\bar{\lambda}}) \leq C_0 2^{-|\mu'|} + \operatorname{diam}(S_{\mu'}) \leq C_0 2^{-|\mu'|} + C_0 2^{-|\mu'|} = C_0 2^{-|\mu|}.$$

Hence, we can take the parent $\lambda \in \mathcal{T}$ of $\bar{\lambda}$ as our reference element for μ.

To confirm the existence of C_2, note that when $\nu, \mu \in \mathcal{L}(\tilde{\mathcal{T}})$ and $S_\nu \cap S_\mu \neq \emptyset$, then $||\nu| - |\mu|| \leq 1$. In fact, suppose that $|\nu| < |\mu| - 1$. Then, for the parent μ' of μ we have $S_\nu \cap S_{\mu'} \neq \emptyset$, since $S_\mu \subset S_{\mu'}$ according to (2.4). Since $\mu' \in \tilde{\mathcal{T}}$ and $|\mu'| > |\nu|$ we conclude $\nu \in \tilde{\mathcal{T}}$, which is a contradiction. This completes the proof. ∎

3.3. The main result in the $\mathcal{A}^s$ case. We first consider the questions raised at the end of the previous section in the $\mathcal{A}^s$ case. Given a tolerance ε we wish to construct a near best tree for approximating $\mathbf{F}(\mathbf{u})$ with accuracy ε, based on the knowledge of $\mathbf{u}$. To this end, suppose that $\mathcal{T}(\varepsilon, \mathbf{u})$ is the near best tree obtained by one of the algorithms in [2] and which satisfies (2.25). For $j = 0, 1, \ldots$, we define

$$\mathcal{T}_j := \mathcal{T}\left(\frac{2^j \varepsilon}{1+j}, \mathbf{u}\right), \tag{3.7}$$

and the corresponding expanded trees $\tilde{\mathcal{T}}_j$ according to the above procedure. By construction these trees are nested in the sense that $\tilde{\mathcal{T}}_j \subset \tilde{\mathcal{T}}_{j-1}$. We define the difference sets

$$\Delta_j := \tilde{\mathcal{T}}_j \setminus \tilde{\mathcal{T}}_{j+1}. \tag{3.8}$$

In order to build a tree which will be adapted to $\mathbf{w} = \mathbf{F}(\mathbf{u})$, we introduce

$$\alpha := \frac{2}{2\gamma - d} > 0, \tag{3.9}$$

where γ is the constant in (3.3), and for each $\mu \in \Delta_j$, we define the *influence set*

$$\Lambda_{\varepsilon,\mu} := \{\lambda \,:\, S_\lambda \cap S_\mu \neq \emptyset \text{ and } |\lambda| \leq |\mu| + \alpha j\}. \tag{3.10}$$

We then define $\mathcal{T}$ by

$$\mathcal{T} := \mathcal{J}_\phi \cup \left(\cup_{\mu \in \tilde{\mathcal{T}}_0} \Lambda_{\varepsilon,\mu}\right). \tag{3.11}$$

We notice that by construction $\mathcal{T}$ has the structure of an expanded tree. Note that an equivalent way of defining $\mathcal{T}$ would be

$$\mathcal{T} := \mathcal{J}_\phi \cup \tilde{\mathcal{T}}_0 \cup \left(\cup_{\mu \in \tilde{\mathcal{T}}_0} \tilde{\Lambda}_{\varepsilon,\mu}\right), \tag{3.12}$$

with

$$\tilde{\Lambda}_{\varepsilon,\mu} := \{\lambda \,:\, S_\lambda \cap S_\mu \neq \emptyset \text{ and } |\mu| \leq |\lambda| \leq |\mu| + \alpha j\}. \tag{3.13}$$

THEOREM 3.2. *Given any* $\mathbf{u}$ *and* $\mathcal{T}$ *defined by* (3.11)*, we have the error estimate*

$$\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{u})|_{\mathcal{T}}\| \lesssim \varepsilon. \tag{3.14}$$

Moreover, if $\mathbf{u} \in \mathcal{A}^s$ *for* $0 < s < \frac{2\gamma - d}{2d}$*, we have the estimate*

$$\#(\mathcal{T}) \lesssim \|\mathbf{u}\|_{\mathcal{A}^s}^{1/s} \varepsilon^{-1/s} + \#(\mathcal{J}_\phi). \tag{3.15}$$

We therefore have $\mathbf{F}(\mathbf{u}) \in \mathcal{A}^s$ *and*

$$\|\mathbf{F}(\mathbf{u})\|_{\mathcal{A}^s} \lesssim 1 + \|\mathbf{u}\|_{\mathcal{A}^s}. \tag{3.16}$$

The constants in these above inequalities depend only on $\|\mathbf{u}\|$*, the space dimension* d*, and the parameter* s*.*

Proof. In order to prove (3.14), we first notice that according to Assumption 1, we have

$$\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{u}|_{\tilde{\mathcal{T}}_0})\| \lesssim \varepsilon, \tag{3.17}$$

with a constant depending on $\|\mathbf{u}\|$. Since one therefore has the trivial estimate $\|(\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{u}|_{\tilde{\mathcal{T}}_0})|_{\mathcal{T}}\| \lesssim \varepsilon$, it suffices to show that $\|\mathbf{w}_\varepsilon - \mathbf{w}_\varepsilon|_{\mathcal{T}}\| \lesssim \varepsilon$, where $\mathbf{w}_\varepsilon := \mathbf{F}(\mathbf{u}|_{\tilde{\mathcal{T}}_0}) = (w_{\varepsilon,\lambda})$. We then remark that Assumption 2 implies the cruder estimate

$$|w_{\varepsilon,\lambda}|^2 \lesssim \sum_{\mu \in \tilde{\mathcal{T}}_0, S_\mu \cap S_\lambda \neq \emptyset} |u_\mu|^2 2^{-2\gamma(|\lambda| - |\mu|)}, \tag{3.18}$$

and for $\lambda \notin \mathcal{T}$ and $\mu \in \Delta_j$ such that $S_\mu \cap S_\lambda \neq \emptyset$, we always have $|\lambda| - |\mu| \geq \alpha j$. Finally we notice that by definition

$$\sum_{\mu \in \Delta_j} |u_\mu|^2 = \|\mathbf{u}|_{\tilde{\mathcal{T}}_j} - \mathbf{u}|_{\tilde{\mathcal{T}}_{j+1}}\|^2 \lesssim \frac{2^{2j}\varepsilon^2}{(1+j)^2}. \tag{3.19}$$

Combining these facts, we obtain

$$\begin{aligned}
\|\mathbf{w}_\varepsilon - \mathbf{w}_\varepsilon|_{\mathcal{T}}\|^2 &= \sum_{\lambda \notin \mathcal{T}} |w_{\varepsilon,\lambda}|^2 \lesssim \sum_{\lambda \notin \mathcal{T}} \sum_{\mu \in \tilde{\mathcal{T}}_0, S_\mu \cap S_\lambda \neq \emptyset} |u_\mu|^2 2^{-2\gamma(|\lambda| - |\mu|)} \\
&= \sum_{\mu \in \tilde{\mathcal{T}}_0} |u_\mu|^2 \sum_{\lambda \notin \mathcal{T}, S_\mu \cap S_\lambda \neq \emptyset} 2^{-2\gamma(|\lambda| - |\mu|)} \\
&= \sum_{j \geq 0} \sum_{\mu \in \Delta_j} |u_\mu|^2 \sum_{\lambda \notin \mathcal{T}, S_\mu \cap S_\lambda \neq \emptyset} 2^{-2\gamma(|\lambda| - |\mu|)} \\
&\lesssim \sum_{j \geq 0} \sum_{\mu \in \Delta_j} |u_\mu|^2 \sum_{k \geq 0} 2^{(d-2\gamma)(\alpha j + k)} \lesssim \sum_{j \geq 0} 2^{(d-2\gamma)\alpha j} \sum_{\mu \in \Delta_j} |u_\mu|^2 \\
&\lesssim \sum_{j \geq 0} 2^{(d-2\gamma)\alpha j} \frac{2^{2j}\varepsilon^2}{(1+j)^2} = \sum_{j \geq 0} \frac{\varepsilon^2}{(1+j)^2} \lesssim \varepsilon^2.
\end{aligned}$$

In order to prove (3.15), we notice that according to (3.12), for each $\mu \in \Delta_j$, we add at most $C 2^{\alpha j d}$ indices to $\mathcal{J}_\phi \cup \tilde{\mathcal{T}}_0$ in the construction of $\mathcal{T}$. Therefore, we have

$$\#(\mathcal{T}) \leq \#(\mathcal{J}_\phi) + \#(\tilde{\mathcal{T}}_0) + C \sum_{j \geq 0} 2^{\alpha j d} \#(\Delta j)$$

$$\lesssim \#(\mathcal{J}_\phi) + \|\mathbf{u}\|_{\mathcal{A}^s}^{1/s}\varepsilon^{-1/s}\left(1+\sum_{j\geq 0}2^{(\alpha d-1/s)j}(1+j)^{1/s}\right)$$
$$\lesssim \#(\mathcal{J}_\phi) + \|\mathbf{u}\|_{\mathcal{A}^s}^{1/s}\varepsilon^{-1/s},$$

where we have used the fact that $\alpha d - 1/s < 0$, which is equivalent to the restriction $s < \frac{2\gamma-d}{2d}$. The estimate (3.16) is then a direct consequence of (3.15). ∎

Note that since we have used Assumptions 1 and 2 in the above proof, the constants in both estimates (3.15) and (3.16) are of the form $C(\|\mathbf{u}\|)$, where $x \mapsto C(x)$ is a positive nondecreasing function.

3.4. The main result in the ${}_t\ell^w_\tau(\mathcal{J})$ case. In order to deal with the ${}_t\ell^w_\tau(\mathcal{J})$ case, we shall build the tree for $\mathbf{F}(\mathbf{u})$ in a slightly different way. To this end, we fix $\eta > 0$ and we define the tree $\mathcal{T}_\eta$ obtained by thresholding the local residuals $\tilde{u}_\lambda$ at level η according to (2.26) and its expanded version $\tilde{\mathcal{T}}_\eta$.

Then for all $\mu \in \tilde{\mathcal{T}}_\eta$, we define the number $n(\mu)$ satisfying

$$\eta 2^{\gamma n(\mu)} \leq |u_\mu| < \eta 2^{\gamma(n(\mu)+1)}. \tag{3.20}$$

We then define the *influence set*

$$\Lambda_{\eta,\mu} := \{\lambda \,:\, S_\lambda \cap S_\mu \neq \emptyset \text{ and } |\lambda| \leq |\mu| + [n(\mu)]_+\} \tag{3.21}$$

and a tree for the approximation of $\mathbf{F}(\mathbf{u})$ by

$$\mathcal{T} := \mathcal{J}_\phi \cup \left(\cup_{\mu\in\tilde{\mathcal{T}}_\eta}\Lambda_{\eta,\mu}\right). \tag{3.22}$$

We notice that by construction $\mathcal{T}$ has the structure of an expanded tree. Note that an equivalent way of defining $\mathcal{T}$ would be

$$\mathcal{T} := \mathcal{J}_\phi \cup \tilde{\mathcal{T}}_\eta \cup \left(\cup_{\mu\in\tilde{\mathcal{T}}_\eta}\tilde{\Lambda}_{\eta,\mu}\right), \tag{3.23}$$

with

$$\tilde{\Lambda}_{\eta,\mu} := \{\lambda \,:\, S_\lambda \cap S_\mu \neq \emptyset \text{ and } |\mu| \leq |\lambda| \leq |\mu| + [n(\mu)]_+\}. \tag{3.24}$$

THEOREM 3.3. *Given any $\mathbf{u} \in \ell_2(\mathcal{J})$ and $\mathcal{T}$ defined by* (3.23)*, one has the coefficient size estimate*

$$|\tilde{w}_\lambda| \lesssim \eta \text{ if } \lambda \notin \mathcal{T}, \tag{3.25}$$

where the $\tilde{w}_\lambda$ are defined for $\mathbf{w} = \mathbf{F}(\mathbf{u})$ according to (2.23)*. If in addition $\mathbf{u} \in {}_t\ell^w_\tau(\mathcal{J})$ for some $d/\gamma < \tau < 2$, then we have the cardinality estimate*

$$\#(\mathcal{T}) \lesssim \|\mathbf{u}\|^\tau_{{}_t\ell^w_\tau(\mathcal{J})}\eta^{-\tau} + \#(\mathcal{J}_\phi). \tag{3.26}$$

Moreover, we have $\mathbf{F}(\mathbf{u}) \in {}_t\ell^w_\tau(\mathcal{J})$ and

$$\|\mathbf{F}(\mathbf{u})\|_{{}_t\ell^w_\tau(\mathcal{J})} \lesssim 1 + \|\mathbf{u}\|_{{}_t\ell^w_\tau(\mathcal{J})}. \tag{3.27}$$

The constants in these above inequalities depend only on $\|\mathbf{u}\|$, the space dimension d, and the parameter τ in the case of (3.26)*.*

Proof. In order to prove (3.25), we first consider the restricted vector $\mathbf{u}_\eta = \mathbf{u}|_{\tilde{\mathcal{T}}_\eta}$ and its image $\mathbf{w}_\eta := \mathbf{F}(\mathbf{u}_\eta) = (w_{\lambda,\eta})$. For $\lambda \notin \mathcal{T}$ and for all $\mu \in \tilde{\mathcal{T}}_\eta$ such that $S_\mu \cap S_\lambda \neq \emptyset$, we have by (3.21) the inequality $|\lambda| - |\mu| \geq [n(\mu)]_+$. Therefore, remarking that $\lambda \in \mathcal{J}_\psi$, the local action assumption (3.3) implies

$$|w_{\lambda,\eta}| \lesssim \eta. \tag{3.28}$$

Moreover, if ν is such that $S_\nu \cap S_\lambda \neq \emptyset$ and $|\nu| = |\lambda| + l$, we also have $|\nu| - |\mu| \geq [n(\mu)]_+ + l$ and therefore, for each $\mu \in \tilde{\mathcal{T}}_\eta$, the better estimate

$$|w_{\nu,\eta}| \lesssim 2^{-\gamma l}\eta. \tag{3.29}$$

It follows that

$$|\tilde{w}_{\lambda,\eta}|^2 \lesssim \eta^2 \left(\sum_{l \geq 0} 2^{(d-2\gamma)l} \right) \lesssim \eta^2, \tag{3.30}$$

since by assumption $\gamma > d/2$.

Next, we remark that for $\lambda \in \mathcal{L}(\mathcal{T})$, we have

$$\begin{aligned} \left| |\tilde{w}_{\lambda,\eta}|^2 - |\tilde{w}_\lambda|^2 \right| &\leq (|\tilde{w}_{\lambda,\eta}| + |\tilde{w}_\lambda|)|\tilde{w}_{\lambda,\eta} - \tilde{w}_\lambda| \\ &\leq (2|\tilde{w}_{\lambda,\eta}| + |\tilde{w}_{\lambda,\eta} - \tilde{w}_\lambda|)|\tilde{w}_{\lambda,\eta} - \tilde{w}_\lambda| \\ &\lesssim (\eta + \|(\mathbf{w} - \mathbf{w}_\eta)|_{\Gamma_\lambda}\|)\|(\mathbf{w} - \mathbf{w}_\eta)|_{\Gamma_\lambda}\|. \end{aligned} \tag{3.31}$$

Now observe that, according to (3.2),

$$\begin{aligned} \|(\mathbf{w} - \mathbf{w}_\eta)|_{\Gamma_\lambda}\| &= \|(\mathbf{w} - \mathbf{w}_\eta)|_{\{\mu : S_\mu \subseteq S_\lambda\}}\| \\ &\lesssim \|(\mathbf{u} - \mathbf{u}_\eta)|_{\{\mu : S_\mu \cap S_\lambda \neq \emptyset\}}\| \\ &= \left(\sum_{\mu \notin \tilde{\mathcal{T}}_\eta, S_\mu \cap S_\lambda \neq \emptyset} |u_\mu|^2 \right)^{1/2} \\ &\leq \left(\sum_{\mu \in \mathcal{L}(\tilde{\mathcal{T}}_\eta), S_\mu \cap S_\lambda \neq \emptyset} |\tilde{u}_\mu|^2 \right)^{1/2} \lesssim \eta, \end{aligned}$$

where the last inequality involves the constant C_2 from Lemma 3.1. Combining this with (3.30) and (3.31), we obtain the size estimate (3.25).

To prove (3.26) we define the trees

$$\tilde{\mathcal{T}}_j := \tilde{\mathcal{T}}_{\eta 2^{\gamma j}}. \tag{3.32}$$

From (2.28) and Lemma 3.1, we infer that $\mathbf{u} \in {}_t\ell^w_\tau(\mathcal{J})$ implies

$$\#(\tilde{\mathcal{T}}_j) \lesssim \eta^{-\tau} 2^{-\gamma\tau j} \|\mathbf{u}\|^\tau_{{}_t\ell^w_\tau(\mathcal{J})}. \tag{3.33}$$

Writing

$$\tilde{\Lambda}_\eta = \bigcup_{j \geq 0} \bigcup_{\mu \in \tilde{\mathcal{T}}_j \setminus \tilde{\mathcal{T}}_{j+1}} \Lambda_{\eta,\mu}, \tag{3.34}$$

so that

$$\bigcup_{\mu\in\tilde{\mathcal{T}}_j\setminus\tilde{\mathcal{T}}_{j+1}}\Lambda_{\eta,\mu}\subseteq\tilde{\mathcal{T}}_j\cup\bigcup_{\mu\in\tilde{\mathcal{T}}_j\setminus\tilde{\mathcal{T}}_{j+1}}\{\lambda:S_\lambda\cap S_\mu\neq\emptyset,\ |\mu|<|\lambda|\le|\mu|+[n(\mu)]_+\},$$

and remarking that, by (3.20) and (3.21), $n(\mu)=j$ for $\mu\in\tilde{\mathcal{T}}_j\setminus\tilde{\mathcal{T}}_{j+1}$, we obtain in view of (3.33)

$$\begin{aligned}\#\left(\bigcup_{\mu\in\tilde{\mathcal{T}}_j\setminus\tilde{\mathcal{T}}_{j+1}}\Lambda_{\eta,\mu}\right)&\lesssim 2^{dj}\#(\tilde{\mathcal{T}}_j\setminus\tilde{\mathcal{T}}_{j+1})+\#(\tilde{\mathcal{T}}_j)\\&\lesssim\|\mathbf{u}\|^\tau_{t\ell^w_\tau(\mathcal{J})}\eta^{-\tau}2^{(d-\gamma\tau)j}.\end{aligned}\tag{3.35}$$

Since $d-\gamma\tau<0$, by summing over $j\ge0$, we obtain

$$\#(\tilde\Lambda_\eta)\lesssim\|\mathbf{u}\|^\tau_{t\ell^w_\tau(\mathcal{J})}\eta^{-\tau},\tag{3.36}$$

and adding the cardinality of $\mathcal{J}_\phi$, we thus obtain (3.26).

In order to obtain the estimate (3.27), we first notice that (3.36) already indicates that we have the estimate

$$\|(\tilde w_\lambda)_{\lambda\in\mathcal{J}_\psi}\|_{\ell^w_\tau(\mathcal{J})}\lesssim\|\mathbf{u}\|_{t\ell^w_\tau(\mathcal{J})}.\tag{3.37}$$

For the remaining indices $\lambda\in\mathcal{J}_\phi$, we can write

$$\|(\tilde w_\lambda)_{\lambda\in\mathcal{J}_\phi}\|_{\ell^w_\tau(\mathcal{J})}\le\|(\tilde w_\lambda)_{\lambda\in\mathcal{J}_\phi}\|_{\ell_\tau}\le[\#(\mathcal{J}_\phi)]^{1/\tau-1/2}\|(\tilde w_\lambda)_{\lambda\in\mathcal{J}_\phi}\|\lesssim\|\mathbf{w}\|,$$

so that we have $\|\mathbf{w}\|_{t\ell^w_\tau(\mathcal{J})}\lesssim\|\mathbf{u}\|_{t\ell^w_\tau(\mathcal{J})}+\|\mathbf{w}\|$. Since by Assumption 1,

$$\|\mathbf{w}\|\lesssim\|\mathbf{F}(0)\|+\|\mathbf{u}\|\lesssim1+\|\mathbf{u}\|\lesssim1+\|\mathbf{u}\|_{t\ell^w_\tau(\mathcal{J})},\tag{3.38}$$

the estimate (3.27) follows. □

Note that since we have used Assumptions 1 and 2 in the above proof, the constants in both estimates (3.25) and (3.27) are of the form $C(\|\mathbf{u}\|)$, where $x\mapsto C(x)$ is a positive nondecreasing function.

Remark 3.1. The limitation $d/\gamma<\tau<2$ in Theorem 3.3 is exactly equivalent to the limitation $0<s<\frac{2\gamma-d}{2d}$ in Theorem 3.2 with $s=1/2-1/\tau$.

Remark 3.2. In both Theorems 3.2 and 3.3, the presence of the constant term in the right side of (3.16) and (3.27), and of the $\#(\mathcal{J}_\phi)$ term on the right side of (3.15) and (3.26), can be avoided if the local action estimate (3.3) remains valid also for $\lambda\in\mathcal{J}_\phi$. The tree $\mathcal{T}$ is then constructed without systematic inclusion of $\mathcal{J}_\phi$, and the proofs of the new estimates are analogous. This is the case for certain classes of linear mappings; see Proposition 7.4 in [11].

3.5. Adaptive evaluation schemes. Adaptive wavelet schemes for variational problems of the type (2.2) rest on two conceptual steps. First (2.2) is formulated in wavelet coordinates as an equivalent problem over $\ell_2(\mathcal{J})$ as follows:

$$\mathbf{Au}+\mathbf{F}(\mathbf{u})=\mathbf{f},\tag{3.39}$$

where $\mathbf{A}=(\langle\nabla\psi_\lambda,\nabla\psi_\nu\rangle)_{\lambda,\nu\in\mathcal{J}}$ is the wavelet representation of Δ and $\mathbf{f}=(\langle f,\psi_\lambda\rangle)_{\lambda\in\mathcal{J}}$. The second step is to devise an iterative scheme for numerically solving (3.39). This

iteration requires the approximate evaluation of $\mathbf{A}\mathbf{u}^n$ and $\mathbf{F}(\mathbf{u}^n)$ with some dynamically updated tolerance, where $\mathbf{u}^n$ is the current finitely supported iterate. How to deal with the linear part $\mathbf{A}\mathbf{u}^n$ has been explained in [9]. The remaining task may therefore be formulated as follows: given a target accuracy $\varepsilon > 0$, and some finitely supported $\mathbf{v} \in \ell_2(\mathcal{J})$, compute $\mathbf{F}(\mathbf{v})$ with accuracy ε at a possibly moderate computational expense.

The way of tackling this task involves two steps : (i) identify an optimal tree such that the restriction of $\mathbf{F}(\mathbf{v})$ to such a tree can be predicted to approximate $\mathbf{F}(\mathbf{v})$ at the desired accuracy, and (ii) numerically compute the coordinates of $\mathbf{F}(\mathbf{v})$ restricted to this tree. We shall not engage (ii) except to mention the paper [15], which treats this topic once (i) has been solved. On the other hand, the results in sections 3.3 and 3.4 allow us to solve (i). We shall again distinguish between near best tree approximation and thresholding the local residual.

When working with near best tree approximation as in section 3.3, Theorem 3.2 provides us with a construction of the required tree, according to the following algorithm.

ALGORITHM EV1. *Given the inputs $\varepsilon > 0$ and $\mathbf{v}$ with finite support do the following:*

Step 1. *Invoke the algorithm in* [2] *to compute the trees*

$$\mathcal{T}_j := \mathcal{T}\left(\frac{2^j\varepsilon}{C_0(j+1)}, \mathbf{v}\right), \tag{3.40}$$

where $C_0 = C_0(\|\mathbf{v}\|)$ is the constant involved in (3.14), *for $j = 0, \dots, J$, and stop for the smallest J such that $\mathcal{T}_J$ is empty (we always have $J \lesssim \log_2(\|\mathbf{v}\|/\varepsilon)$).*

Step 2. *Derive the expanded trees $\tilde{\mathcal{T}}_j$, the layers Δ_j, and the outcome tree $\mathcal{T}$ according to* (3.11).

The following theorem summarizes the properties of Algorithm EV1.

THEOREM 3.4. *Given the inputs $\epsilon > 0$, a nonlinear function F such that $\mathbf{F}$ satisfies Assumptions* 1 *and* 2, *and a finitely supported vector $\mathbf{v}$, the output tree $\mathcal{T}$ has the following properties:*

P1: $\|\mathbf{F}(\mathbf{v}) - \mathbf{F}(\mathbf{v})|_{\mathcal{T}}\| \le \epsilon$.

P2: *For any $0 < s < \frac{2\gamma - d}{2d}$ (see Theorem* 3.2*),*

$$\#(\mathcal{T}) \le C\|\mathbf{v}\|_{\mathcal{A}^s}^{1/s}\epsilon^{-1/s} + \#(\mathcal{J}_\phi) =: N_\varepsilon, \tag{3.41}$$

with C a constant depending only on the constants appearing in Theorem 3.2.

P3: *Moreover, the number of computations needed to find $\mathcal{T}$ is bounded by $C(N_\varepsilon + \#\mathcal{T}(\mathbf{v}))$, where N_ε is the right-hand side of* (3.41) *and $\mathcal{T}(\mathbf{v})$ is the smallest tree containing* $\operatorname{supp}\mathbf{v}$.

Proof. Properties P1 and P2 directly follow from Theorem 3.2. Property P3 is a consequence of the optimality property of the algorithm in [2]. First, the application of this algorithm requires the array $\tilde{\mathbf{v}}$ whose entries serve as local error terms. $\tilde{\mathbf{v}}$ can be computed by summing the squares of the v_λ starting from the leaves of $\mathcal{T}(\mathbf{v})$ towards the roots. The number of operations remains proportional to $\#(\mathcal{T}(\mathbf{v}))$. Moreover, the additional cost of computing each tree $\mathcal{T}_j$, making use of the previously computed tree, is bounded by $C\#(\mathcal{T}_j)$ and therefore by $C\|\mathbf{v}\|_{\mathcal{A}^s}^{1/s}\epsilon^{-1/s}2^{-j/s}(j+1)^{1/s}$. Summing over j we obtain that the total cost remains bounded by the right side of (3.41) and $\#(\mathcal{T}(\mathbf{v}))$. ∎

When working with trees obtained by thresholding the local residuals as in section 3.4, Theorem 3.2 does not provide us with a direct construction of the required tree.

In order to build this tree, we first note that if $\mathcal{T}$ is the tree obtained by (3.23) for some given threshold η, and if $\mathcal{L}$ is the set of its outer leaves, we have the estimate

$$\|\mathbf{v} - \mathbf{F}(\mathbf{v})\|^2 \leq C_0^2 \#(\mathcal{L})\eta^2, \tag{3.42}$$

where C_0 is the constant of the inequality (3.25). Therefore we can invoke the following algorithm based on geometrically updated thresholds $\eta_j = 2^{-j}$.

ALGORITHM EV2. *Given the inputs $\epsilon > 0$ and $\mathbf{v}$ with finite support, initialize with $j = 0$ and do the following:*

Step 1. *Given j compute the predicted tree $\mathcal{T}$ and its outer leaves $\mathcal{L}$ for $\eta = \eta_j = 2^{-j}$. Compute the corresponding error estimator $\varepsilon_0 = C_0^2 \#(\mathcal{L})\eta_j^2$ and proceed to Step* 2.

Step 2. *If $\epsilon_j \leq \epsilon$, terminate the algorithm and take $\mathcal{T}$ as output. If $\epsilon_j > \epsilon$, replace j by $j + 1$ and return to Step* 1.

The following theorem summarizes the properties of Algorithm EV2.

THEOREM 3.5. *Given the inputs $\epsilon > 0$, a nonlinear function F such that $\mathbf{F}$ satisfies Assumptions* 1 *and* 2, *and a finitely supported vector $\mathbf{v}$, the output tree $\mathcal{T}$ has the following properties:*

P1: $\|\mathbf{F}(\mathbf{v}) - \mathbf{F}(\mathbf{v})|_{\mathcal{T}}\| \leq \epsilon$.

P2: *For any $d/\gamma < \tau < 2$ (see Theorem* 3.3*),*

$$\#(\mathcal{T}) \leq C\|\mathbf{v}\|^{1/s}_{{}_t\ell^w_\tau(\mathcal{J})}\epsilon^{-1/s} + \#(\mathcal{J}_\phi), \tag{3.43}$$

with C a constant depending only on the constants appearing in Theorem 3.3.

P3: *Moreover, the number of computations needed to find $\mathcal{T}$ is also bounded by the right side of* (3.43) *plus $\#(\mathcal{T}(\mathbf{v}))$, where $\mathcal{T}(\mathbf{v})$ denotes again the smallest tree containing the support of $\mathbf{v}$.*

Proof. Since the vector $\mathbf{v}$ is finite, it belongs to all ${}_t\ell^w_\tau(\mathcal{J})$. From (3.26) of Theorem 3.3, we have at step j

$$\#(\mathcal{T}) \lesssim 2^{j\tau}\|\mathbf{v}\|^\tau_{{}_t\ell^w_\tau(\mathcal{J})} + \#(\mathcal{J}_\phi), \tag{3.44}$$

from which it follows that ϵ_j tends to 0 as $j \to \infty$. Therefore, the algorithm must terminate at some finite value j^*. It follows from (3.42) that for the tree $\mathcal{T}$ obtained at step j^*, we have

$$\|\mathbf{F}(\mathbf{v}) - \mathbf{F}(\mathbf{v})|_{\mathcal{T}}\|^2 \leq \epsilon^2, \tag{3.45}$$

which proves P1.

In order to prove P2, we start with (3.26), which gives for $\mathcal{T}$ obtained at step j^*

$$\#(\mathcal{T}) \lesssim \|\mathbf{v}\|^\tau_{{}_t\ell^w_\tau(\mathcal{J})}\eta_{j^*}^{-\tau} + \#(\mathcal{J}_\phi) = C_0^\tau\|\mathbf{v}\|^\tau_{{}_t\ell^w_\tau(\mathcal{J})}\epsilon_{j^*}^{-\tau}(\#(\mathcal{L}))^{\tau/2} + \#(\mathcal{J}_\phi). \tag{3.46}$$

We continue under the assumption that the first term on the far right side of (3.46) is bigger than the second, since otherwise we are done. We recall now that $\#(\mathcal{T}) \sim \#(\mathcal{L})$ and that $\epsilon \leq \epsilon_{j^*-1} \leq 4\varepsilon_{j^*}$ because the sets $\mathcal{L}$ increase when j increases. Using this information back in (3.46) gives

$$\#(\mathcal{T})^{1-\tau/2} \lesssim \|\mathbf{v}\|^\tau_{{}_t\ell^w_\tau(\mathcal{J})}\epsilon^{-\tau}. \tag{3.47}$$

Therefore P2 follows by raising both sides of (3.47) to the power $1/(1 - \tau/2)$ because $s = 1/\tau - 1/2$, and hence $1 - \tau/2 = s\tau$.

Finally, to prove P3 we note that the thresholding requires the knowledge of $\tilde{\mathbf{v}}$. The same arguments as in the proof of P3 in Theorem 3.4 yield the contribution of $\#(\mathcal{T}(\mathbf{v}))$ to the operations count. Moreover, once $\tilde{\mathbf{v}}$ is known, the number of additional computations used in the algorithm is bounded by $C(\|\mathbf{v}\|^{\tau}_{t\ell^w_\tau(\mathcal{J})}\eta_0^{-\tau}+\#(\mathcal{J}_\phi))$ at iteration 0. Given the tree at stage j, which allows one to avoid corresponding comparisons, the additional work needed to determine the tree at stage $j+1$ is bounded by $C\|\mathbf{v}\|^{\tau}_{t\ell^w_\tau(\mathcal{J})}\eta_j^{-\tau}$. Summing up over j, we derive P3. ∎

4. Verification of the basic assumptions. We shall show in this section that Assumptions 1 and 2 hold for nonlinear mappings of the form $\mathcal{F}(u)(x) = \mathcal{F}(u(x))$, where $\mathcal{F}$ is a univariate function which satisfies growth conditions at infinity of the type

$$|\mathcal{F}^{(n)}(x)| \le C(1+|x|)^{[p-n]_+}, \quad x \in \mathbb{R},\ n = 0, 1, \dots, n^*, \tag{4.1}$$

for some $p \ge 0$ and n^* a positive integer. Clearly $\mathcal{F}(u) = u^p$ is of this type for all n^* if p is an integer, and with n^* the integer part of p otherwise.

4.1. Verification of Assumption 1. The verification of Assumption 1 is a classical result in the case where $H = H^t(\Omega)$, $t \ge 0$, or when H is a closed subspace of $H^t(\Omega)$ determined, e.g., by homogeneous boundary conditions, such as $H_0^t(\Omega)$ (the closure in the $\|\cdot\|_{H^t}$-norm of smooth functions with compact support in the open bounded domain Ω).

PROPOSITION 4.1. *Assume that $\mathcal{F}$ satisfies* (4.1) *for some $p \ge 0$ and $n^* \ge 0$. Then $\mathcal{F}$ maps H to H' under the restriction*

$$0 \le p \le p^* := \frac{d+2t}{d-2t} \tag{4.2}$$

when $t < d/2$, and with no restriction otherwise. If in addition $n^* \ge 1$, then we also have under the same restriction

$$\|\mathcal{F}(u) - \mathcal{F}(v)\|_{H'} \le C\|u-v\|_H, \tag{4.3}$$

where $C = C(\max\{\|u\|_H, \|v\|_H\})$ and $x \to C(x)$ is nondecreasing, and therefore Assumption 1 *holds.*

Proof. For $u \in H$ and $\varphi \in H$, we write

$$|\langle \mathcal{F}(u), \varphi\rangle| \le C\left[\int_\Omega |\varphi| + \int_\Omega |\varphi||u|^p\right]. \tag{4.4}$$

The first term is bounded according to

$$\int_\Omega |\varphi| \le |\Omega|^{1/2}\|\varphi\|_{L_2} \le |\Omega|^{1/2}\|\varphi\|_H. \tag{4.5}$$

For the second term, we use Hölder's inequality to obtain

$$\int_\Omega |\varphi||u|^p \le \|\varphi\|_{L_q}\|u\|^p_{L_{pq'}}, \tag{4.6}$$

where $\frac{1}{q} + \frac{1}{q'} = 1$. Taking q such that $q = pq' = pq/(q-1)$, i.e., $q = p+1$, this gives

$$\int_\Omega |\varphi||u|^p \le \|\varphi\|_{L_{p+1}}\|u\|^p_{L_{p+1}}. \tag{4.7}$$

We then note that when $t < d/2$, $H = H^t$ is continuously embedded in L_{p+1} if and only if $p \leq p^*$, and this embedding holds for all $p \geq 0$ when $t \geq d/2$. We therefore conclude that

$$\|\mathcal{F}(u)\|_{H'} \leq C(1 + \|u\|_H^p). \tag{4.8}$$

Therefore, $\mathcal{F}$ maps H to H' provided that $p \leq p^*$ when $t < d/2$, and for all $p \geq 0$ otherwise.

For the stability property, we use the inequality

$$|\mathcal{F}(u) - \mathcal{F}(v)| \leq C|u - v|(1 + |u| + |v|)^{[p-1]_+}, \tag{4.9}$$

which is a consequence of (4.1) with $n = 1$. Therefore, one has for all $\varphi \in H$

$$|\langle \mathcal{F}(u) - \mathcal{F}(v), \varphi \rangle| \leq C \left[\int_\Omega |\varphi||u - v| + \int |\varphi||u - v|(|u| + |v|)^{[p-1]_+} \right]. \tag{4.10}$$

The first term is simply bounded by

$$\int_\Omega |\varphi||u - v| \leq \|\varphi\|_{L_2} \|u - v\|_{L_2} \leq \|\varphi\|_H \|u - v\|_H. \tag{4.11}$$

If $p \leq 1$, the second term is bounded analogously. If $p > 1$, we apply Hölder's inequality twice, again with $q = p + 1$, to obtain

$$\int_\Omega |\varphi||u - v|(|u| + |v|)^{p-1} \leq \|\varphi\|_{L_{p+1}} \|u - v\|_{L_{p+1}} \left(\|u\|_{L_{p+1}} + \|v\|_{L_{p+1}} \right)^{p-1}. \tag{4.12}$$

Using again the Sobolev embedding, these factors are controlled by $\|\varphi\|_H$, $\|u - v\|_H$, and $(\|u\|_H + \|v\|_H)^{p-1}$, so that we obtain

$$\|\mathcal{F}(u) - \mathcal{F}(v)\|_{H'} \leq C\|u - v\|_H, \tag{4.13}$$

which is exactly (3.1). ∎

Next we want to prove the local version (3.2) of Assumption 1. For a given subdomain D, we define a vector $\bar{\mathbf{v}} = (\bar{v}_\lambda)$ such that $\bar{v}_\lambda = v_\lambda$ if $S_\lambda \cap D \neq \emptyset$, and $\bar{v}_\lambda = u_\lambda$ otherwise. It follows that

$$\|(\mathbf{u} - \mathbf{v})|_{\{\lambda : S_\lambda \cap D \neq \emptyset\}}\| = \|\mathbf{u} - \bar{\mathbf{v}}\|. \tag{4.14}$$

Denoting by $v, \bar{v}$ the corresponding functions $v = \sum_{\lambda \in \mathcal{J}} v_\lambda \psi_\lambda$, $\bar{v} = \sum_{\lambda \in \mathcal{J}} \bar{v}_\lambda \psi_\lambda$, we clearly have $v = \bar{v}$ on D so that

$$\mathbf{F}(\bar{\mathbf{v}})_\lambda = \langle \mathcal{F}(\bar{v}), \psi_\lambda \rangle = \langle \mathcal{F}(v), \psi_\lambda \rangle = \mathbf{F}(\mathbf{v})_\lambda \tag{4.15}$$

whenever $S_\lambda \subset D$. It follows that

$$\|(\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v}))|_{\{\lambda : S_\lambda \subset D\}}\| \leq \|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\bar{\mathbf{v}})\|. \tag{4.16}$$

Therefore, the local stability estimate (3.2) follows by combining (4.14) and (4.16) together with the global stability estimate (3.1). ∎

4.2. Verification of Assumption 2. For the verification of Assumption 2, we shall assume that either (4.1) is valid or that

$$|\mathcal{F}^{(n)}(x)| \le C(1+|x|)^{p-n}, \quad n \le p, \qquad \text{and } \mathcal{F}^{(n)}(x) = 0, \quad n > p, \tag{4.17}$$

with p an integer. We shall show in the next theorem that whenever $\mathcal{F}$ satisfies either (4.1) or (4.17), then it satisfies Assumption 2 with $\gamma = r + t + d/2$. Our reason for separating the two cases (4.1) and (4.17) is that in the latter case we can take a larger value for r. We recall the critical index p^* defined by (4.2).

THEOREM 4.1. *Assume that the wavelets ψ_λ belong to C^m and have (for those $\lambda \in \mathcal{J}_\psi$) vanishing moments of order m (i.e., are orthogonal to $\mathbb{P}_{m-1}$ the space of polynomials of total degree at most $m-1$) for some positive integer m. Then Assumption 2 holds for $\gamma = r + t + d/2$ with the following value of r:*

(i) *If $t \ge d/2$ and $\mathcal{F}$ satisfies (4.1) for some $p \ge 0$, then $r = \min\{m, n^*\}$.*
(ii) *If $t < d/2$ and $\mathcal{F}$ satisfies (4.1) with $0 \le p < p^*$, then $r = \lceil \min\{m, p, n^*\} \rceil$.*
(iii) *If $t \ge d/2$ and $\mathcal{F}$ satisfies (4.17) for some $p > 0$, then $r = m$.*
(iv) *If $t < d/2$ and $\mathcal{F}$ satisfies (4.17) for some $0 \le p < p^*$, then $r = m$.*

Proof. Suppose that u has a finite wavelet expansion. We assume that $r \ge 1$ and leave the simpler case $r = 0$ to the reader (this case only occurs in (ii) when $p < 1$). Since the wavelets $\psi_\lambda \in \mathcal{J}_\psi$ have at least r vanishing moments, we have

$$\begin{aligned} |w_\lambda| &= |\langle w, \psi_\lambda \rangle| = \inf_{P \in \Pi_{r-1}} |\langle w - P, \psi_\lambda \rangle| \\ &\lesssim |w|_{W^r(L_\infty(S_\lambda))} 2^{-r|\lambda|} 2^{-(t+d/2)|\lambda|} = |w|_{W^r(L_\infty(S_\lambda))} 2^{-\gamma|\lambda|}, \end{aligned} \tag{4.18}$$

where $w(x) = \mathcal{F}(u(x))$. Using the chain rule, any rth order derivative of w can be written as a finite sum of functions of the form

$$\mathcal{F}^{(k)}(u) D^{\beta_1} u \cdots D^{\beta_k} u, \quad k = 1, \dots, r, \tag{4.19}$$

where $|\beta_1| + \cdots + |\beta_k| = r$ with the usual notation $|\beta_i| := \beta_{i,1} + \cdots + \beta_{i,d}$. Therefore, one has

$$|w|_{W^r(L_\infty(S_\lambda))} \lesssim \max_{k=1,\dots,r} \max_{|\beta_1|+\cdots+|\beta_k|=r} \|\mathcal{F}^{(k)}(u)\|_{L_\infty(S_\lambda)} \prod_{i=1}^{k} \|D^{\beta_i} u\|_{L_\infty(S_\lambda)}. \tag{4.20}$$

To bound the right side of (4.20), we recall that

$$\|D^{\beta_i} u\|_{L_\infty(S_\lambda)} \lesssim \|u\|_{L_\infty(S_\lambda)}^{1-|\beta_i|/r} |u|_{W^r(L_\infty(S_\lambda))}^{|\beta_i|/r}. \tag{4.21}$$

This gives for $k = 1, \dots, r$

$$\begin{aligned} \|\mathcal{F}^{(k)}(u)\|_{L_\infty(S_\lambda)} \prod_{i=1}^{k} \|D^{\beta_i} u\|_{L_\infty(S_\lambda)} &\lesssim \|\mathcal{F}^{(k)}(u)\|_{L_\infty(S_\lambda)} \|u\|_{L_\infty(S_\lambda)}^{k-1} \\ &\quad \times |u|_{W^r(L_\infty(S_\Lambda))}. \end{aligned} \tag{4.22}$$

We shall finish the proof by separating it into two cases depending on the size of $\|u\|_{L_\infty(S_\lambda)}$.

Case $\|u\|_{L_\infty(S_\lambda)} \ge 1$. In this case, (4.20), (4.22), and the bounds (4.1) and (4.17) give

$$|w|_{W^r(L_\infty(S_\lambda))} \lesssim \|u\|_{L_\infty(S_\lambda)}^{M} |u|_{W^r(L_\infty(S_\lambda))}, \tag{4.23}$$

where $M := \max\{p, r\} - 1$ in cases (i) and (ii), and $M := p - 1$ in cases (iii) and (iv). We can bound the norms on the right side of (4.23) by using the Besov spaces $B^s_\infty(L_\infty)$, $s \geq 0$. They satisfy the norm equivalences

$$\|v\|_{B^s_\infty(L_\infty(S_\lambda))} \sim \sup_{S_\mu \cap S_\lambda \neq \emptyset} \left(2^{s|\mu|}\|v_\mu \psi_\mu\|_{L_\infty}\right) = \sup_{S_\mu \cap S_\lambda \neq \emptyset} \left(2^{(s+\delta)|\mu|}|v_\mu|\right) \tag{4.24}$$

with $\delta := \frac{d}{2} - t$. Here we used the fact that the H-normalization of the wavelets implies $\|\psi_\mu\|_{L_\infty} \sim 2^{\delta|\mu|}$. We also recall the embedding estimates

$$\|u\|_{W^s(L_\infty(S_\lambda))} \lesssim \|u\|_{B^{s+\varepsilon}_\infty(L_\infty(S_\lambda))} \tag{4.25}$$

for any fixed $\varepsilon \in]0, 1[$ and all $s \geq 0$. Using all of this in (4.23), we obtain

$$\begin{aligned} |w|_{W^r(L_\infty(S_\lambda))} &\lesssim \left(\sup_{S_\mu \cap S_\lambda \neq \emptyset} 2^{(\delta+\varepsilon)|\mu|}|u_\mu|\right)^M \left(\sup_{S_\mu \cap S_\lambda \neq \emptyset} 2^{(r+\delta+\varepsilon)|\mu|}|u_\mu|\right) \\ &= \left(2^{(\delta+\varepsilon)|\mu_0|}|u_{\mu_0}|\right)^M \left(2^{(r+\delta+\varepsilon)|\mu_1|}|u_{\mu_1}|\right), \end{aligned} \tag{4.26}$$

where μ_0 and μ_1 are the maximizing indices. If $\delta < 0$ (i.e., $t > d/2$), we can take $\epsilon < |\delta|$ and obtain the bound

$$|w|_{W^r(L_\infty(S_\lambda))} \lesssim \|\mathbf{u}\|^M (2^{(r+\delta+\varepsilon)|\mu_1|}|u_{\mu_1}|) \leq \|\mathbf{u}\|^M \sup_{S_\mu \cap S_\lambda \neq \emptyset} 2^{r|\mu|}|u_\mu|, \tag{4.27}$$

which verifies Assumption 2 in this case. If $\delta > 0$, then $|\mu_1| \geq |\mu_0|$ and $p < p_*$ and $M = p - 1$, and so we obtain

$$|w|_{W^r(L_\infty(S_\lambda))} \lesssim \|\mathbf{u}\|^M (2^{(r+p\delta+p\varepsilon)|\mu_1|}|u_{\mu_1}|) \leq \|\mathbf{u}\|^M \sup_{S_\mu \cap S_\lambda \neq \emptyset} 2^{(r+t+d/2)\mu|}|u_\mu|, \tag{4.28}$$

provided $p\epsilon < (p^* - p)\delta$. So we have completed the proof in this case.

Case $\|u\|_{L_\infty(S_\lambda)} < 1$. In this case, starting from (4.22) and using either (4.1) or (4.17), we obtain

$$\begin{aligned} |w|_{W^r(L_\infty(S_\lambda))} &\lesssim |u|_{W^r(L_\infty(S_\lambda))} \lesssim \|u\|_{B^{r+\varepsilon}_\infty(L_\infty(S_\lambda))} \\ &\lesssim \sup_{S_\mu \cap S_\lambda \neq \emptyset} 2^{(r+\delta+\epsilon)|\mu|}|u_\mu| \lesssim \sup_{S_\mu \cap S_\lambda \neq \emptyset} 2^{(r+d/2+t)\mu|}|u_\mu|, \end{aligned} \tag{4.29}$$

provided $\varepsilon < 2t$. Therefore, we have verified Assumption 2 in this case as well. ∎

5. Multiple arguments and derivatives. In this final section, we shall extend the previous results to more general nonlinear operators of the form

$$(u_1, \ldots, u_n) \mapsto w = \mathcal{F}(D^{\alpha_1}u_1, \ldots, D^{\alpha_n}u_n), \tag{5.1}$$

acting from $H \times \cdots \times H$ to its dual H' (note that $\alpha_i = (\alpha_{i,1}, \ldots, \alpha_{i,d})$ are multi-indices). These include multilinear operators as particular cases. Here we shall indicate the appropriate generalizations of the results in the two previous sections with brief sketches of proofs since they are quite similar but notationally heavier.

5.1. Sparsity preserving discrete operators. Denoting by $\mathbf{u}_i = (u_{i,\lambda})$ the arrays of the wavelet coefficients of the function u_i, $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_n)$, and $\mathbf{F}$ the corresponding discrete mapping

$$\mathbf{F}(\mathbf{u}) := (\langle \mathcal{F}(D^{\alpha_1} u_1, \dots, D^{\alpha_n} u_n), \psi_\lambda \rangle_{\lambda \in \mathcal{J}}, \tag{5.2}$$

we introduce the following generalization of the basic assumptions.

Assumption 1. $\mathbf{F}$ is a Lipschitz map from $(\ell_2)^n$ into ℓ_2:

$$\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| \le C \sum_{i=1}^n \|\mathbf{u}_i - \mathbf{v}_i\|, \tag{5.3}$$

with $C = C(\max_i\{\|\mathbf{u}_i\|, \|\mathbf{v}_i\|\})$, where $x \mapsto C(x)$ is a positive nondecreasing function.

The local version of this stability assumption now reads

$$\|(\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v}))|_{\{\lambda : S_\lambda \subset D\}}\| \le C \sum_{i=1}^n \|(\mathbf{u}_i - \mathbf{v}_i)|_{\{\lambda : S_\lambda \cap D \neq \emptyset\}}\| \tag{5.4}$$

for any domain D.

Assumption 2. For any finitely supported $\mathbf{u}$ (i.e., with all $\mathbf{u}_i$ finitely supported) and $\mathbf{w} = \mathbf{F}(\mathbf{u})$, we have the estimate

$$|w_\lambda| \le C \sup_{\mu \,:\, S_\lambda \cap S_\mu \neq \emptyset} \left(\sum_{i=1}^n |u_{i,\mu}| \right) 2^{-\gamma(|\lambda| - |\mu|)} \tag{5.5}$$

for all $\lambda \in \mathcal{J}_\psi$, where $\gamma > d/2$, $C = C(\max_i \|\mathbf{u}_i\|)$, and $x \mapsto C(x)$ is a positive nondecreasing function.

In order to generalize the construction of the near best approximation tree from section 3.3, we construct for a prescribed accuracy ε the trees $\tilde{\mathcal{T}}_{j,i}$ for each component $\mathbf{u}_i$ in the same way as we constructed $\tilde{\mathcal{T}}_j$ in section 3.3. We then define

$$\tilde{\mathcal{T}}_j := \cup_{i=0}^n \tilde{\mathcal{T}}_{j,i} \ \text{ and } \ \Delta_j := \tilde{\mathcal{T}}_j \setminus \tilde{\mathcal{T}}_{j+1}. \tag{5.6}$$

The tree $\mathcal{T}$ is then constructed as before, according to (3.11).

THEOREM 5.1. *With this definition of* $\mathcal{T}$ *and under the above generalized Assumptions* 1 *and* 2, *if* $\mathbf{u} \in (\mathcal{A}^s)^n$, *we obtain that the same conclusions as in Theorem* 3.2 *also hold.*

Sketch of proof. As in the proof of Theorem 3.2, we start by invoking the stability property, which leads us to estimate $\|\mathbf{w}_\varepsilon - \mathbf{w}_\varepsilon|_{\mathcal{T}}\|$, where $\mathbf{w}_\varepsilon := \mathbf{F}(\mathbf{u}|_{\tilde{\mathcal{T}}_0})$. We use the estimate

$$|w_{\varepsilon,\lambda}|^2 \ \lesssim \ \sum_{\mu \in \tilde{\mathcal{T}}_0, S_\mu \cap S_\lambda \neq \emptyset} \left[\sum_{i=1}^n |u_{i,\mu}|^2 \right] 2^{-2\gamma(|\lambda| - |\mu|)}, \tag{5.7}$$

together with

$$\sum_{\mu \in \Delta_j} \sum_{i=1}^n |u_{i,\mu}|^2 = \|\mathbf{u}|_{\tilde{\mathcal{T}}_j} - \mathbf{u}|_{\tilde{\mathcal{T}}_{j+1}}\|^2 \ \lesssim \ \frac{2^{2j}\varepsilon^2}{(1+j)^2}, \tag{5.8}$$

in order to derive $\|\mathbf{w}_\varepsilon - \mathbf{w}_\varepsilon|_{\mathcal{T}}\| \lesssim \varepsilon$ in a similar way as in the proof of Theorem 3.2. The estimate on $\#(\mathcal{T})$ also remains the same using that $\#(\Delta_j) \le \#(\mathcal{T}_j) \le \sum_{i=1}^n \#(\mathcal{T}_{j,i})$. ∎

In order to generalize the construction of the residual thresholding tree from section 3.4, we fix a threshold $\eta > 0$ and define for all $\mu \in \mathcal{J}$ the number $n(\mu)$ satisfying

$$\eta 2^{\gamma n(\mu)} \leq \max_i |u_{i,\mu}| < \eta 2^{\gamma(n(\mu)+1)}, \tag{5.9}$$

where $u_{i,\mu}$ are the coefficients of $\mathbf{u}_i$. We then define the influence sets $\Lambda_{\eta,\mu}$ and the tree $\mathcal{T}$ in a similar way as in (3.21) and (3.23), with $\tilde{\mathcal{T}}_\eta = \cup_{i=1}^n \tilde{\mathcal{T}}_\eta(\mathbf{u}_i)$ and $\tilde{\mathcal{T}}_\eta(\mathbf{u}_i)$ the expansion of the tree $\mathcal{T}_\eta(u_i)$.

THEOREM 5.2. *With this definition of $\mathcal{T}$ and under the above generalized Assumptions* 1 *and* 2*, if $\mathbf{u} \in (\ell^w_\tau(\mathcal{J}))^n$, we obtain that the same conclusions as in Theorem* 3.3 *hold.*

Sketch of proof. As in the proof of Theorem 3.3, in order to prove (3.25) we first consider the restricted vector $\mathbf{u}_\eta = \mathbf{u}|_{\tilde{\mathcal{T}}_\eta}$ and its image $\mathbf{w}_\eta := \mathbf{F}(\mathbf{u}_\eta) = (w_{\lambda,\eta})$. Using (5.5), we obtain that for any $\lambda \notin \mathcal{T}$, we have $|\tilde{w}_{\lambda,\eta}| \lesssim \eta$. We then use (5.4) in a similar way in order to derive (3.25).

In order to prove (3.26), we use the trees $\tilde{\mathcal{T}}_j := \tilde{\mathcal{T}}_{\eta 2^{\gamma j}}$ to decompose $\mathcal{T}$ into layers indexed by j as in the proof of Theorem 3.3. We then proceed in a similar way to derive (3.26), remarking that

$$\#(\tilde{\mathcal{T}}_j) \lesssim \eta^{-\tau} 2^{-\gamma\tau j} \sup_i \|\mathbf{u}_i\|^\tau_{{}_t\ell^w_\tau(\mathcal{J})}, \tag{5.10}$$

and that, according to (5.9) and (3.21), $n(\mu) = j$ for $\mu \in \tilde{\mathcal{T}}_j \setminus \tilde{\mathcal{T}}_{j+1}$.

Finally, we prove

$$\|\mathbf{w}\|_{{}_t\ell^w_\tau(\mathcal{J})} \lesssim 1 + \sup_i \|\mathbf{u}_i\|_{{}_t\ell^w_\tau(\mathcal{J})} \tag{5.11}$$

by the same arguments as in the proof for Theorem 3.3. □

The generalization of Algorithms EV1 and EV2 is straightforward, as is the following proposition.

PROPOSITION 5.1. *If $\mathbf{u} \in \mathcal{X}$ with $\mathcal{X} = (\mathcal{A}^s)^n$ (resp., $({}_t\ell^w_\tau(\mathcal{J}))^n$), the tree $\mathcal{T}$ produced by Algorithm* EV1 *(resp.,* EV2*) for target accuracy ε satisfies*

$$\#(\mathcal{T}) \lesssim \sup_i \|\mathbf{u}_i\|^{1/s}_{\mathcal{X}} \varepsilon^{-1/s} + \#(\mathcal{J}_\phi), \tag{5.12}$$

with $s = 1/\tau - 1/2$, under the restriction $d/\gamma < \tau < 2$.

5.2. Verification of the basic assumptions. Recalling that the nonlinear map has the form $\mathcal{F}(D^{\alpha_1}u_1, \ldots, D^{\alpha_n}u_n)$, we shall therefore replace (4.1) by growth assumptions of the type

$$|D^\beta \mathcal{F}(x_1, \ldots, x_n)| \leq C \prod_{i=1}^n (1 + |x_i|)^{[p_i - \beta_i]_+}, \quad |\beta| = 0, 1, \ldots, n^*, \tag{5.13}$$

for some $p_i \geq 0$ and n^* a positive integer. For notational simplicity, we shall write

$$\mathcal{F}(u) = \mathcal{F}(D^{\alpha_1}u_1, \ldots, D^{\alpha_n}u_n), \quad \text{with } u = (u_1 \ldots, u_n). \tag{5.14}$$

We then obtain the following generalization of Proposition 4.1.

PROPOSITION 5.2. *Assume that the growth assumptions* (5.13) *hold at least with* $n^* = 0$. *Then* $\mathcal{F}$ *maps* $H \times \cdots \times H$ *to* H' *whenever* $H = H^t$ *and* $t \geq 0$ *satisfies*

$$\left[\frac{1}{2} - \frac{t}{d}\right]_+ + \sum_{i=1}^n p_i \left[\frac{1}{2} - \frac{t}{d} + \frac{|\alpha_i|}{d}\right]_+ < 1. \tag{5.15}$$

If in addition $n^* = 1$, *then we also have under the same restriction*

$$\|\mathcal{F}(u) - \mathcal{F}(v)\|_{H'} \leq C \sum_{i=1}^n \|u_i - v_i\|_H, \tag{5.16}$$

where $C = C(\max_i\{\|u_i\|_H, \|v_i\|_H\})$ *and* $x \to C(x)$ *is nondecreasing, and therefore Assumption* 1 *holds.*

Sketch of proof. For $u_i \in H$ and $\varphi \in H$, we write

$$|\langle \mathcal{F}(u), \varphi\rangle| \leq C \int_\Omega |\varphi| \prod_{i=1}^n (1 + |D^{\alpha_i} u_i|)^{p_i}. \tag{5.17}$$

In view of (5.15), we can choose positive numbers r and r_i, $i = 1, \ldots, n$, such that $\frac{1}{r} + \sum_{i=1}^n \frac{p_i}{r_i} = 1$ and

$$\frac{1}{r} > \frac{1}{2} - \frac{t}{d} \ \text{ and } \ \frac{1}{r_i} > \frac{1}{2} - \frac{t}{d} + \frac{|\alpha_i|}{d}. \tag{5.18}$$

It follows that H^t is continuously embedded in L_r and $W^{|\alpha_i|}(L_{r_i})$. We can apply Hölder's inequality to obtain

$$|\langle \mathcal{F}(u), \varphi\rangle| \leq C\|\varphi\|_{L_r} \prod_{i=1}^n (1 + \|D^{\alpha_i} u_i\|_{L_{r_i}}^{p_i}), \tag{5.19}$$

where we have used the fact that Ω is a bounded domain in order to control $\int_\Omega 1$ by a constant. In this way, we obtain

$$\|\mathcal{F}(u)\|_{H'} \leq C \prod_{i=1}^n (1 + \|D^{\alpha_i} u_i\|_H^{p_i}). \tag{5.20}$$

For the stability property, we use the inequality

$$|\mathcal{F}(u) - \mathcal{F}(v)| \leq C \sum_{i=1}^n |D^{\alpha_i} u_i - D^{\alpha_i} v_i| \prod_{k=1}^n (1 + |D^{\alpha_k} u_k| + |D^{\alpha_k} v_k|)^{[p_k - \delta_{i,k}]_+}, \tag{5.21}$$

with δ the Kronecker delta. Therefore, when estimating $|\langle \mathcal{F}(u) - \mathcal{F}(v), \varphi\rangle|$ for $\varphi \in H$, we are led to expressions of the form

$$E_i = \int_\Omega |\varphi| |D^{\alpha_i} u_i - D^{\alpha_i} v_i| \prod_{k=1}^n (1 + |D^{\alpha_k} u_k| + |D^{\alpha_k} v_k|)^{[p_k - \delta_{i,k}]_+} \tag{5.22}$$

for each i. Using Hölder's inequality, we obtain

$$E_i \leq C\|\varphi\|_{L_r} \|D^{\alpha_i} u_i - D^{\alpha_i} v_i\|_{L_q} \prod_{k=1}^n \left(1 + \|D^{\alpha_k} u_k\|_{L_{r_k}}^{[p_k - \delta_{i,k}]_+} + \|D^{\alpha_k} v_k\|_{L_{r_k}}^{[p_k - \delta_{i,k}]_+}\right)$$

whenever

$$\frac{1}{r}+\frac{1}{q}+\sum_{k=1}^{n}\frac{[p_k-\delta_{i,k}]_+}{r_k}=1. \tag{5.23}$$

In view of (5.15), we can choose positive numbers r, q, and r_i satisfying condition (5.23) such that

$$\frac{1}{r}>\frac{1}{2}-\frac{t}{d},\quad \frac{1}{q}>\frac{1}{2}-\frac{t}{d}+\frac{|\alpha_i|}{d},\quad \text{and}\quad \frac{1}{r_k}>\frac{1}{2}-\frac{t}{d}+\frac{|\alpha_k|}{d}. \tag{5.24}$$

Therefore, the Sobolev embedding gives

$$E_i\le C\|\varphi\|_H\|u_i-v_i\|_H, \tag{5.25}$$

and therefore Assumption 1 holds. ∎

The local version (3.2) of Assumption 1 is derived in the same way as in section 4.2. Note that the condition (5.15) does not yield the optimal condition (4.2) in the simple case $n=1$ and $\alpha_1=0$ due to the strict inequality but that we anyway need this strict inequality in order to obtain the validity of Assumption 2 according to Theorem 4.1.

For the proof of Assumption 2, we again treat separately the polynomial case for which we have the growth condition

$$|D^\beta\mathcal{F}(x_1,\dots,x_n)|\le C\prod_{i=1}^{n}(1+|x_i|)^{p_i-\beta_i},\quad \beta_i\le p_i, \tag{5.26}$$

and $D^\beta\mathcal{F}=0$ if $\beta_i>p_i$ for some i, where p_i are positive integers.

THEOREM 5.3. *Assume that the wavelets belong to C^m and have vanishing moments of order m (i.e., are orthogonal to $\mathbb{P}_{m-1}$ the space of polynomials of total degree at most $m-1$) for some positive integer m. Then Assumption 2 holds for $\gamma=r+t+d/2$ with the following values of r:*

(i) *If $\mathcal{F}$ satisfies (5.13) with p such that $\sum_{i=1}^n p_i[d/2-t+|\alpha_i|]_+<d/2+t$, then $r=\lceil\min\{m,n^*,p^*\}\rceil$, where $p^*=\min\{p_i\ :\ i \text{ s.t. } d/2-t+|\alpha_i|>0\}$.*

(ii) *If $\mathcal{F}$ satisfies (5.26) with p such that $\sum_{i=1}^n p_i[d/2-t+|\alpha_i|]_+<d/2+t$, then $r=m$.*

Sketch of proof. We shall prove (i); the other case is similar. We shall also assume that $r\ge 1$ and leave the simpler case $r=0$ to the reader. As in the proof of Theorem 4.1 we start from the estimate

$$|w_\lambda|\ \lesssim\ |w|_{W^r(L_\infty(S_\lambda))}2^{-\gamma|\lambda|}, \tag{5.27}$$

where $w(x)=\mathcal{F}(u(x))$. Using the chain rule, any rth order derivative of w can be written as a finite sum of functions of the form

$$D^\nu\mathcal{F}(D^{\alpha_1}u_1,\dots,D^{\alpha_n}u_n)G_\nu,\quad |\nu|=1,\dots,r, \tag{5.28}$$

where

$$G_\nu=\prod_{i=1}^{n}\prod_{j=1}^{\nu_i}D^{\beta_{i,j}+\alpha_i}u_i, \tag{5.29}$$

and $\sum_{i=1}^n \sum_{j=1}^{\nu_i} |\beta_{i,j}| = r$. Therefore, one has

$$|w|_{W^r(L_\infty(S_\lambda))} \lesssim \max_{|\nu|\le r} A_\nu, \tag{5.30}$$

where

$$A_\nu := \|D^\nu \mathcal{F}(D^{\alpha_1}u, \dots, D^{\alpha_n}u_n)\|_{L_\infty(S_\lambda)} \|G_\nu\|_{L_\infty(S_\lambda)}. \tag{5.31}$$

To bound $\|G_\nu\|_{L_\infty(S_\lambda)}$, we use the estimate for intermediate derivatives (4.21) and find with $r_i := \sum_{j=1}^{\nu_i} |\beta_{i,j}|$ that

$$\|G_\nu\|_{L_\infty(S_\lambda)} \lesssim \prod_{i=1}^n |u_i|^{\nu_i - 1}_{W^{|\alpha_i|}(L_\infty(S_\lambda))} |u_i|_{W^{r_i+|\alpha_i|}(L_\infty(S_\lambda))}. \tag{5.32}$$

We now invoke (5.13) and obtain that

$$\begin{aligned} A_\nu &\le \prod_{i=1}^n (1 + |u_i|_{W^{|\alpha_i|}(L_\infty(S_\lambda))})^{(p_i-\nu_i)_+} |u_i|^{\nu_i-1}_{W^{|\alpha_i|}(L_\infty(S_\lambda))} |u|_{W^{r_i+|\alpha_i|}(L_\infty(S_\lambda))} \\ &\le \prod_{i=1}^n |u_i|^{M_i}_{W^{|\alpha_i|}(L_\infty(S_\lambda))} |u|_{W^{r_i+|\alpha_i|}(L_\infty(S_\lambda))}, \end{aligned} \tag{5.33}$$

where $M_i = \max(p_i, r_i) - 1$ if $|u_i|_{W^{|\alpha_i|}(L_\infty(S_\lambda))} \ge 1$, and $M_i = 0$ otherwise.

Each term appearing in the last product in (5.33) can be bounded by Besov norms. The arguments used in deriving (4.28) and (4.29) give

$$|u_i|^{M_i}_{W^{|\alpha_i|}(L_\infty(S_\lambda))} |u|_{W^{r_i+|\alpha_i|}(L_\infty(S_\lambda))} \lesssim \|\mathbf{u}\|^{M_i} 2^{(r_i+(M_i+1)(|\alpha_i|+\delta+\epsilon))|\mu_i|} |u_{i,\mu_i}|, \tag{5.34}$$

where μ_i is a maximizing index. Let $\mu^* := \max_i \mu_i$. We place (5.34) into (5.33). Each term $|u_{i,\mu_i}|$, $\mu_i \ne \mu^*$, we pull out of the product by the majorant $\|\mathbf{u}\|$. This then gives

$$A_\nu \lesssim \prod_{i=1}^n \|\mathbf{u}\|^{M_i} 2^{(r_i+(M_i+1)(|\alpha_i|+\delta+\epsilon))|\mu_i|} |u_{i,\mu_i}| \lesssim \|\mathbf{u}\|^M \left[\sum_{i=1}^n |u_{i,\mu^*}|\right] 2^{\tilde\gamma|\mu^*|}, \tag{5.35}$$

with $M = \sum_{i=1}^n M_i$ and

$$\tilde\gamma = r + \sum_{i=1}^n (1 + M_i)(\varepsilon + [d/2 - t + |\alpha_i|]_+). \tag{5.36}$$

Now, consider any term in the sum which is not zero. If $M_i \ne 0$, then $M_i + 1 = \max(p_i, r_i) \le \max(p_i, r) = p_i$ because $r \le p_i$. If $M_i = 0$, then $M_i + 1 = 1 \le p_i$ because by definition $r \le p^* \le p_i$, and we have assumed $r \ge 1$. Using this information in (5.36) shows that

$$\tilde\gamma \le r + \sum_{i=1}^n p_i(\varepsilon + [d/2 - t + |\alpha_i|]_+) \le \gamma, \tag{5.37}$$

provided ϵ is sufficiently small. □

REFERENCES

[1] S. Bertoluzza, *Adaptive wavelet collocation for the solution of steady state equation*, in SPIE Proc. Wavelet Appl. II, 2491 (1995), pp. 947–956.

[2] P. Binev and R. DeVore, *Fast computations in adaptive tree approximation*, Numer. Math., to appear.

[3] G. Bourdaud, *Fonctions qui opèrent sur les espaces de Besov et de Triebel*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (1993), pp. 413–422.

[4] G. Bourdaud and D. Kateb, *Fonctions qui opèrent sur les espacs de Besov*, Math. Ann., 303 (1995), pp. 653–675.

[5] C. Canuto and I. Cravero, *Wavelet-based adaptive methods for advection-diffusion problems*, Math. Models Methods Appl. Sci., 7 (1997), pp. 265–289.

[6] C. Canuto, A. Tabacco, and K. Urban, *The wavelet element method, part* I: *Construction and analysis*, Appl. Comput. Harmon. Anal., 6(1999), pp. 1–52.

[7] A. Cohen and R. Masson, *Wavelet adaptive methods for second order elliptic problems, boundary conditions and domain decomposition*, Numer. Math., 86 (2000), pp. 193–238.

[8] A. Cohen, W. Dahmen, I. Daubechies, and R. DeVore, *Tree approximation and optimal encoding*, Appl. Comp. Harm. Anal., 11 (2001), pp. 192–226.

[9] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods for elliptic operator equations—Convergence rates*, Math. Comp., 70 (2001), pp. 27–75.

[10] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods* II: *Beyond the elliptic case*, Found. Comput. Math., 2 (2002), pp. 203–245.

[11] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive Wavelet Schemes for Nonlinear Variational Problems*, IGPM Report 221, RWTH Aachen, Aachen, Germany, 2002.

[12] A. Cohen, S.M. Kaber, S. Müller, and M. Postel, *Fully adaptive multiresolution finite volume schemes for conservation laws*, Math. Comp., 72 (2003), pp. 183–225.

[13] W. Dahmen, *Adaptive approximation by smooth multivariate splines*, J. Approx. Theory, 36 (1982), pp. 119–140.

[14] W. Dahmen and R. Schneider, *Composite Wavelet Bases for Operator Equations*, Math. Comp., 68 (1999), pp. 1533–1567.

[15] W. Dahmen, R. Schneider, and Y. Xu, *Nonlinear functions of wavelet expansions: Adaptive reconstruction and fast evaluation*, Numer. Math., 86 (2000), pp. 49–101.

[16] R. DeVore, *Nonlinear Approximation*, in Acta Numer. 7, Cambridge University Press, Cambridge, UK, 1998, pp. 51–150.

[17] M. Farge, M. Griebel, F. Koster, and K. Schneider, *Adaptive wavelet methods for the Navier-Stokes equations*, in Numerical Flow Simulation II, E. H. Hirschel, ed., Notes Numer. Fluid Mech. 75, Springer-Verlag, Berlin, 2001, pp. 303–318.

[18] M. Farge and K. Schneider, *Numerical simulation of a mixing layer in adaptive wavelet basis*, C. R. Acad. Sci. Paris Sér. II Fasc. b Méc., 328 (2000), pp. 263–269.

[19] A. Harten, *Adaptive multiresolution schemes for shock computations*, J. Comput. Phys., 115 (1994), pp. 319–338.

[20] Y. Maday, V. Perrier, and J.C. Ravel, *Adaptativité dynamique sur bases d'ondelettes pour l'approximation d'équations aux dérivées partielles*, C. R. Acad. Sci. Paris Sér. I, 1 (1991), pp. 405–410.

[21] T. Runst and W. Sickel, *Sobolev Spaces of Fractional Order, Nemitskij Operators, and Nonlinear Partial Differential Equations*, de Gruyter Ser. Nonlinear Anal. Appl. 3, de Gruyter, New York, 1996.

[22] W. Sickel, *Necessary conditions on composition operators acting between Besov spaces. The case* $1 < s < n/p$, Forum Math., 10 (1998), pp. 303–327.

SIAM J. MATH. ANAL.
Vol. 35, No. 2, pp. 304–317

UNIQUE CONTINUATION FOR AN ELASTICITY SYSTEM WITH RESIDUAL STRESS AND ITS APPLICATIONS*

GEN NAKAMURA† AND JENN-NAN WANG‡

Abstract. In this paper we prove the unique continuation property for an elasticity system with small residual stress. The constitutive equation of this elasticity system differs from that of the isotropic elasticity system by $T + (\nabla u)T$, where T is the residual stress tensor. It turns out this elasticity system becomes anisotropic due to the existence of residual stress T. The main technique in the proof is Carleman estimates. Having proved the unique continuation property, we study the inverse problem of identifying the inclusion or cavity.

Key words. unique continuation property, residual stress, probe method

AMS subject classifications. 35B60, 35R30, 74B10

DOI. 10.1137/S003614100139974X

1. Introduction.

Let $\mathcal{B}$ be an isotropic elastic body with residual stress, and let the reference configuration of $\mathcal{B}$ be Ω, a bounded open set in $\mathbb{R}^n$ with smooth boundary. The residual stress is modeled by a symmetric, smooth, second-rank tensor $T(x) = (t_{ij}(x))_{1\le i,j\le n}$ satisfying

$$\partial_{x_j} t_{ij} = 0 \quad \text{in } \Omega, \quad 1 \le i \le n, \tag{1.1}$$

and

$$t_{ij}\nu_j = 0 \quad \text{on } \partial\Omega, \quad 1 \le i \le n, \tag{1.2}$$

where $\nu = (\nu_1, \dots, \nu_n)$ is the unit outer normal to $\partial\Omega$. Hereafter, we adopt the summation convention. Let $u : \Omega \to \mathbb{R}^n$ be the displacement vector; then the first Piola–Kirchhoff stress is written as

$$\begin{aligned}\sigma = {} & T + (\nabla u)T + \lambda(\mathrm{tr}\epsilon)I + 2\mu\epsilon + \beta_1(\mathrm{tr}\epsilon)(\mathrm{tr}T)I + \beta_2(\mathrm{tr}T)\epsilon \\ & + \beta_3((\mathrm{tr}\epsilon)T + \mathrm{tr}(\epsilon T)I) + \beta_4(\epsilon T + T\epsilon),\end{aligned}$$

where λ, μ are the Lamé moduli, $\beta_1, \dots, \beta_4$ are material parameters, and

$$\epsilon = Sym(\nabla u) = \frac{1}{2}(\nabla u + (\nabla u)^t)$$

is the strain tensor [18]. Moreover, we assume that the Lamé moduli satisfy the strong ellipticity condition

$$\mu(x) > \delta > 0, \ \lambda(x) + 2\mu(x) > \delta > 0 \qquad \forall x \in \Omega \tag{1.3}$$

*Received by the editors December 13, 2001; accepted for publication (in revised form) August 6, 2002; published electronically July 18, 2003. This work was done when the authors were participating in the Inverse Problems Program at MSRI and was partially supported by NSF grant DMS-9810361.
http://www.siam.org/journals/sima/35-2/39974.html

†Department of Mathematics, Hokkaido University, Sapporo 060-0810, Japan (gnaka@math.sci.hokudai.ac.jp). This author was partially supported by Grant in Aid for Scientific Research (C) (12640153) of Japan Society for the Promotion of Science.

‡Department of Mathematics, National Taiwan University, Taipei 106, Taiwan (jnwang@math.ntu.edu.tw). This author was partially supported by the National Science Council of Taiwan.

and

$$\beta_3 = \beta_4 = 0,$$

i.e.,

$$\sigma = T + (\nabla u)T + \tilde{\lambda}(\mathrm{tr}\epsilon)I + 2\tilde{\mu}\epsilon, \tag{1.4}$$

where

$$\tilde{\lambda} = \lambda + \beta_1(\mathrm{tr}T), \quad \tilde{\mu} = \mu + \frac{1}{2}\beta_2(\mathrm{tr}T).$$

With the constitutive equation (1.4), the elasticity system considered here is quite close to the one studied by Robertson [22]. Hoger [8] also considered an elasticity system with residual stress where she used the constitutive equation

$$\sigma = T + (\nabla u)T - \frac{1}{2}(\epsilon T + T\epsilon) + \tilde{\lambda}(\mathrm{tr}\epsilon)I + 2\tilde{\mu}\epsilon$$

in her study.

Now the stationary elasticity system is expressed as

$$(Lu)_i = (\nabla \cdot \sigma)_i + \omega^2\rho(x)u_i = \partial_j\sigma_{ij} + \omega^2\rho(x)u_i = 0 \quad \text{in } \Omega,\ 1 \le i \le n,\ \omega \in \mathbb{R}, \tag{1.5}$$

where $\rho(x) > 0$ is the density of the medium. In another setting, if we define the elasticity tensor $\mathbb{C}$ with components

$$C_{ijkl} = \tilde{\lambda}\delta_{ij}\delta_{kl} + (\tilde{\mu}\delta_{jl} + t_{jl})\delta_{ik} + \tilde{\mu}\delta_{il}\delta_{jk} \tag{1.6}$$

and denote

$$(\mathbb{C}E)_{ij} = C_{ijkl}E_{kl} \quad \text{for any matrix} \quad E,$$

then (1.5) is equivalent to

$$(Lu)_i = (\nabla \cdot \mathbb{C}\nabla u)_i + \omega^2\rho u_i = \partial_j(C_{ijkl}\partial_l u_k) + \omega^2\rho u_i = 0 \quad \text{in } \Omega,\ 1 \le i \le n.$$

It is clear to see that (1.5) is an anisotropic elasticity system. In this paper, we will investigate the (weak) unique continuation property (UCP) for the system (1.5); i.e., if $u \in H^2_{\mathrm{loc}}(\Omega)$ is a solution to (1.5) in Ω and vanishes in a nonempty open subset of Ω, then u vanishes identically in Ω.

The UCP for differential equations has a long history. Many deep results about scalar elliptic equations or elliptic systems have been established. We refer the reader to [3] and references therein for details. Recently, few attempts have been made at studying the UCP for systems of equations in mathematical physics such as the Dirac equations and the Maxwell equations [4], [15], [20], [23], [24]. Here we mention two interesting articles [24] and [20] in which Vogelsang and Ōkaji, respectively, proved the strong UCP for the Maxwell system with anisotropic coefficients. In this paper we pay attention to the elasticity system. Several results of weak continuation property for the inhomogeneous isotropic elasticity have been obtained in [1], [5] (stationary) and [6], [14] (nonstationary). Moreover, a strong UCP was recently proven by Alessandrini and Morassi [2]. Unlike the isotropic case, the UCP for the inhomogeneous anisotropic elasticity has not been fully explored.

Our study of the UCP for the inhomogeneous anisotropic elasticity is motivated by its application to inverse problems. It was first recognized by Lax [17] that the Runge approximation property is a consequence of the weak UCP. The Runge approximation property is shown to be a useful technique in dealing with some inverse

problems, especially the inverse problem of recovering inclusions or cavities (see [13], [9], [10], [11], [12], [16], and references therein). It should be noted that the Runge approximation property with constraint for the anisotropic elasticity were proved in [11] and [12]. However, the elasticity tensor there is assumed to be either homogeneous or real-analytic. The weak UCP is an obvious fact in these two situations.

To prove the UCP for the general inhomogeneous anisotropic elasticity is very challenging and difficult. Here we want to consider the system (1.5) which has the simplest form of anisotropy. It turns out we are able to establish the UCP for (1.5), provided the residual stress is sufficiently small. Our main idea comes from Weck's recent article [25], where he proved the UCP for the isotropic elasticity system with zeroth or first order perturbations which contains the results previous obtained by [1], [5]. Weck actually proved something more, namely, he established the UCP for a rather general system of second order differential inequalities with the Laplacian principal part. Like much of the literature on the UCP, the key step in [25] is to prove appropriate Carleman estimates. Here we will adopt Weck's approach to (1.5) with small residual stress, but we have to work a little harder to derive the desired Carleman estimates because we need to deal with variable coefficients second order principal parts due to the presence of residual stress. As indicated previously, having established the UCP, we can prove the Runge approximation property for (1.5) with constraints on Dirichlet data. With this tool at hand, we can solve the inverse problem of identifying inclusions or cavities inside an elastic body with small residual stress by the localized Dirichlet-to-Neumann map using the methods in [11] and [12].

This paper is organized as follows. In section 2, we state and prove the UCP for (1.5) with small residual stress based on suitable Carleman estimates. The derivation of these Carleman estimates is given in section 3. In section 4, we will discuss the application of UCP for (1.5) to the aforementioned inverse problem. In the paper, C stands for a generic constant, and its value may vary from line to line.

2. Unique continuation. To begin, let us denote $v_i = u_i$ for $1 \le i \le n$ and $v_{n+1} = \partial_i u_i$. Then, it follows from (1.5) that

$$
\begin{aligned}
0 &= (Lu)_i \\
&= (\tilde{\mu}\Delta + t_{kj}\partial_j\partial_k)v_i + (\tilde{\lambda}+\tilde{\mu})\partial_i v_{n+1} + (\partial_j t_{kj})\partial_k v_i + (\partial_i\tilde{\lambda})v_{n+1} \\
&\quad + (\partial_j\tilde{\mu})(\partial_i v_j + \partial_j v_i) + \omega^2\rho v_i \\
&= (\tilde{\mu}\Delta + t_{kj}\partial_j\partial_k)v_i + R_i^{(1)}(v_1,\dots,v_n,v_{n+1}) \quad \text{in } \Omega,\ 1 \le i \le n,
\end{aligned} \tag{2.1}
$$

where $R_i^{(1)}$'s are some first order differential operators. Next, by taking the divergence of (1.5), we obtain that

$$
\begin{aligned}
0 &= \partial_i(Lu)_i \\
&= ((\tilde{\lambda}+2\tilde{\mu})\Delta + t_{kj}\partial_j\partial_k)v_{n+1} + 2(\partial_i\tilde{\mu})\Delta v_i + (\partial_i t_{kj})\partial_j\partial_k v_i + 2\partial_i(\tilde{\lambda}+\tilde{\mu})\partial_i v_{n+1} \\
&\quad + (\partial_j t_{kj})\partial_k v_{n+1} + (\partial_i\partial_j t_{kj})\partial_k v_i + (\Delta\tilde{\lambda})v_{n+1} + (\partial_i\partial_j\tilde{\mu})(\partial_i v_j + \partial_j v_i) \\
&\quad + \omega^2(\partial_i\rho)v_i + \omega^2\rho v_{n+1} \\
&= ((\tilde{\lambda}+2\tilde{\mu})\Delta + t_{kj}\partial_j\partial_k)v_{n+1} + R^{(2)}(v_1,\dots,v_n) + R_{n+1}^{(1)}(v_1,\dots,v_{n+1}),
\end{aligned} \tag{2.2}
$$

where $R^{(2)}$ is a pure second order differential operator and $R^{(1)}_{n+1}$ is a first order differential operator, respectively. It should be mentioned that $R^{(2)}$ acts only on $v_1, \dots, v_n$. In view of (1.3), we can see that if

$$\max_{kj} \|t_{kj}\|_{L^\infty(\Omega)} < \varepsilon \tag{2.3}$$

with $\varepsilon \ll 1$, then

$$\tilde{\mu} > \delta' > 0 \quad \text{and} \quad \tilde{\lambda} + 2\tilde{\mu} > \delta' > 0 \quad \forall x \in \Omega.$$

With (2.1) and (2.2) in mind, motivated by Weck's paper [25], we will prove the UCP for the following system of differential inequalities:

$$\begin{aligned} &|A_1(x,\partial)u^1| \le CQ(u^1,u^2)^{1/2}, \\ &|A_2(x,\partial)u^2| \le C\left\{\sum_{ijk}|\partial_i\partial_j u^1_k| + Q(u^1,u^2)^{1/2}\right\}, \end{aligned} \tag{2.4}$$

where $u^l : \Omega \to \mathbb{R}^{m_l}, m_l \in \mathbb{Z}_+$(positive integers) and $A_l(x,\partial) = a^l_{ij}\partial_i\partial_j$ with real symmetric matrix (a^l_{ij}), $l = 1, 2$, and $Q(u^1,u^2) = \sum_{ikl}(|\partial_i u^l_k|^2 + |u^l_k|^2)$.

THEOREM 2.1. *Let $a^l_{ij} \in W^{1,\infty}(\Omega)$ and $(u^1,u^2) \in H^2_{loc}(\Omega)\times H^2_{loc}(\Omega)$ satisfy* (2.4). *Then there exists an $\varepsilon > 0$ such that if*

$$\max_{ij} \|a^l_{ij}(x) - \delta_{ij}\|_{L^\infty(\Omega)} < \varepsilon, \tag{2.5}$$

then (u^1,u^2) vanishes identically in Ω if it vanishes in a nonempty open subset of Ω.

Theorem 2.1 immediately implies the UCP for (1.5) with small residual stress.

COROLLARY 2.2. *Let coefficients $\lambda, \mu, \beta_1, \beta_2, t_{kj}$ belong to $W^{2,\infty}(\Omega)$, and let ρ be in $W^{1,\infty}(\Omega)$. Then there exists an $\varepsilon > 0$ such that if* (2.3) *is satisfied with this ε, then the system* (1.5) *possesses the UCP.*

The proof of Theorem 2.1 relies on the following Carleman estimates.

PROPOSITION 2.3. *Assume that the differential operators A_1 and A_2 satisfy the assumptions in Theorem* 2.1. *Let $r_0 < 1$ and $U_{r_0} = \{u \in C^\infty_0(\mathbb{R}^n \setminus \{0\}) : \operatorname{supp}(u) \subset B_{r_0}\}$, where B_{r_0} is the ball centered at the origin with radius r_0. Then there exist positive constants β_0 and ε_0 such that if* (2.5) *is satisfied with $\varepsilon \le \varepsilon_0$, then for all $\beta \ge \beta_0$ and $u \in U_{r_0}$ we have that*

$$\int r^{-\sigma}\psi^2 \sum_{ij}|\partial_i\partial_j u|^2 dx \le C\int r^{-\sigma}\psi^2(\beta^2 r^{-2\beta-2}|\nabla u|^2 + |A_l u|^2)dx \tag{2.6}$$

and

$$\beta^2\int r^{-\sigma-\beta-1}\psi^2(|\nabla u|^2 + |u|^2)dx \le C\int r^{-\sigma}\psi^2|A_l u|^2 dx \tag{2.7}$$

for $l = 1, 2$, where $r = |x|$, $\psi = \exp(r^{-\beta})$, and $\sigma = \sigma_0 + c\beta$ with $\sigma_0, c \in \mathbb{R}$.

The proof of Proposition 2.3 is postponed until the next section. Here we want to prove Theorem 2.1 based on this proposition.

Proof of Theorem 2.1. It suffices to prove the theorem for the case $m_1 = m_2 = 1$. Let (u^1,u^2) vanish in a neighborhood of $x_0 \in \Omega$. Without loss of generality, we assume

$x_0 = 0$. We set $\tilde{r} = \min\{1/2, \mathrm{dist}(0, \partial\Omega)\}$. Now let $\chi \in C_0^\infty(\mathbb{R}^n)$ be a cut-off function satisfying $0 \le \chi \le 1$, $\chi|_{B_{\tilde{r}/2}} = 1$, and $\mathrm{supp}(\chi) \subset B_{\tilde{r}}$. Denote $v_l = \chi u^l$, $l = 1, 2$. From (2.4) we have that

$$
\begin{aligned}
&|A_1 v_1| \le C(e(v_1) + e(v_2))^{1/2} + f_1, \\
&|A_2 v_2| \le C\left[\sum_{ij} |\partial_i \partial_j v_1| + (e(v_1) + e(v_2))^{1/2}\right] + f_2,
\end{aligned} \tag{2.8}
$$

where $e(v) = |\nabla v|^2 + |v|^2$ and f_l is supported in $B_{\tilde{r}} \setminus B_{\tilde{r}/2}$ for $l = 1, 2$. It follows from (2.8) that

(2.9)
$$
I := \gamma \int r^{-\beta} \psi^2 |A_1 v_1|^2 dx + \int r \psi^2 |A_2 v_2|^2 dx \le C\left(F + G + \int r\psi^2 \sum_{ij} |\partial_i \partial_j v_1|^2 dx\right),
$$

where

$$
\begin{aligned}
F &= \gamma \int r^{-\beta} \psi^2 f_1^2 dx + \int r \psi^2 f_2^2 dx, \\
G &= \int (r + \gamma r^{-\beta}) \psi^2 (e(v_1) + e(v_2)) dx.
\end{aligned}
$$

Here γ is a large positive parameter which will be chosen later on. By the standard approximation argument, we can see that v_1 and v_2 satisfy estimates (2.6) and (2.7). Taking $\sigma = -1$ in the estimate (2.6) for $l = 1$ and substituting it into (2.9) yield

$$
I \le C\left(F + G + \int r\psi^2 |A_1 v_1|^2 dx + \beta^2 \int r^{-2\beta - 1} \psi^2 |\nabla v_1|^2 dx\right). \tag{2.10}
$$

Replacing the last term of (2.10) with the help of (2.7) for $\sigma = \beta$ and $l = 1$, we obtain that

$$
I \le C\left(F + G + \int r^{-\beta} \psi^2 |A_1 v_1|^2 dx\right). \tag{2.11}
$$

Now taking γ sufficiently large, we can absorb the last term of (2.11) and get

$$
I \le C(F + G). \tag{2.12}
$$

From now on we fix the parameter γ.

Next using $\sigma = \beta$ in (2.7) for $l = 1$ and $\sigma = -1$ in (2.7) for $l = 2$, we find that

$$
\begin{aligned}
H &:= \beta^2 \int r^{-2\beta - 1} \psi^2 e(v_1) dx + \beta^2 \int r^{-\beta} \psi^2 e(v_2) dx \\
&\le C\left(\int r^{-\beta} \psi^2 |A_1 v_1|^2 dx + \int r \psi^2 |A_2 v_2|^2 dx\right).
\end{aligned} \tag{2.13}
$$

Combining (2.12) and (2.13) gives

$$
H \le C(F + G) \le C\left(F + \int (r + \gamma r^{-\beta}) \psi^2 (e(v_1) + e(v_2)) dx\right). \tag{2.14}
$$

Now observing that $r < r^{-\beta} < \beta r^{-\beta} < \beta r^{-2\beta-1}$ when $r \le \tilde{r}$ and $\beta > 1$, we obtain from (2.14) that

$$H \le C\left(F + \beta \int r^{-2\beta-1}\psi^2 e(v_1)dx + \beta \int r^{-\beta}\psi^2 e(v_2)dx\right). \tag{2.15}$$

Taking β sufficiently large in (2.15), we get that

$$H \le CF,$$

i.e.,

$$\beta^2 \int r^{-2\beta-1}\psi^2 e(v_1)dx + \beta^2 \int r^{-\beta}\psi^2 e(v_2)dx \le C\left(\int r^{-\beta}\psi^2 f_1^2 dx + \int r\psi^2 f_2^2 dx\right),$$

from which we immediately have

$$\beta^2 \int_{B_{\tilde{r}/2}} r^{-\beta}\psi^2(v_1^2 + v_2^2)dx \le C \int_{B_{\tilde{r}}\setminus B_{\tilde{r}/2}} r^{-\beta}\psi^2(f_1^2 + f_2^2)dx. \tag{2.16}$$

Since $r^{-\beta}\psi^2$ is a strictly decreasing function, (2.16) implies that

$$\beta^2 \int_{B_{\tilde{r}/2}} (v_1^2 + v_2^2)dx \le C \int_{B_{\tilde{r}}\setminus B_{\tilde{r}/2}} (f_1^2 + f_2^2)dx,$$

and therefore $(v_1, v_2) = 0$ on $B_{\tilde{r}/2}$ if we choose β sufficiently large. Clearly, (u^1, u^2) must be zero throughout Ω. □

3. Proof of Carleman estimates. This section is devoted to the proof of Proposition 2.3. It suffices to prove (2.6) and (2.7) for A_1. Therefore, we denote $a_{ij}^1 = a_{ij}$ and $A_1 = A$. To prove (2.6), we first recall the following estimate in [25]:

$$\int r^{-\sigma}\psi^2 \sum_{ij} |\partial_i\partial_j u|^2 dx \le C \int r^{-\sigma}\psi^2(\beta^2 r^{-2\beta-2}|\nabla u|^2 + |\Delta u|^2)dx$$

(see [25, Lemma 2]), from which we have that

$$\begin{aligned}\int r^{-\sigma}\psi^2 \sum_{ij} |\partial_i\partial_j u|^2 dx &\le C \int r^{-\sigma}\psi^2(\beta^2 r^{-2\beta-2}|\nabla u|^2 + |Au|^2 + |\Delta u - Au|^2)dx \\ &\le C \int r^{-\sigma}\psi^2\left(\beta^2 r^{-2\beta-2}|\nabla u|^2 + |Au|^2 + \varepsilon^2 \sum_{ij} |\partial_i\partial_j u|^2\right)dx.\end{aligned}$$

Thus, choosing ε small enough immediately implies the estimate (2.6).

The proof of (2.7) is lengthy. Here we will adopt some techniques from [21], [25], and [26]. Let $\phi = \psi^{-1}$ and $u = r^{\tau/2}\phi z$. Then

$$\begin{aligned} r^{-\sigma/2}\psi Au &= r^{-\sigma/2}\psi A(r^{\tau/2}\phi z) \\ &= r^{-\sigma/2}\psi[r^{\tau/2}\phi Az + 2a_{ij}\partial_i z\partial_j(r^{\tau/2}\phi) + zA(r^{\tau/2}\phi)].\end{aligned}$$

By virtue of the inequality $(a+b+c)^2 \ge 2ab + 2bc$, we have that

$$\begin{aligned}\int r^{-\sigma}\psi^2|Au|^2 dx &\ge 4\int r^{-\sigma}\psi^2 a_{ij}\partial_i z\partial_j(r^{\tau/2}\phi)r^{\tau/2}\phi Az dx \\ &\quad + 4\int r^{-\sigma}\psi^2 a_{ij}\partial_i z\partial_j(r^{\tau/2}\phi)zA(r^{\tau/2}\phi)dx.\end{aligned} \tag{3.1}$$

With the choice of $\tau = \sigma + \beta + 2$, we can compute

$$\begin{aligned} I &:= \int r^{-\sigma}\psi^2 a_{ij}\partial_i z \partial_j (r^{\tau/2}\phi) r^{\tau/2}\phi A z dx \\ &= \beta \int a_{ij}\partial_i z x_j A z dx + \tau/2 \int r^{\beta} a_{ij}\partial_i z x_j A z dx. \end{aligned}$$

It is readily seen that the leading term (for large β) of I is $\beta \int a_{ij}\partial_i z x_j A z dx$. Repeated integration by parts shows that

$$\begin{aligned} 2\int a_{ij}\partial_i z x_j A z dx &= 2\int a_{ij}\partial_i z x_j a_{kl}\partial_k\partial_l z dx \\ &= -\int \partial_i z \partial_l (a_{kl}a_{ij}x_j)\partial_k z dx + \int \partial_k z \partial_i (a_{kl}a_{ij}x_j)\partial_l z dx \\ &\quad - \int \partial_l z \partial_k (a_{kl}a_{ij}x_j)\partial_i z dx. \end{aligned} \tag{3.2}$$

Using (3.2), we obtain that

$$\begin{aligned} |I| &\le C\beta \left| \int \partial_i z \partial_l (a_{kl}a_{ij}x_j)\partial_k z dx \right| \\ &\le C\beta \|\nabla z\|^2 \\ &\le C\beta (\|\nabla (r^{-\tau/2}\psi)u\|^2 + \|r^{-\tau/2}\psi \nabla u\|^2) \\ &\le C\left(\beta^3 \int r^{-\sigma-3\beta-4}\psi^2 |u|^2 dx + \beta \int r^{-\sigma-\beta-2}\psi^2 |\nabla u|^2 dx\right). \end{aligned} \tag{3.3}$$

Next we observe that

$$\begin{aligned} J &:= \int r^{-\sigma}\psi^2 a_{ij}\partial_i z \partial_j (r^{\tau/2}\phi) z A (r^{\tau/2}\phi) dx \\ &= \beta \int r^{-\sigma+\tau/2-\beta-2}\psi a_{ij}\partial_i z x_j z A (r^{\tau/2}\phi) dx \\ &\quad + \tau/2 \int r^{-\sigma+\tau/2-2}\psi a_{ij}\partial_i z x_j z A (r^{\tau/2}\phi) dx. \end{aligned}$$

Straightforward calculations show that

$$\partial_i\partial_j \phi = (\beta^2 x_i x_j r^{-2\beta-4} + \beta\delta_{ij} r^{-\beta-2} - \beta(\beta+2) x_i x_j r^{-\beta-4})\phi$$

and

$$\partial_i\partial_j r^{\tau/2} = (\tau/2)(\tau/2-2) r^{\tau/2-4} x_i x_j + (\tau/2) r^{\tau/2-2}\delta_{ij}.$$

So the leading term of J is

$$\beta^3 \int r^{-2\beta-4} a_{ij}\partial_i z x_j a_{kl} x_k x_l z dx.$$

Note that we have chosen $\tau = \sigma + \beta + 2$. Performing the integration by parts, we can

see that

$$\begin{aligned}
&\beta^3 \int r^{-2\beta-4} a_{ij}\partial_i z x_j a_{kl} x_k x_l z dx \\
&= -\frac{1}{2}\beta^3 \int z\partial_i (r^{-2\beta-4} a_{ij} x_j a_{kl} x_k x_l) z dx \\
&\geq (1-o(\beta))\beta^4 \int r^{-2\beta-6} a_{ij} x_i x_j a_{kl} x_k x_l |z|^2 dx \\
&\geq (1-o(\beta))\beta^4 (1-O(\varepsilon)) \int r^{-2\beta-2} |z|^2 dx \\
&\geq (1-o(\beta))\beta^4 (1-O(\varepsilon)) \int r^{-\sigma-3\beta-4} \psi^2 |u|^2 dx,
\end{aligned}$$

where $0 \leq o(\beta) \to 0$ as $\beta \to \infty$ and $O(\varepsilon)$ is a positive constant bounded by $C\varepsilon$. In other words, we have that

$$J \geq (1-o(\beta))\beta^4 (1-O(\varepsilon)) \int r^{-\sigma-3\beta-4} \psi^2 |u|^2 dx. \tag{3.4}$$

Notice that we need to keep track of the leading constant here in order to obtain the desired estimate. Combining (3.1), (3.3), and (3.4) gives

$$\begin{aligned}
&\int r^{-\sigma}\psi^2 |Au|^2 dx + C\left(\beta^3 \int r^{-\sigma-3\beta-4}\psi^2 |u|^2 dx + \beta \int r^{-\sigma-\beta-2}\psi^2 |\nabla u|^2 dx\right) \\
&\geq 4(1-o(\beta))\beta^4 (1-O(\varepsilon)) \int r^{-\sigma-3\beta-4}\psi^2 |u|^2 dx,
\end{aligned}$$

from which we can derive that

$$\begin{aligned}
&\int r^{-\sigma}\psi^2 |Au|^2 dx + C\beta \int r^{-\sigma-\beta-2}\psi^2 |\nabla u|^2 dx \\
&\qquad \geq 4(1-o(\beta))\beta^4 (1-O(\varepsilon)) \int r^{-\sigma-3\beta-4}\psi^2 |u|^2 dx.
\end{aligned} \tag{3.5}$$

By the ellipticity condition and performing the integration by parts, we can get that

$$\begin{aligned}
&(1-O(\varepsilon)) \int r^{-\sigma-\beta-2}\psi^2 |\nabla u|^2 dx \\
&\leq \int r^{-\sigma-\beta-2}\psi^2 a_{ij}\partial_i u \partial_j u dx \\
&\leq \left| \int u\partial_i (r^{-\sigma-\beta-2}\psi^2) a_{ij}\partial_j u dx \right| + \left| \int r^{-\sigma-\beta-2}\psi^2 u \partial_i (a_{ij})\partial_j u dx \right| \\
&\quad + \left| \int r^{-\sigma-\beta-2}\psi^2 u a_{ij}\partial_i\partial_j u dx \right| \\
&:= K_1 + K_2 + K_3.
\end{aligned} \tag{3.6}$$

Using the relation $|ab| \le (a^2+b^2)/2$, we can estimate

$$
\begin{aligned}
K_1 &= \left| \int u \partial_i (r^{-\sigma-\beta-2}\psi^2) a_{ij} \partial_j u dx \right| \\
&\le \int (2+o(\beta))\beta r^{-\sigma-2\beta-4}\psi^2 |u a_{ij} x_i \partial_j u| dx \\
&\le (2+o(\beta))\beta^2(1+O(\varepsilon)) \int r^{-\sigma-3\beta-4}\psi^2 |u|^2 dx \\
&\quad + (1+O(\varepsilon))/2 \int r^{-\sigma-\beta-2}\psi^2 |\nabla u|^2 dx.
\end{aligned} \tag{3.7}
$$

Likewise, for K_2 and K_3, we have that

$$
K_2 \le C \left(r_0^\beta \int r^{-\sigma-3\beta-4}\psi^2 |u|^2 dx + r_0^{\beta+2} \int r^{-\sigma-\beta-2}\psi^2 |\nabla u|^2 dx \right) \tag{3.8}
$$

and

$$
K_3 \le C \left(r_0^\beta \beta^2 \int r^{-\sigma-3\beta-4}\psi^2 |u|^2 dx + \beta^{-2} \int r^{-\sigma}\psi^2 |Au|^2 dx \right). \tag{3.9}
$$

Plugging (3.7), (3.8), and (3.9) into (3.6) and multiplying the new inequality by β^2, we obtain that

$$
\begin{aligned}
&\beta^2(1-O(\varepsilon)) \int r^{-\sigma-\beta-2}\psi^2 |\nabla u|^2 dx \\
&\le (2+o(\beta))\beta^4(1+O(\varepsilon)) \int r^{-\sigma-3\beta-4}\psi^2 |u|^2 dx + \beta^2(1+O(\varepsilon))/2 \int r^{-\sigma-\beta-2}\psi^2 |\nabla u|^2 dx \\
&\quad + C \left(r_0^\beta \beta^2 \int r^{-\sigma-3\beta-4}\psi^2 |u|^2 dx + r_0^{\beta+2}\beta^2 \int r^{-\sigma-\beta-2}\psi^2 |\nabla u|^2 dx \right) \\
&\quad + C \left(r_0^\beta \beta^4 \int r^{-\sigma-3\beta-4}\psi^2 |u|^2 dx + \int r^{-\sigma}\psi^2 |Au|^2 dx \right).
\end{aligned} \tag{3.10}
$$

Adding (3.10) to (3.5) and taking β sufficiently large and ε small enough, we conclude that

$$
\beta^4 \int r^{-\sigma-3\beta-4}\psi^2 |u|^2 dx + \beta^2 \int r^{-\sigma-\beta-2}\psi^2 |\nabla u|^2 dx \le C \int r^{-\sigma}\psi^2 |Au|^2 dx,
$$

which immediately implies (2.7).

4. Applications to inverse problems. In this section we will discuss the application of the UCP for (1.5) to the inverse problem of identifying inclusions or cavities by boundary measurements. To begin, assume that D is an open subset of Ω with Lipschitz boundary such that $\Omega \setminus \bar{D}$ is connected. The domain D stands forthe region of the inclusion or cavity embedded in Ω. Let the reference elasticity tensor $\mathbb{C}(x)$ with components $C_{ijkl}(x)$ be defined by (1.6), i.e.,

$$
C_{ijkl} = \tilde{\lambda}\delta_{ij}\delta_{kl} + (\tilde{\mu}\delta_{jl} + t_{jl})\delta_{ik} + \tilde{\mu}\delta_{il}\delta_{jk},
$$

where $\tilde{\lambda} = \lambda + \beta_1(\mathrm{tr}T)$ and $\tilde{\mu} = \mu + (1/2)\beta_2(\mathrm{tr}T)$. Here we require that the Lamé moduli satisfy the strong convexity condition

$$
\mu(x) > \delta > 0 \quad \text{and} \quad n\lambda(x) + 2\mu(x) > \delta > 0 \quad \forall\, x \in \Omega \tag{4.1}
$$

and T satisfies

$$ET \cdot E \geq (\varepsilon/2)|E|^2,$$

which is equivalent to

$$\mathbb{C}(x)E \cdot E \geq \kappa E_{ij}E_{ij} = \kappa|E|^2, \quad \kappa(\varepsilon) > 0, \ \forall\ x \in \Omega \tag{4.2}$$

for all matrices E, provided that ε in (2.3) is sufficiently small. It is obvious that (4.1) implies (1.3). Next we assume that $\tilde{\mathbb{C}}$ is some fourth-rank tensor such that $\mathbb{C} + \chi_D\tilde{\mathbb{C}}$ satisfies the strong convexity condition (4.2), where χ_D denotes the characteristic function of D. Moreover, suppose that $\tilde{\mathbb{C}}$ satisfies the jump condition

(4.3)
$$\forall\ x \in \partial D, \ \exists\ C_x > 0, \ \exists\ \delta_x > 0 \text{ such that } \tilde{\mathbb{C}}(y)E \cdot E \geq C_x|E|^2 \text{ or } \tilde{\mathbb{C}}(y)E \cdot E \leq -C_x|E|^2$$

for almost all $y \in B_{\delta_x}(x) \cap D$ and all real matrices E. Let all components of $\mathbb{C}(x)$ and $\tilde{\mathbb{C}}(x)$ be in $L^\infty(\Omega)$. Then it is easy to show that there exists a unique solution $u \in H^1(\Omega)$ to

$$\begin{cases} \nabla \cdot ((\mathbb{C} + \chi_D\tilde{\mathbb{C}})\nabla u) = 0 & \text{in } \Omega, \\ u = f & \text{on } \partial\Omega \end{cases}$$

for any $f \in H^{1/2}(\partial\Omega)$. In this case, the domain D is an inclusion. So we can define the Dirichlet-to-Neumann (displacement-to-traction) map $\Lambda_I : H^{1/2}(\partial\Omega) \to H^{-1/2}(\partial\Omega)$ by

$$\Lambda_I(f) = (\mathbb{C}\nabla u)\nu|_{\partial\Omega}.$$

Equivalently, Λ_I can be defined by the formula

$$\langle \Lambda_I(f), g\rangle = \int_\Omega (\mathbb{C} + \chi_D\tilde{\mathbb{C}})\nabla u \cdot \nabla v dx,$$

where $v \in H^1(\Omega)$ with $v|_{\partial\Omega} = g$. We are interested in the following inverse problem:

IP.A. *Reconstruct the inclusion D from the knowledge of $\Lambda_I(f)_{\Gamma_0}$ for infinitely many $f \in H^{1/2}(\partial\Omega)$ with $\operatorname{supp}(f) \subset \Gamma_0$, where Γ_0 is a nonempty subset of $\partial\Omega$.*

Likewise, in the extreme case, if the tensor $\tilde{\mathbb{C}}$ becomes $-\mathbb{C}$, then the domain D corresponds to a cavity. In the same way, we can prove that there exists a unique solution $u \in H^1(\Omega \setminus \bar{D})$ to the boundary value problem

$$\begin{cases} \nabla \cdot (\mathbb{C}\nabla u) = 0 & \text{in } \Omega \setminus \bar{D}, \\ (\mathbb{C}\nabla u)\nu = 0 & \text{on } \partial D, \quad (\mathbb{C}\nabla u)\nu = g \quad \text{on } \partial\Omega \end{cases}$$

for any $g \in H^{1/2}(\partial\Omega)$. Therefore, we can define the Dirichlet-to-Neumann map $\Lambda_C : H^{1/2}(\partial\Omega) \to H^{-1/2}(\partial\Omega)$ by

$$\Lambda_C(g) = (\mathbb{C}\nabla u)\nu|_{\partial\Omega}.$$

Similarly, we will consider the following inverse problem:

IP.B. *Reconstruct the cavity D from the knowledge of $\Lambda_C(g)|_{\Gamma_0}$ for infinitely many g with $\operatorname{supp}(g) \subset \Gamma_0$.*

Note that uniqueness theorems of determining the inclusion or cavity embedded in an elastic body have been established in [11] and [12], where the reference medium is assumed to be either inhomogeneous isotropic or anisotropic with homogeneous or analytic elasticity tensors. Besides, a reconstruction algorithm for recovering the cavity is given in [12]. A similar algorithm can be developed for the inclusion case. Here we want to extend their results to the elasticity system with residual stress (1.5). To this end, we will need the Runge approximation property with constraint for (1.5), which is a consequence of the UCP (see Corollary 2.2). Its proof can be found in [12].

PROPOSITION 4.1. *Assume that all coefficients of* $\mathbb{C}$ *are in* $W^{2,\infty}(\Omega)$ *and the residual stress satisfies* (2.3) *with* ε *given in Corollary* 2.2. *Let* U *and* Ω *be two open bounded domains with Lipschitz and* C^2 *boundaries, respectively, such that* $\bar{U} \subset \Omega$. *Denote* Γ_0 *a subset of the boundary* $\partial\Omega$. *Let* $u \in H^1(U)$ *satisfy*

$$\nabla \cdot (\mathbb{C}\nabla u) = 0 \quad \text{in} \quad U.$$

Then for any compact subset $K \subset U$ *such that* $\Omega \setminus K$ *is connected and any* $\tilde{\varepsilon} > 0$ *there exists* $w \in H^1(\Omega)$ *satisfying*

$$\nabla \cdot (\mathbb{C}\nabla w) = 0 \quad \text{in} \quad \Omega$$

with $\operatorname{supp}(w|_{\partial\Omega}) \subset \Gamma_0$ *such that*

$$\|w - u\|_{H^1(K)} < \tilde{\varepsilon}.$$

Remark. The reason for using C^2 boundary on Ω is that we want to extend all coefficients of $\mathbb{C}$ into a larger domain $\tilde{\Omega}$ and the newly extended coefficients have the same regularity $W^{2,\infty}$ in $\tilde{\Omega}$.

Having the Runge approximation property Proposition 4.1 at hand, we now can apply the methods in [11] and [12] to solve IP.A and IP.B. It should be pointed out that the reference elasticity tensor in [11] and [12] satisfies the full symmetry properties, i.e.,

$$C_{ijkl} = C_{klij} = C_{jikl}.$$

Nevertheless, it is not hard to check that the proofs in [11] and [12] are still valid if we only assume $C_{ijkl} = C_{klij}$, which is the case for the elasticity system with residual stress (1.5). For IP.A, we prove the following theorem (see [11]).

THEOREM 4.2 (identification of inclusion). *Let the domain* Ω *have* C^2 *boundary. Assume that the elasticity tensor* $\mathbb{C}$ *given by* (1.6) *possesses* $W^{2,\infty}(\Omega)$ *coefficients satisfying* (4.1). *Furthermore, the residual stress tensor* T *in* $\mathbb{C}$ *satisfies the smallness condition described in Corollary* 2.2. *Let* $(D_1, \tilde{\mathbb{C}}_1)$ *and* $(D_2, \tilde{\mathbb{C}}_2)$ *be two inclusions such that* $\mathbb{C} + \chi_{D_i}\tilde{\mathbb{C}}_i$ *and* $\tilde{\mathbb{C}}_i$ *satisfy* (4.2) *and* (4.3), *respectively, and* $\Omega \setminus \bar{D}_i$ *is connected,* $i = 1, 2$. *If*

$$\Lambda_{I_1}(f) = \Lambda_{I_2}(f) \quad \text{on} \quad \Gamma_0$$

for all $f \in H^{1/2}(\partial\Omega)$ *with* $\operatorname{supp}(f) \subset \Gamma_0$, *then*

$$D_1 = D_2.$$

The proof of Theorem 4.2 is based on integral inequalities

$$\int_D \{\mathbb{C}^{-1} - (\mathbb{C} + \tilde{\mathbb{C}})^{-1}\}\mathbb{C}\nabla w \cdot \mathbb{C}\nabla w dx \le \langle (\Lambda_I - \Lambda_0)f, f\rangle \le \int_D \tilde{\mathbb{C}}\nabla w \cdot \nabla w dx, \tag{4.4}$$

where $w \in H^1(\Omega)$ solves

$$\begin{cases} \nabla \cdot (\mathbb{C}\nabla w) = 0 \quad \text{in} \quad \Omega, \\ w|_{\partial\Omega} = f. \end{cases} \tag{4.5}$$

Here $\mathbb{C}^{-1}$ (or $(\mathbb{C}+\tilde{\mathbb{C}})^{-1}$) is called the compliance tensor (see, e.g., [7]). Notice that we do not assume $\tilde{\mathbb{C}}_1 = \tilde{\mathbb{C}}_2$ in Theorem 4.2. Also, the regularity of the medium inside of the inclusions is only assumed to be essentially bounded. Theorem 4.2 provides the uniqueness of determining the inclusion embedded in an elastic body with small residual stress by the localized Dirichlet-to-Neumann map. For the sake of completeness, we want to briefly describe a reconstruction algorithm for identifying the inclusion. Let $y \in \Omega$ and $G_0(\cdot; y)$ be the fundamental solution for the operator $\nabla \cdot \mathbb{C}(y)\nabla$ (see, e.g., [19]). One can find $e(\cdot; y)$ such that

$$\nabla \cdot (\mathbb{C}(x)\nabla e(\cdot; y)) = 0 \quad \text{in} \quad \Omega \setminus \{y\}$$

and

$$(e(\cdot; y) - G_0(\cdot - y; y)b)_{y\in\Omega} \quad \text{is bounded in } H^1(\Omega),$$

where b is a nonzero constant vector. Note that if $y \in \partial D$, then

$$\int_{D\cap B_r(y)} |\nabla\{G_0(x-y;y)b\}|^2 dx = \infty \tag{4.6}$$

for any ball $B_r(y)$ centered at y with radius r and nonzero vector b. The symmetric version of (4.6) has been proved in [11], i.e.,

$$\int_{D\cap B_r(y)} |Sym\nabla\{G_0(x-y;y)b\}|^2 dx = \infty,$$

which clearly implies (4.6).

A continuous map $c : [0,1] \to \bar{\Omega}$ is called a *needle* if it satisfies (i) $c(0), c(1) \in \partial\Omega$; (ii) $c(t) \in \Omega$ for $0 < t < 1$. In view of Proposition 4.1, we can see that for each needle and $t \in (0,1)$, there exists a sequence $\{f_j\} = \{f_j(\cdot; c(t))\}$ in $H^{1/2}(\partial\Omega)$ with $\text{supp}(f_j) \subset \Gamma_0$ such that the solution w_j of (4.5) with $f = f_j$ satisfies $w_j \to e(\cdot; c(t))$ in $H^1_{loc}(\Omega \setminus \{c(t') : 0 < t' \le t\})$ as $j \to \infty$. We call $\{f_j\}$ a fundamental sequence with respect to Γ_0. For each needle c, define

$$t(c) = \sup\{0 < s < 1 : C(t) \in \Omega \setminus \bar{D} \ (0 < t < s)\}.$$

It should be noted that $0 < t(c) \le 1$, and if $t(c) = 1$, then c never touches ∂D. On the other hand, if $t(c) < 1$, then c touches ∂D at $t = t(c)$ at the first time. Since $\Omega \setminus \bar{D}$ is connected, we have that

$$\partial D = \{c(t(c)) : c \text{ is a needle and } t(c) < 1\}. \tag{4.7}$$

Let Λ_0 be the Dirichlet-to-Neumann map associated with the boundary value problem (4.5). Denote

$$\mathcal{I}_I(t,c) = \lim_{j\to\infty} \langle (\Lambda_I - \Lambda_0) f_j(\cdot; c(t)), f_j(\cdot; c(t)) \rangle$$

and

$$\mathcal{T}_I(c) = \left\{ 0 < s < 1 : \mathcal{I}_I \text{ exists } \forall\ 0 < t < s \text{ and } \sup_{0<t<s} |\mathcal{I}_I(t,c)| < \infty \right\}.$$

Using (4.3), (4.4), and (4.6) and pursuing the arguments in [11], we can show that $\mathcal{T}_I(c) = (0, t(c))$, and therefore $t(c) = \sup \mathcal{T}_I(c)$ (see similar arguments in [12]). In summary, we have a reconstruction algorithm for determining the inclusion as follows.

Reconstruction algorithm for IP.A.

(i) *For each needle c and each $t \in (0,1)$, find the fundamental sequence $\{f_j(\cdot; c(t))\}$ with respect to Γ_0.*

(ii) *Compute $\mathcal{T}_I(c)$ and set $t(c) = \sup \mathcal{T}_I(c)$.*

(iii) *Use the formula* (4.7) *to reconstruct ∂D.*

Now for IP.B, we show the following (see [12]).

Theorem 4.3 (identification of cavity). *Let the assumptions in Theorem* 4.2 *on Ω and $\mathbb{C}$ hold. Assume that D_1 and D_2 are two cavities and $\Omega \setminus \bar{D}_1$ and $\Omega \setminus \bar{D}_2$ are connected. Let*

$$\Lambda_{C_1}(f) = \Lambda_{C_2}(f) \quad \text{on} \quad \Gamma_0$$

for all $f \in H^{1/2}(\partial\Omega)$ with $\operatorname{supp}(f) \subset \Gamma_0$. Then $D_1 = D_2$.

As for reconstructing the cavity, we follow the lines of the above algorithm and define

$$\mathcal{I}_C(t,c) = \lim_{j\to\infty} \langle (\Lambda_0 - \Lambda_C) f_j(\cdot; c(t)), f_j(\cdot; c(t)) \rangle$$

and

$$\mathcal{T}_C(c) = \left\{ 0 < s < 1 : \mathcal{I}_C \text{ exists } \forall\ 0 < t < s \text{ and } \sup_{0<t<s} \mathcal{I}_C(t,c) < \infty \right\}.$$

Note that $\langle (\Lambda_0 - \Lambda_C) f, f \rangle \geq 0$ for all $f \in H^{1/2}(\partial\Omega)$. Now using (4.6) and the inequalities

$$\frac{1}{M} \int_D |\nabla e(x; c(t))|^2 dx \leq \mathcal{I}_C(t,c) \leq M \int_D |\nabla e(x; c(t))|^2 dx$$

for some constant $M > 0$, one can prove that $\mathcal{T}_C(c) = (0, t(c))$ and thus $t(c) = \sup \mathcal{T}_C(c)$ (see the arguments in [12]). So a reconstruction algorithm for identifying the cavity is described as follows.

Reconstruction algorithm for IP.B.

(i) *For each needle c and each $t \in (0,1)$, find the fundamental sequence $\{f_j(\cdot; c(t))\}$ with respect to Γ_0.*

(ii) *Compute $\mathcal{T}_C(c)$ and set $t(c) = \sup \mathcal{T}_C(c)$.*

(iii) *Use the formula* (4.7) *to reconstruct ∂D.*

Acknowledgment. The authors would like to thank MSRI for providing a very stimulating research environment and partial financial support.

REFERENCES

[1] D. D. Ang, M. Ikehata, D. D. Trong, and M. Yamamoto, *Unique continuation for a stationary isotropic Lamé system with variable coefficients*, Comm. Partial Differential Equations, 23 (1998), pp. 371–385.

[2] G. Alessandrini and A. Morassi, *Strong unique continuation for the Lamé system of elasticity*, Comm. Partial Differential Equations, 26 (2001), pp. 1787–1810.
[3] L. Bers, F. John, and M. Schechter, *Partial Differential Equation*, John Wiley, New York, 1964.
[4] L. De Carli and T. Ōkaji, *Strong unique continuation property for the Dirac equation*, Publ. Res. Inst. Math. Sci., 35 (1999), pp. 825–846.
[5] B. Dehman and L. Rabbiano, *La propriété du prolongement unique pour un système elliptique: Le système de Lamé*, J. Math. Pures Appl., 72 (1993), pp. 475–492.
[6] M. Eller, V. Isakov, G. Nakamura, and D. Tataru, *Uniqueness and stability in the Cauchy problem for Maxwell and elasticity systems*, in Nonlinear Partial Differential Equations and Applications, Collége de France Seminar, Vol. 14, D. Cioranescu and J.-L. Lions, eds., Stud. Math. Appl. 31, North-Holland, Amsterdam, 2002, pp. 329–349.
[7] M. Gurtin, *The Linear Theory of Elasticity*, Mechanics of Solids, Vol. II, C. Thruesdell, ed., Springer-Verlag, Berlin, 1984.
[8] A. Hoger, *On the determination of residual stress in an elastic body*, J. Elasticity, 16 (1986), pp. 303–324.
[9] M. Ikehata, *Identification of the shape of the inclusion having essentially bounded conductivity*, J. Inverse Ill-Posed Probl., 7 (1999), pp. 533–540.
[10] M. Ikehata, *Reconstruction of the shape of the inclusion by boundary measurements*, Comm. Partial Differential Equations, 23 (1998), pp. 1459–1474.
[11] M. Ikehata, G. Nakamura, and K. Tanuma, *Identification of the shape of the inclusion in the anisotropic elastic body*, Appl. Anal., 72 (1999), pp. 17–26.
[12] M. Ikehata and G. Nakamura, *Reconstruction of Cavity from Boundary Measurements*, preprint.
[13] V. Isakov, *On uniqueness of recovery of a discontinuous conductivity coefficients*, Comm. Pure Appl. Math., 41 (1988), pp. 865–877.
[14] V. Isakov, *A non-hyperbolic Cauchy problem for $\Box_b\Box_c$ and its applications to elasticity theory*, Comm. Pure Appl. Math., 39 (1986), pp. 747–767.
[15] D. Jerison, *Carleman inequalities for the Dirac and Laplace operator and unique continuation*, Adv. in Math., 63 (1986), pp. 118–134.
[16] R. Kohn and M. Vogelius, *Determining conductivity by boundary measurements* II. *Interior results*, Comm. Pure Appl. Math., 38 (1985), pp. 643–667.
[17] P. Lax, *A stability theorem for solutions of abstract differential equations and its application to the study of local behavior of solutions of elliptic equations*, Comm. Pure Appl. Math., 9 (1956), pp. 747–766.
[18] C. S. Man, *Hartig's law and linear elasticity with initial stress*, Inverse Problems, 14 (1998), pp. 313–319.
[19] G. Nakamura and K. Tanuma, *A formula for the fundamental solution of anisotropic elasticity*, Quart. J. Mech. Appl. Math., 50 (1997), pp. 179–194.
[20] T. Ōkaji, *Strong unique continuation property for time harmonic Maxwell equations*, J. Math. Soc. Japan, 54 (2002), pp. 89–122.
[21] M. Protter, *Unique continuation for elliptic equations*, Trans. Amer. Math. Soc., 95 (1960), pp. 81–91.
[22] R. Robertson, *Boundary identifiability of residual stress via the Dirichlet to Neumann map*, Inverse Problems, 13 (1997), pp. 1107–1119.
[23] V. Vogelsang, *Absence of embedded eigenvalues of the Dirac equation for long range potentials*, Analysis, 7 (1987), pp. 259–274.
[24] V. Vogelsang, *On the strong continuation principle for inequalities of Maxwell type*, Math. Ann., 289 (1991), pp. 285–295.
[25] N. Weck, *Unique continuation for systems with Lamé principal part*, Math. Methods Appl. Sci., 24 (2001), pp. 595–605.
[26] N. Weck, *Unique continuation for some systems of partial differential equations*, Appl. Anal., 13 (1982), pp. 53–63.

SIAM J. MATH. ANAL.
Vol. 35, No. 2, pp. 318–356

INSTABILITY OF SOME IDEAL PLANE FLOWS*

ZHIWU LIN†

Abstract. We prove the instability of large classes of steady states of the two-dimensional Euler equation. For an odd shear flow, beginning with the Rayleigh equation, we define a family of operators depending on some positive parameter. Then we use infinite determinants to keep track of the signs of the eigenvalues of these operators. The existence of purely growing modes follows from a continuation argument. Employing a new analysis of neutral modes together with a rigorous justification of Tollmien's classical method, we obtain a sharp condition for linear and hence nonlinear instability of a general class of bounded shear flows. We obtain similar results for bounded rotating flows and unbounded shear flows.

Key words. Rayleigh's equation, neutral limiting mode, shear flow instability

AMS subject classifications. 76E05, 76E09

DOI. 10.1137/S0036141002406266

1. Introduction. In this paper, we study the hydrodynamic stability problem for plane shear flows and rotating flows. The purpose is to get some sufficient conditions for linear instability and hence nonlinear instability. For plane shear flows, this problem has a long history, going back to scientists such as Rayleigh and Kelvin in the nineteenth century. The vorticity form of the incompressible two-dimensional Euler equation in a bounded domain D with smooth boundary ∂D is

$$\partial_t \omega + u \cdot \bigtriangledown \omega = 0 \quad \text{in } \mathbf{R}_t \times D$$

or

$$\partial_t \Delta\psi + \frac{\partial \psi}{\partial y}\frac{\partial}{\partial x}\Delta\psi - \frac{\partial \psi}{\partial x}\frac{\partial}{\partial y}\Delta\psi = 0, \tag{1}$$

where ψ is the stream function, $\omega = -\Delta\psi$ is the vorticity, and $u = (\frac{\partial \psi}{\partial y}, -\frac{\partial \psi}{\partial x})$ is the velocity. We consider the basic steady state flow $\mathbf{U}_0 = U(y)\,\mathbf{i}$, a parallel shear flow in the x-direction, in the flow domain $D = \{(x,y) \mid y_1 \le y \le y_2\}$ with rigid walls at $y = y_1, y_2$. This means u is tangential, or ψ is constant on each wall. The linearized equation of (1) around $\mathbf{U}_0$ is

$$\partial_t \Delta\tilde{\psi} + U\frac{\partial}{\partial x}\Delta\tilde{\psi} - U''\frac{\partial \tilde{\psi}}{\partial x} = 0, \tag{2}$$

where $\tilde{\psi}$ is constant on $y = y_j$ $(j = 1, 2)$. Taking $\tilde{\psi} = \phi(y)\, e^{i\alpha(x-ct)}$ with α the wave number (positive real) in the x-direction and $c = c_r + ic_i$ the complex wave speed, we obtain from (2) the Rayleigh equation

$$(U - c)\left(\frac{d^2}{dy^2} - \alpha^2\right)\phi - U''\phi = 0 \tag{3}$$

*Received by the editors April 24, 2002; accepted for publication (in revised form) January 31, 2003; published electronically July 18, 2003.
http://www.siam.org/journals/sima/35-2/40626.html
†Department of Mathematics, Brown University, Providence, RI 02912 (linzw@math.brown.edu).

with $\phi(y_1) = \phi(y_2) = 0$. We will also consider unbounded shear flows where one of y_1, y_2 is infinity, with the boundary condition $\phi(y) \to 0$ as $y \to \infty$.

So for shear flows, the instability problem is reduced to studying the Rayleigh equation (3). The flow is linearly unstable if some nontrivial solution to (3) with $c_i > 0$ exists. A classical result of Lord Rayleigh [16] is the necessary condition for instability that the basic velocity profile should have an inflection point at some point $y = y_s$, that is, $U''(y_s) = 0$. This condition was later improved by Fjørtoft [10]. Howard's semicircle theorem [14] says that any unstable eigenvalue $c = c_r + ic_i$ must lie in the semicircle

$$\left(c_r - \frac{1}{2}(U_{\min} + U_{\max})\right)^2 + c_i^2 \le \left(\frac{1}{2}(U_{\min} - U_{\max})\right)^2. \tag{4}$$

However, very few sufficient conditions for instability are known. In 1935, Tollmien [23] obtained an unstable solution to (3) by formally perturbing around a neutral mode (c real) for symmetric flows in class $\mathcal{K}^+$ (defined below). The original presentation was improved by C. C. Lin [17] and the asymptotic growth rate was found. Even in recent treatises such as [20], the main instability result mentioned is Tollmien's. However, as indicated by Friedlander and Howard [12], in all these references the existence of an unstable mode had to be *assumed* in a neighborhood of the neutral mode. The assumption of analytic dependence between the parameters α and c (complex) also lacked justification. These assumptions are rigorously justified in this paper. Here we get a sharp condition for the instability of a class of flows.

Let us describe the setting of the problem. First we define a class of flows having some inflection point. By an inflection value we mean the value of U at an inflection point.

DEFINITION 1.1. *The flow $U(y)$ is in class $\mathcal{K}$ if U is a $\mathbf{C}^2$ function on a interval $[y_1, y_2]$, and there exists some inflection value U_s such that*

$$K(y) := -U''(y) / (U(y) - U_s) \tag{5}$$

is nonnegative and bounded in $[y_1, y_2]$. If K is positive on $[y_1, y_2]$, we say that U is in class $\mathcal{K}^+$.

A typical example of such a flow is $U = \cos my$ or $\sin my$. Now we consider any flow in class $\mathcal{K}$. If (ϕ_s, α_s) is a solution to the Sturm–Liouville problem

$$\phi_s'' - \alpha_s^2 \phi_s + K\phi_s = 0, \ \phi = 0 \quad \text{at } y = y_1, y_2, \tag{6}$$

then $(\phi, \alpha, c) = (\phi_s, \alpha_s, U_s)$ is a special solution (a so-called neutral mode) to the Rayleigh equation (3). Let $\alpha_{\max}$ be the largest wave number so that a neutral mode exists. That is,

$$-\alpha_{\max}^2 = \inf_{\phi \in \mathbf{H}_0^1(y_1, y_2)} \frac{\int_{y_1}^{y_2} \left(|\phi'|^2 - K(y)|\phi|^2\right) dy}{\int_{y_1}^{y_2} |\phi|^2 \, dy}. \tag{7}$$

Throughout this paper, we assume that the right-hand side of (7) is negative. Otherwise, the shear flow was proved to be linearly stable by Drazin and Howard [8]. It was also proved in [8] that instability is possible only for wave numbers α such that $0 < \alpha < \alpha_{\max}$. Howard [15] estimated the maximal number of possible unstable modes for a fixed wave number. However, it still was not clear whether there exists

some unstable mode for each α in that range. Recently, Friedlander and Howard [12] studied the special flow $U(y) = \cos my$, using a continued fractions technique and a numerical method. For this flow they proved that for all $0 < \alpha < \alpha_{\max}$, there exists some growing mode for the Rayleigh equation.

In this paper, we rigorously prove that for any flow of class $\mathcal{K}^+$ and for all $0 < \alpha < \alpha_{\max}$, there does indeed exist an unstable solution to the Rayleigh equation (3). This is our main theorem.

THEOREM 1.2. *Suppose the steady state is in class $\mathcal{K}^+$. Let $-\alpha_{\max}^2$ be the lowest eigenvalue of $-\frac{d^2}{dy^2} - K(y)$, which is assumed to be negative. For all $\alpha \in (0, \alpha_{\max})$, there is an unstable solution (with $\operatorname{Im} c > 0$) to* (3).

The unstable interval $(0, \alpha_{\max})$ is sharp in the sense that there is linear stability if $\alpha \geq \alpha_{\max}$ or $-\frac{d^2}{dy^2} - K(y)$ is nonnegative. We can also treat plane rotating flows in an annulus. In this case, the analogue of the Rayleigh equation becomes

$$(\Omega - c)\left(D_* D - n^2/r^2\right)\phi - r^{-1}(rD^2\Omega + 3D\Omega)\phi = 0, \tag{8}$$

with $\phi(R_1) = \phi(R_2) = 0$, $0 < R_1 \leq r \leq R_2$. Here $\Omega(r)$ is the angular velocity of the steady state, $D_* = \frac{d}{dr} + \frac{1}{r}$, $D = \frac{d}{dr}$, and n is some integer. We have the following result analogous to Theorem 1.2.

THEOREM 1.3. *For the rotating case, if*

$$K(r) := -\left(rD^2\Omega + 3D\Omega\right)/(\Omega - \Omega_s) \tag{9}$$

is positive and $\Omega(R_1) \neq \Omega(R_2)$, then a necessary and sufficient condition for instability is that there exists $\alpha > 1$ such that the equation

$$\left(D_* D - \alpha^2/r^2\right)\phi + r^{-1}K(r)\phi = 0 \tag{10}$$

has some nontrivial solution with $\phi(R_1) = \phi(R_2) = 0$. This is equivalent to the condition

$$-\alpha_{\max}^2 := \inf_{\phi \in \mathbf{H}_0^1(R_1, R_2)} \frac{\int_{R_1}^{R_2} r\left(\frac{d}{dr}\phi\right)^2 dr - \int_{R_1}^{R_2} K(r)\phi^2 dr}{\int_{R_1}^{R_2} \frac{1}{r}\phi^2 dr} < -1. \tag{11}$$

In the case that $K(r)$ is positive and $\Omega(R_1) = \Omega(R_2)$, a sufficient condition for instability is that

$$-\alpha_{\max}^2 := \inf_{\phi \in \mathbf{H}_0^1(R_1, R_2)} \frac{\int_{R_1}^{R_2} r\left(\frac{d}{dr}\phi\right)^2 dr - \int_{R_1}^{R_2} K(r)\phi^2 dr}{\int_{R_1}^{R_2} \frac{1}{r}\phi^2 dr} < -4. \tag{12}$$

Let us return to shear flows that are not in class $\mathcal{K}^+$. If a shear flow is odd but there is no assumption on the sign of $K(y)$, we can still get a sufficient condition for instability.

THEOREM 1.4. *Assume $U(y)$ is odd in $[-a, a]$ and define*

$$K(y) := -U''(y)/U(y). \tag{13}$$

If K is bounded and the operator $-\frac{d^2}{dy^2} - K(y)$ with zero boundary values at $\pm a$ has a negative eigenvalue, then there is a solution to the Rayleigh equation (3) *with $c = i\lambda_0$ (here $\lambda_0 > 0$) for some range of wave numbers. Specifically, if $-\alpha_0^2 < -\alpha_1^2 < \cdots < -\alpha_{k_0}^2 < 0$ denote all the negative eigenvalues of $-\frac{d^2}{dy^2} - K(y)$, then we have a*

purely growing instability for α *belonging to the intervals* $(\alpha_1, \alpha_0) \cup \cdots \cup (\alpha_{2k-1}, \alpha_{2k-2}) \cdots \cup (\alpha_{k_0}, \alpha_{k_0-1})$ *(if* k_0 *odd) or to the intervals* $(\alpha_1, \alpha_0) \cup \cdots \cup (\alpha_{2k-1}, \alpha_{2k-2}) \cdots \cup (\alpha_{k_0-1}, \alpha_{k_0-2})$ *(if* k_0 *even).*

We can extend Theorems 1.2 and 1.4 to the case of unbounded shear flows.

THEOREM 1.5. (i) *(class* $\mathcal{K}^+$*) Assume* $U(y)$ *is in* $\mathbf{C}^2(-\infty, +\infty)$, $U(y) \to U(\pm\infty)$ *as* $y \to \pm\infty$, *and* $U(y)$ *takes the values* $U(\pm\infty)$ *at only a finite number of points. We consider the flows such that* $K(y)$ *defined by* (5) *is bounded, positive, and* $\lim_{y\to\pm\infty} K(y) = 0$. *Let* $-\alpha_0^2$ *be the lowest eigenvalue of* $-\frac{d^2}{dy^2} - K(y)$ *on* $\mathbf{H}^2(\mathbf{R})$, *which is assumed to be negative. Then for each* α *in* $(0, \alpha_0)$, *there is instability. This condition is sharp in the sense that if* $\alpha \geq \alpha_0$ *or* $-\frac{d^2}{dy^2} - K(y)$ *is nonnegative, then there is linear stability. The same result holds for the shear flows defined in the half line.*

(ii) *(odd flows) Assume* $U(y)$ *is in* $\mathbf{C}^2(-\infty, +\infty)$, *odd, and* $K(y)$ *defined by* (13) *is bounded and* $\lim_{y\to\pm\infty} K(y) = 0$. *If* $-\alpha_0^2 < -\alpha_1^2 < \cdots < -\alpha_k^2 < \cdots < 0$ *denote all the negative eigenvalues of the operator* $-\frac{d^2}{dy^2} - K(y)$ *on* $\mathbf{H}^2(\mathbf{R})$, *then we have a purely growing instability for* α *belonging to the intervals* $(\alpha_1, \alpha_0) \cup (\alpha_3, \alpha_2) \cup \cdots \cup (\alpha_{2k-1}, \alpha_{2k-2}) \cdots$.

Now let us sketch the main ideas of the proofs. For the proof of Theorem 1.4, we define a family of elliptic operators A_λ depending on the positive parameter λ where $c = i\lambda$. The problem is reduced to finding some λ_0 such that A_{λ_0} has a kernel. The operator A_λ is nonnegative when λ is large and A_λ has an odd number of eigenvalues when λ tends to 0. The idea then is to use an infinite determinant to keep track of the sign of the eigenvalues of A_λ as λ varies from 0 to ∞.

For the proofs of Theorems 1.2 and 1.3, we carefully use the neutral modes. In the literature, neutral modes have usually been used as the base modes, from which unstable modes have been obtained by the perturbation argument of Tollmien. The novelty of this paper is to utilize a different property of neutral modes: the neutral wave numbers are the possible boundary points of the set of all unstable wave numbers. Thus if we knew all these possible neutral wave numbers and the instability properties around them, we could deduce the stability properties at all the wave numbers. Indeed for our purpose we only need to understand the neutral modes from which the unstable modes can issue. We call them the neutral limiting modes to distinguish them from the usual neutral modes, which are just the solutions to the Rayleigh equation with real c.

DEFINITION 1.6. *The triple* (c_s, α_s, ϕ_s) *with* c_s *real and* α_s *positive is said to be a neutral limiting mode if it is the limit of the growing solution sequence* (c_k, α_k, ϕ_k) *(with* $\operatorname{Im} c_k > 0$*) of the Rayleigh equation* (3). *The precise notions of convergence of* ϕ_k *to* ϕ_s *will be made clear in Lemma* 3.6. *Formally* (c_s, α_s, ϕ_s) *ought to satisfy the Rayleigh equation*

$$(U - c_s)\left(\frac{d^2}{dy^2} - \alpha_s^2\right)\phi_s - U''\phi_s = 0. \tag{14}$$

We call c_s *the neutral limiting phase speed and* α_s *the neutral limiting wave number.*

Here in the above definition, the convergence of $\{c_k\}$ is guaranteed by Howard's semicircle theorem (4). From (4) we also know that c_s must lie in the range of $U(y)$. The importance of neutral limiting modes lies in the fact that the neutral limiting wave numbers are the possible boundary points of the set of all unstable wave numbers (see Theorem 3.9). The knowledge of the instability near every neutral limiting wave number will allow us to determine the instability in the whole range of wave numbers.

For that purpose, first we need to know what all the neutral limiting modes are. In general, it is difficult to get a simple answer. But we have the following simple characterization in case the flow is in class $\mathcal{K}$.

THEOREM 1.7. *If the flow is in class $\mathcal{K}$, then for any neutral limiting mode (c_s, α_s, ϕ_s) with positive α_s, the phase speed must be $c_s = U_s$ and the function ϕ_s must solve*

$$-\frac{d^2}{dy^2}\phi_s + \frac{U''}{U - U_s}\phi_s = -\alpha_s^2 \phi_s \tag{15}$$

with $\phi_s(y_1) = \phi_s(y_2) = 0$.

In the physics literature [17], [6], for a monotone flow it was shown heuristically by using Reynolds stress that the neutral limiting phase speed must be U_s. Using some lemmas of Sattinger [22], we rigorously prove the same result for large classes of flows. For flows in class $\mathcal{K}$, we have some uniform a priori bound on the $\mathbf{H}^2$ norm of unstable eigenfunctions. This enables us to deduce the other conclusions in the theorem.

Furthermore, for a flow in the class $\mathcal{K}^+$, we also obtain the instability property near the neutral wave numbers. This is done by rigorously verifying Tollmien's argument. By combining it with the boundary point property of neutral limiting wave numbers (Theorem 3.9), we obtain an unstable mode for each α in $(0, \alpha_{\max})$.

To prove Theorem 1.5, we truncate the unbounded flow to get a sequence of bounded flows. Then by applying Theorems 1.2 and 1.4 to truncated flows, we get a sequence of approximating unstable solutions. We can show that the sequence obtained converges to a nontrivial function, which is an unstable solution to the Rayleigh equation in the unbounded case.

In [1] Bardos, Guo, and Strauss rigorously proved nonlinear instability from the existence of growing modes under a certain assumption for flows defined on bounded domains. For rotating flows as in Theorem 1.3, that assumption is satisfied. For shear flows as in Theorem 1.2, we assume the x-direction is P-periodic, with the wave number α being multiples of $\frac{2\pi}{P}$. Then the result in [1] can still apply. The nonlinear instability proved in [1] is in the $\mathbf{L}^2$ norm of the vorticity. In [13], Grenier proved nonlinear instability from the existence of growing modes for very general shear flows. In particular, nonlinear instability of shear flows in [13] can be proved in unbounded spaces. Thus the flows in Theorem 1.5 are also nonlinearly unstable. Note that the nonlinear instability in [13] is in the $\mathbf{L}^\infty$ and $\mathbf{L}^2$ norms of velocity.

We can generalize most of Theorem 1.2 of this paper to general shear flows in the class $\mathcal{F}$ (see Definition 3.1). Thus we can treat any flow with a monotone velocity profile $U(y)$ or any flow that satisfies a differential equation $U''(y) = g(U(y))\,k(y)$ for some function $k(y) > 0$. The details will appear in a forthcoming paper. In [19], we use the method of section 2 to treat linear instability of general ideal plane flows.

The paper is organized as follows. In section 2, we prove Theorem 1.4 for odd flows. We study the neutral limiting modes in section 3. Section 4 is devoted to the proof of Theorem 1.2. In section 5, we give the proof of Theorem 1.3 for the rotating case. We treat unbounded shear flows in section 6.

2. Odd flows. We divide the proof of Theorem 1.4 into several steps. First we reduce the problem to the eigenvalue problem of an ODE system. Let $c = i\lambda\,(\lambda > 0)$ and $\phi = f + ih$; then (3) becomes

$$\left(-\frac{d^2}{dy^2} + \alpha^2\right)(f + ih) + \left(\frac{U''U}{U^2 + \lambda^2} + i\frac{\lambda U''}{U^2 + \lambda^2}\right)(f + ih) = 0.$$

Comparing the real and imaginary parts of (3) and using the definition of $K(y)$, we get

$$-\frac{d^2}{dy^2}f+\alpha^2 f-K(y)f+K(y)\frac{\lambda^2}{\lambda^2+U(y)^2}f+K(y)\frac{\lambda U(y)}{\lambda^2+U(y)^2}h=0, \tag{16a}$$

$$-\frac{d^2}{dy^2}h+\alpha^2 h-K(y)h-K(y)\frac{\lambda U(y)}{\lambda^2+U(y)^2}f+K(y)\frac{\lambda^2}{\lambda^2+U(y)^2}h=0 \tag{16b}$$

with $f=h=0$ at $y=-a,a$. If we denote

$$A_0=\begin{pmatrix} -\frac{d^2}{dy^2}+\alpha^2-K(y) & 0 \\ 0 & -\frac{d^2}{dy^2}+\alpha^2-K(y) \end{pmatrix}$$

and

$$B_\lambda=K(y)\begin{pmatrix} \frac{\lambda^2}{\lambda^2+U(y)^2} & \frac{\lambda U(y)}{\lambda^2+U(y)^2} \\ -\frac{\lambda U(y)}{\lambda^2+U(y)^2} & \frac{\lambda^2}{\lambda^2+U(y)^2} \end{pmatrix},$$

$A_\lambda=A_0+B_\lambda$. Then (16) becomes

$$A_\lambda\begin{pmatrix} f \\ h \end{pmatrix}=0.$$

The common domain for the operators A_λ is

$$\mathcal{H}=\left\{(f,h)\mid f,h\in\left(\mathbf{H}^2(-a,a)\cap\mathbf{H}_0^1(-a,a)\right)\text{ and }f\text{ odd, }h\text{ even}\right\}.$$

Let

$$\mathcal{X}=\left\{(f,h)\mid f,h\in\mathbf{L}^2(-a,a)\text{ with }f\text{ odd, }h\text{ even}\right\}.$$

Here $\mathcal{H}$, $\mathcal{X}$ are complex spaces. Due to the oddness of $U(y)$, $A_\lambda:\mathcal{H}\to\mathcal{X}$. In the following $\|.\|$ denotes the $\mathbf{L}^2$ norm. We have the following simple characterization of A_λ.

LEMMA 2.1. *A_λ is a densely defined closed operator, and for any ξ in its resolvent set $\rho(A_\lambda)$, $(\xi-A_\lambda)^{-1}$ is a trace class operator. The eigenvalues of A_λ appear in complex conjugate pairs and are all discrete with finite multiplicity.*

Proof. Denote

$$A=\begin{pmatrix} -\frac{d^2}{dy^2} & 0 \\ 0 & -\frac{d^2}{dy^2} \end{pmatrix}$$

with $D(A)=\mathcal{H}$. Then clearly $(\xi-A)^{-1}$ is a trace class operator for any $\xi\in\rho(A)$ and we have

$$\left\|(A+k)^{-1}\right\|\le\frac{1}{k}$$

for any $k>0$. On the other hand, $A_\lambda-A$ are uniformly bounded operators, and suppose $\|A_\lambda-A\|\le M$. We have

$$A_\lambda+k=A+k+A_\lambda-A=\left(1+(A_\lambda-A)(A+k)^{-1}\right)(A+k).$$

If $M < k$, then $-k \in \rho(A_\lambda)$ and

$$(A_\lambda + k)^{-1} = (A+k)^{-1}\left(1 + (A_\lambda - A)(A+k)^{-1}\right)^{-1}.$$

This is the multiplication of a bounded operator with a trace class operator, so it is also in trace class. For any $\xi \in \rho(A_\lambda)$, from formula

$$(\xi - A_\lambda)^{-1} = (-k - A_\lambda)^{-1} + (\xi + k)(\xi - A_\lambda)^{-1}(-k - A_\lambda)^{-1},$$

we can see that $(\xi - A_\lambda)^{-1}$ is in trace class.

Now the conclusions about the eigenvalues of A_λ follow from the trace class property just proved and the fact that the coefficients of A_λ are real. □

Now we study the semigroup generated by $-A_\lambda$. Notice that $-A_\lambda$ is a bounded perturbation of

$$A = \begin{pmatrix} \frac{d^2}{dy^2} & 0 \\ 0 & \frac{d^2}{dy^2} \end{pmatrix},$$

which generates the diffusion semigroup. Then by the bounded perturbation theorem of semigroups, we know that $-A_\lambda$ generates a strongly continuous semigroup. Denote $T_\lambda(t) = \exp(-tA_\lambda)$. Then there exists some C, ω positive (independent of λ) such that

$$\|T_\lambda(t)\| \le Ce^{\omega t}.$$

We have the following characterization of $T_\lambda(t)$.

LEMMA 2.2. *For all $t > 0$, $T_\lambda(t)$ is in trace class.*

Proof. First we claim that $A_\lambda T_\lambda(t)$ is a bounded operator. Assuming the claim, the theorem follows easily since we have for any $\xi \in \rho(A)$

$$T_\lambda(t) = (\xi - A_\lambda)^{-1}\left((\xi - A_\lambda)T_\lambda(t)\right),$$

which is the multiplication of a trace class operator with a bounded operator, so it is in trace class.

We shall now prove the claim, which is due to the smoothing effect of $T_\lambda(t)$. We need to show only that $AT_\lambda(t)$ is bounded. For this purpose we study the evolution equation associated with $T_\lambda(t)$.

$$\frac{d}{dt}f = \frac{d^2}{dy^2}f - \alpha^2 f + K(y)f - K(y)\frac{\lambda^2}{\lambda^2 + U(y)^2}f - K(y)\frac{\lambda U(y)}{\lambda^2 + U(y)^2}h, \tag{17a}$$

$$\frac{d}{dt}h = \frac{d^2}{dy^2}h - \alpha^2 h + K(y)h + K(y)\frac{\lambda U(y)}{\lambda^2 + U(y)^2}f - K(y)\frac{\lambda^2}{\lambda^2 + U(y)^2}h \tag{17b}$$

with $f(0) = f_0$, $h(0) = h_0$. Now to show the claim, it suffices to prove

$$\left\|\frac{d^2}{dy^2}f(t)\right\|_2, \left\|\frac{d^2}{dy^2}h(t)\right\|_2 \le C(t)\left(\|f_0\|_2 + \|h_0\|_2\right).$$

We denote (17) by

$$\frac{d}{dt}f = \frac{d^2}{dy^2}f + R_1(f,h), \tag{18a}$$

$$\frac{d}{dt}h = \frac{d^2}{dy^2}h + R_2(f,h). \tag{18b}$$

Then it is easy to see that

$$\begin{aligned}\|R_1\|_2, \|R_2\|_2 &\le C_0 \left(\|f(t)\|_2 + \|h(t)\|_2\right) \\ &\le C_0 C e^{\omega t} \left(\|f_0\|_2 + \|h_0\|_2\right).\end{aligned}$$

So from the regularity theory of the linear parabolic equation, we have

$$\begin{aligned}\left\|\frac{d^2}{dy^2} f(t)\right\|_2, \left\|\frac{d^2}{dy^2} h(t)\right\|_2 &\le C'(t) \left(\|f(t)\|_2 + \|h(t)\|_2 + \|R_1\|_2 + \|R_2\|_2\right) \\ &\le C''(t) \left(\|f_0\|_2 + \|h_0\|_2\right).\end{aligned}$$

Thus the claim is proved. □

From Lemmas 2.1 and 2.2, we know that the eigenvalues of A_λ and $T_\lambda(t)$ are discrete with finite multiplicity and that

$$\sigma(T_\lambda(t)) \setminus \{0\} = \exp\left(-t\sigma(A_\lambda)\right).$$

Now denote all the distinct eigenvalues of A_λ (arranged with nondecreasing real part) by $\mu_1(\lambda), \mu_2(\lambda), \ldots, \mu_k(\lambda), \ldots$, with multiplicities $n_1, n_2, \ldots, n_k, \ldots$. We define the infinite determinant of $Id - T_\lambda(1)$ as

$$d(\lambda) = \prod_{k=1}^{\infty} \left(1 - \exp\left(-\mu_k(\lambda)\right)\right)^{n_k}.$$

Since $T_\lambda(1)$ is a trace class operator and $\mu_k(\lambda)$ appears in complex conjugate pairs, $d(\lambda)$ is a finite real number. From the definition of $d(\lambda)$, we know that the sign of $d(\lambda)$ is determined only by the number of negative real eigenvalues of A_λ. If this number is odd, then $d(\lambda)$ is negative. And $d(\lambda)$ is positive if the number is even. Here we always assume A_λ has no kernel, since otherwise we have already obtained a solution to the Rayleigh equation.

We define three sets

$$S_- = \{\lambda > 0 |\ d(\lambda) < 0\}, \quad S_+ = \{\lambda > 0 |\ d(\lambda) > 0\}, \quad S_0 = \{\lambda > 0 |\ d(\lambda) = 0\}.$$

We will show that S_-, S_+ are nonempty open sets. Then the theorem follows easily, as we shall now show.

Proof of Theorem 1.4. We claim that S_0 is nonempty. Otherwise we would have $(0, +\infty) = S_- \cup S_+$, which is impossible, since S_-, S_+ are two disjoint open sets. So there must exist some $\lambda_0 > 0$ such that $d(\lambda_0) = 0$. Then there exists k so that $1 - \exp\left(-\mu_k(\lambda_0)\right) = 0$. So $\mu_k(\lambda_0) = 0$ and A_{λ_0} has a nontrivial kernel (f, h). This means that $c = i\lambda_0$, $\phi = f + ih$ is a solution to Rayleigh's equation (3). □

The next several lemmas prove the properties of S_-, S_+ that we need.

LEMMA 2.3. *S_+ is nonempty.*

Proof. Because for any real vector (f, h),

$$\begin{aligned}\left((f,h), A_\lambda \begin{pmatrix} f \\ h \end{pmatrix}\right) &= \left(\left(-\frac{d^2}{dy^2} + \alpha^2 - K(y) + K(y) \frac{\lambda^2}{\lambda^2 + U(y)^2}\right) f, f\right) \\ &\quad + \left(\left(-\frac{d^2}{dy^2} + \alpha^2 - K(y) + K(y) \frac{\lambda^2}{\lambda^2 + U(y)^2}\right) h, h\right) \\ &> 0\end{aligned}$$

when λ is large, A_λ is a positive operator. Thus all its real eigenvalues are positive, so that $d(\lambda) > 0$ for λ large. □

LEMMA 2.4. *S_- is nonempty.*

Proof. From the assumptions of Theorem 1.4 and the definition of operator A_0, we know that $d(0) < 0$. We will show that for λ small, $d(\lambda) < 0$.

First we claim that

(i) for any eigenvalue $\mu(\lambda)$ of A_λ, we have $|\mathrm{Im}\,\mu(\lambda)| < \|K\|_\infty$;

(ii) there exists positive ε_1, δ_1 such that if $0 \le \lambda < \delta_1$, then for any eigenvalue $\mu(\lambda)$ of A_λ, we have $|\mathrm{Re}\,\mu(\lambda)| > \varepsilon_1$.

Proof of claim (i). Let (f, h) be the eigenfunction with $\|f\|_2 + \|h\|_2 = 1$. Taking inner products with the conjugate $(\bar{f},\ \bar{h})$ on both sides of

$$A_\lambda \begin{pmatrix} f \\ h \end{pmatrix} = \mu(\lambda) \begin{pmatrix} f \\ h \end{pmatrix} \tag{19}$$

and comparing the imaginary parts, we get

$$\begin{aligned} |\mathrm{Im}\,\mu(\lambda)| &\le \left| 2\|K\|_\infty \mathrm{Im} \int_{-a}^{a} \frac{\lambda U(y)}{\lambda^2 + U(y)^2} f\bar{h}dy \right| \\ &\le \|K\|_\infty \frac{1}{2} (\|f\|_2 + \|h\|_2)^2 = \frac{1}{2}\|K\|_\infty . \end{aligned}$$

Proof of claim (ii). Supposing it is not true, we could find a sequence $\lambda_n \to 0$, μ_n being an eigenvalue of A_{λ_n}, and $\mathrm{Re}\,\mu_n \to 0$. Let (f_n, h_n) be the corresponding eigenfunction and $\|f_n\|_2 + \|h_n\|_2 = 1$. By (i), $\{\mu_n\}$ is a bounded sequence. We can find a subsequence such that $\mu_{n_k} \to \mu_0$, so that μ_0 is purely imaginary. We still denote the subsequence by $\{\mu_n\}$.

From the equation satisfied by the eigenfunction (f_n, h_n), we get

$$\|f_n\|_{\mathbf{H}^2}, \|g_n\|_{\mathbf{H}^2} \le C(\|f\|_2 + \|h\|_2) = C$$

from elliptic regularity theory by noticing that the coefficients in (19) are uniformly bounded. Thus there exists a subsequence such that $(f_{n_k}, g_{n_k}) \to (f_0, g_0)$ weakly in $\mathbf{H}^2$ and strongly in $\mathbf{H}^1$. Moreover,

$$\begin{aligned} \left\| (A_0 - \mu_0) \begin{pmatrix} f_{n_k} \\ h_{n_k} \end{pmatrix} \right\| \le & \|K\|_\infty \left(\left\| \frac{\lambda_{n_k}^2}{\lambda_{n_k}^2 + U(y)^2} \right\|_2 + \left\| \frac{\lambda_{n_k} U(y)}{\lambda_{n_k}^2 + U(y)^2} \right\|_2 \right) \\ & \times \left(\|f_{n_k}\|_\infty + \|h_{n_k}\|_\infty \right) + |\mu_{n_k} - \mu_0| \left(\|f_{n_k}\|_2 + \|h_{n_k}\|_2 \right) \\ \le & C \left(\left\| \frac{\lambda_{n_k}^2}{\lambda_{n_k}^2 + U(y)^2} \right\|_2 + \left\| \frac{\lambda_{n_k} U(y)}{\lambda_{n_k}^2 + U(y)^2} \right\|_2 + |\mu_{n_k} - \mu_0| \right) \end{aligned}$$

tends to zero as $\lambda_{n_k} \to 0$. Thus we have $\mu_0 \in \sigma(A_0)$, which is a contradiction to the fact that A_0 has no eigenvalue lying on the imaginary axis. So claim (ii) is proved.

Let Λ be the infimum of real part of eigenvalues of A_λ. Λ is finite since A_λ is uniformly bounded from below. Define

$$D = \left\{ (x, y) \ | \Lambda - 1 < x < -\frac{\varepsilon_1}{2},\ -\|K\|_\infty < y < \|K\|_\infty \right\}$$

and $\Gamma = \partial D$. From claim (ii), if $\lambda < \delta_1$, all eigenvalues of A_λ with negative real part lie in D. Define the Riesz projection as

$$P_\lambda = \frac{1}{2\pi i} \oint_\Gamma (k - A_\lambda)^{-1} \, dk \tag{20}$$

and $R(P_\lambda)$ its range, where $\lambda \geq 0$ and the Γ-integral is in the counterclockwise sense. Then by the definition of $d(\lambda)$

$$sign \; d(\lambda) = (-1)^{\dim R(P_\lambda)}. \tag{21}$$

To prove the lemma, it suffices to show that $\|P_\lambda - P_0\| \to 0$ as $\lambda \to 0$. If so, then $\dim(R(P_\lambda)) = \dim(R(P_0))$ if λ is small enough. By the definition of P_0, $\dim(R(P_0))$ is the number of negative eigenvalues of A_0 on the space $\mathcal{H}$, which is equal to that of the operator $-\frac{d^2}{dy^2} + \alpha^2 - K(y)$ on the space $\mathbf{H}^2(-a,a) \cap \mathbf{H}_0^1(-a,a)$. This is due to the fact that any eigenfunction of $-\frac{d^2}{dy^2} + \alpha^2 - K(y)$ is either odd or even when $K(y)$ is even. With α lying in the intervals of Theorem 1.4, $-\frac{d^2}{dy^2} + \alpha^2 - K(y)$ has an odd number of negative eigenvalues, so $\dim(R(P_0))$ is odd. Thus when λ is small enough, $\dim(R(P_\lambda))$ is odd, which implies that $d(\lambda)$ is negative by (21) so that S_- is not empty.

To show $\|P_\lambda - P_0\| \to 0$, we note that

$$\begin{aligned}
&\left\| B_\lambda \begin{pmatrix} f \\ h \end{pmatrix} \right\| \leq \|K\|_\infty \left(\left\| \frac{\lambda^2}{\lambda^2 + U(y)^2} \right\|_2 + \left\| \frac{\lambda U(y)}{\lambda^2 + U(y)^2} \right\|_2 \right) (\|f\|_\infty + \|h\|_\infty) \\
&\leq C \|K\|_\infty \left(\left\| \frac{\lambda^2}{\lambda^2 + U(y)^2} \right\|_2 + \left\| \frac{\lambda U(y)}{\lambda^2 + U(y)^2} \right\|_2 \right) \left(\left\| A_0 \begin{pmatrix} f \\ h \end{pmatrix} \right\| + \left\| \begin{pmatrix} f \\ h \end{pmatrix} \right\| \right) \\
&= C(\lambda) \left(\left\| A_0 \begin{pmatrix} f \\ h \end{pmatrix} \right\| + \left\| \begin{pmatrix} f \\ h \end{pmatrix} \right\| \right),
\end{aligned}$$

where

$$C(\lambda) = C \|K\|_\infty \left(\left\| \frac{\lambda^2}{\lambda^2 + U(y)^2} \right\|_2 + \left\| \frac{\lambda U(y)}{\lambda^2 + U(y)^2} \right\|_2 \right) \to 0$$

as $\lambda \to 0$ by dominant convergence. Since $\Gamma \subset \sigma(A_\lambda)$ if $\lambda < \delta_1$ and Γ is compact, it follows that $\| (\xi - A_\lambda)^{-1} \|$ is uniformly bounded by some constant M independent of $\xi \in \Gamma$. Then we have

$$\begin{aligned}
\left\| (\xi - A_\lambda)^{-1} - (\xi - A_0)^{-1} \right\| &= \left\| (\xi - A_\lambda)^{-1} B_\lambda (\xi - A_0)^{-1} \right\| \\
&\leq \left\| (\xi - A_\lambda)^{-1} \right\| \left\| B_\lambda (\xi - A_0)^{-1} \right\| \\
&\leq MC(\lambda) \left(\left\| A_0 (\xi - A_0)^{-1} \right\| + \left\| (\xi - A_0)^{-1} \right\| \right) \\
&\leq MC(\lambda) \left(1 + \left\| \xi (\xi - A_0)^{-1} \right\| + \left\| (\xi - A_0)^{-1} \right\| \right).
\end{aligned}$$

So as $\lambda \to 0$, $\| (\xi - A_\lambda)^{-1} - (\xi - A_0)^{-1} \| \to 0$ uniformly for $\xi \in \Gamma$. Thus $\|P_\lambda - P_0\| \to 0$ if $\lambda \to 0$. □

LEMMA 2.5. *S_- and S_+ are open sets.*

Proof. We will show that S_- is open. The proof for S_+ is the same. Suppose $\lambda_0 \in S_-$. Let $b > 0$ be such that there is no eigenvalue of A_{λ_0} with real part b. Then by the same argument as in the last lemma, there exists $\varepsilon_1, \delta_1 > 0$ such that if $|\lambda - \lambda_0| < \delta_1$, then for any eigenvalue $\mu(\lambda)$ of A_λ, we have $|\operatorname{Re}\mu(\lambda) - b| > \varepsilon_1$. Let Λ be the infimum of real part of eigenvalues of A_λ. Define

$$D = \left\{(x,y) \ |\Lambda - 1 < x < -\frac{\varepsilon_1}{2} + b, \ -\|K\|_\infty < y < \|K\|_\infty\right\}$$

and $\Gamma = \partial D$. Then all eigenvalues of A_λ with real part smaller than b lie in D and $\Gamma \subset \sigma(A_\lambda)$ provided $|\lambda - \lambda_0| < \delta_1$. Define P_λ by (20). Then $\|P_\lambda - P_{\lambda_0}\| \to 0$ as $|\lambda - \lambda_0| \to 0$, since A_λ is analytic for $\lambda > 0$. So $\dim(R(P_\lambda)) = \dim(R(P_{\lambda_0}))$ if $|\lambda - \lambda_0|$ is small enough. Let $\mu_1, \mu_2, \ldots, \mu_N$ be all the distinct eigenvalues of A_{λ_0} in D. Let m_k be the multiplicity of μ_k. Now for each μ_k, we can take a small ball $B_k = B(\mu_k; r_k)$ such that there are no other eigenvalues of A_{λ_0} in it besides μ_k. And by taking r_k small enough we can suppose that B_k does not intersect with the imaginary axis if $\operatorname{Re}\mu_k \neq 0$, and B_k does not intersect with the real axis if $\operatorname{Re}\mu_k = 0$. Also B_k does not intersect with Γ. They are disjoint with others, and for the conjugate eigenvalue we take the same radius. Then if $|\lambda - \lambda_0|$ is small enough, by analytic perturbation theory, there are exactly m_k eigenvalues (counting multiplicity) of A_λ in each B_k. Since $\dim(R(P_\lambda)) = \dim(R(P_{\lambda_0}))$, these are all the eigenvalues of A_λ in D. Now notice that for each B_k and its conjugate one, if we multiply all the eigenvalues of A_λ in them, the sign is the same as for A_{λ_0}. So in the definition of $d(\lambda)$, the part corresponding to the multiplication of all eigenvalues of A_λ with real part smaller than b is of the same sign with the λ_0 case. Thus it is negative if $|\lambda - \lambda_0|$ is small. While the other part of multiplication is always positive, we proved that $d(\lambda)$ is negative when $|\lambda - \lambda_0|$ is small. This finishes the proof of the lemma. □

It is easy to see that we can get the following abstract version by the same proof.

THEOREM 2.6. *Consider a family of real operators $A_\lambda = -A + B_\lambda$ $(\lambda \in (0, +\infty))$ with the same domain $\mathcal{H}$. We assume the following:*

(I) *B_λ is bounded and norm continuous for positive λ.*

(II) *A generates a generalized parabolic semigroup; that is, $\exp(tA)$ is in trace class and $A\exp(tA)$ is bounded.*

(III) *When λ is sufficiently large, A_λ has no eigenvalue with negative real part.*

(IV) *When λ tends to 0, A_λ tend to A_0 in the sense that*

$$\|(A_\lambda - A_0)\phi\| \leq c(\lambda)\left(\|A_0\phi\| + \|\phi\|\right),$$

$c(\lambda) \to 0$ as $\lambda \to 0^+$ for any function $\phi \in \mathcal{H}$. Then if A_0 has an odd number of negative eigenvalues and no kernel, there must exist some $\lambda_0 > 0$ such that A_{λ_0} has a nontrivial kernel.

We can also treat the periodic and Neumann boundary conditions for the Rayleigh equation by the same method. The conclusion and the proofs are direct analogues of Theorem 1.4.

Example 2.7 (doubly symmetric flows). Theorem 1.4 could be used to treat some nonodd flows.

Suppose that $U(y)$ is even on $(0, 2d)$ with respect to its midpoint d and is odd on $(0, d)$ with respect to its midpoint $\frac{d}{2}$. In that case, we could treat the sinuous (even mode) and varicose (odd mode) separately by studying the Rayleigh equation on $[0, d]$, taking the boundary condition at d to be either $\phi'(d) = 0$ or $\phi(d) = 0$. We could treat the varicose case by Theorem 1.4.

The flow $U(y) = \cos(my)$ on $[-\pi, \pi]$ was treated in [12]. If m is odd, then $U(y)$ is in the class we described above. For varicose modes, we can restrict the problem to $[0, \pi]$ and furthermore restrict the function space to be the space $\mathcal{P}_j$ spanned by $\sin ny$ $(n = j + mp)$. Here j is a fixed integer in $[1, [m/2]]$. Then the space $\mathcal{P}_j$ is invariant under the operator A_λ corresponding to $U(y) = \cos(my)$. Notice that if

$$m^2 - (m-j)^2 < \alpha^2 < m^2 - j^2,$$

then $-\frac{d^2}{dy^2} + \alpha^2 - m^2$ has only one negative eigenvalue on $\mathcal{P}_j$. Thus from Theorem 1.4, we know that there is a purely growing unstable mode. This was proved in [12] by a continued fractions technique. It was also shown in [12] by numerical computation that if α is small, there is no purely growing mode.

3. Neutral limiting modes. In this section, we study properties of the possible neutral limiting modes. For a certain class of flows, we get a simple characterization of the neutral limiting phase speed c_s. For flows of class $\mathcal{K}$, we get a complete characterization as in Theorem 1.7.

DEFINITION 3.1. *A velocity profile $U(y)$ is said to be in class $\mathcal{F}$ if for each number c in the range of U but not an inflection value, U'' takes the same sign at all points where $U(y) = c$.*

Some examples in class $\mathcal{F}$ are a monotone flow, a symmetric flow with monotone half part, and a flow satisfying $U''(y) = g(U(y))\, k(y)$ for some function g and $k(y) > 0$. It is readily seen that $\mathcal{K} \subset \mathcal{F}$.

Remark 3.2. We mention two simple facts we will use later.

(i) For a $\mathbf{C}^2$ flow $U(y)$, if c is not an inflection value, then $U(y) = c$ can only hold at a finite number of points.

(ii) For a $\mathbf{C}^2$ flow $U(y)$, if there exists some inflection value U_s such that the function

$$K(y) := -U''(y) / (U(y) - U_s) \tag{22}$$

is bounded on $[y_1, y_2]$, then $U(y) - U_s = 0$ can only hold at a finite number of points.

For the proof of (i), we notice that $U''(y_0) \neq 0$ at any point $y_0 \in \{U(y) = c\}$, since c is not an inflection value. So y_0 is an isolated point of $\{U(y) = c\}$. Therefore $\{U(y) = c\}$ is a finite set. For (ii) we observe that $\phi = U(y) - U_s$ solves a second order regular ODE

$$\phi'' + K(y)\,\phi = 0$$

on $[y_1, y_2]$. So the zeros of ϕ cannot cluster in the interval.

THEOREM 3.3. *If $U(y)$ is in class $\mathcal{F}$, then the neutral limiting phase speed must be an inflection value.*

Note that in Definition 1.6, α_s is positive. If $\alpha_s = 0$, then the neutral limiting phase speed might not be the inflection value. A counterexample is $U(y) = \cos(6y)$, $y \in (-\pi, \pi)$. The numerical computation in [12] indicated that when $\alpha_s = 0$, the neutral limiting phase speed is $c_s = -1$ while the inflection value is 0.

For the proof of this theorem, we need several lemmas from the literature, which we state without proof. The first one is an important equality which was first used to prove Rayleigh's criterion.

LEMMA 3.4. *Let ϕ be a solution of* (3) *with complex eigenvalue $c = c_r + ic_i$ $(c_i \neq 0)$, and let*

$$J_q(\phi) = \int_{y_1}^{y_2} \left(|\phi'|^2 + \alpha^2 |\phi|^2 + \frac{U''(U-q)}{|U-c|^2} |\phi|^2 \right) dy. \tag{23}$$

Then $J_q(\phi) = 0$ for every real number q.

Proof. We multiply the Rayleigh equation

$$\left(\frac{d^2}{dy^2} - \alpha^2 \right) \phi - \frac{U''}{U-c}\phi = 0$$

by ϕ^* ($*$ denotes the complex conjugate) and integrate it to get

$$\int_{y_1}^{y_2} \left(|\phi'|^2 + \alpha^2 |\phi|^2 + \frac{U''}{U-c} |\phi|^2 \right) dy = 0.$$

Comparing real and imaginary parts, we get

$$\int_{y_1}^{y_2} \left(|\phi'|^2 + \alpha^2 |\phi|^2 + \frac{U''(U-c_r)}{|U-c|^2} |\phi|^2 \right) dy = 0, \tag{24}$$

$$\int_{y_1}^{y_2} \frac{U''}{|U-c|^2} |\phi|^2 \, dy = 0. \tag{25}$$

Combining (24) and (25), we get the conclusion. □

We also need some results from [22]. In the following we use the notation in [22]. Let c be any real number in the range of $U(y)$ and let $z_1 < z_2 < \cdots < z_{k_c}$ be the zeros of $U(y) - c$. Here we assume k_c is finite. In the following we always consider the cases in Remark 3.2, so this assumption is valid. We denote by S_0 the complement of the set of points $\{z_i\}$ in the interval $[y_1, y_2]$. Let $z_0 = y_1$ and $z_{k_c+1} = y_2$. Then we have the following lemma.

LEMMA 3.5. *Let ϕ satisfy* (3) *with positive α and c as above on S_0, where ϕ is sectionally continuous on the open intervals (z_j, z_{j+1}), $j = 0, 1, \ldots, k_c$. Then ϕ cannot vanish at both endpoints of any of the intervals (z_j, z_{j+1}) unless it vanishes identically on that interval.*

Proof. This lemma was proved in [22], where it was used for a different purpose, namely, to show that for a fixed wave number there are only a finite number of unstable eigenvalues of the Rayleigh equation under some conditions. Here we give the proof for completeness.

The Rayleigh equation (3) can be rewritten as

$$\left((U-c)\phi' - U'\phi \right)' = \alpha^2 (U-c)\phi. \tag{26}$$

Suppose $\phi(z_i+) = \phi(z_{i+1}-) = 0$ and study (26) in $[z_i, z_{i+1}]$. From the definition of z_i, $U - c$ has constant sign in (z_i, z_{i+1}).

If $z_i \neq y_1$ $(i \neq 0)$, then $U(z_i) - c = 0$. Let $\tilde{z} \leq z_{i+1}$ be the nearest zero of ϕ in $(z_i, z_{i+1}]$. Since (26) is a real equation, we may assume ϕ is real and nonnegative on the interval $(z_i, \tilde{z})$ and that $\phi'(z_i) \geq 0$ and $\phi'(\tilde{z}) \leq 0$. Integrating (26) over $(z_i, \tilde{z})$, we get

$$(U(\tilde{z}) - c)\phi'(\tilde{z}) = \alpha^2 \int_{z_i}^{\tilde{z}} \phi(U-c)dz',$$

since ϕ vanishes at the endpoints $z_i, \tilde{z}$.

If $\tilde{z} = z_{i+1,}$ then the left-hand side above must be zero. Hence ϕ is identically zero on (z_i, z_{i+1}). On the other hand, if $\tilde{z} < z_{i+1}$, then $U(\tilde{z}) \neq c$ and

$$\phi'(\tilde{z}) = \alpha^2 \int_{z_i}^{\tilde{z}} \frac{(U(z') - c)}{(U(\tilde{z}) - c)} \phi(z')\, dz',$$

which could not hold true unless $\phi \equiv 0$ on $[z_i, \tilde{z}]$. But the second order ODE (26) is regular on (z_i, z_{i+1}). Thus z could not be a cluster point of a nontrivial solution ϕ. Thus ϕ must be identically zero on (z_i, z_{i+1}).

If $i = 0$, then we repeat the same argument with the right endpoint of the interval (y_1, z_1). □

LEMMA 3.6. *Let* $\{(c_k, \alpha_k, \phi_k)\,(\text{with } \operatorname{Im} c_k > 0)\}_{k=1}^{\infty}$ *be the solutions to the Rayleigh equation* (3) *and* $\|\phi_k\| = 1$, *and* (c_k, α_k) *converges to* (c_s, α_s) *with positive* α_s. *Then* ϕ_k *converges uniformly to a function* ϕ_s *on any compact subset of* S_0, ϕ_s'' *exists on* S_0, *and* ϕ_s *satisfies* (14).

The case when α_k is independent of k was proved in [22], but the proof can be applied to the current case without much change. The basic idea is that on compact subsets of S_0, the function $1/(U(y) - c_k)$ is uniformly bounded, so we get a uniform bound on the derivatives of ϕ_k up to second order.

Proof of Theorem 3.3. Let (c_s, α_s, ϕ_s) be a neutral limiting mode and assume c_s is not an inflection value. First we show that the ϕ_s obtained by Lemma 3.6 is not identically zero. Otherwise suppose $\phi_s \equiv 0$. Let $z_1, z_2, \ldots, z_m$ be all the zeros of $U(y) - c_s$, which by Remark 3.2 is finite. Then by the assumption of the theorem and the definition of class $\mathcal{F}$, all $U''(z_i)$ have the same sign, say positive. Let $E_\delta = \{y \in [y_1, y_2] \mid |y - z_i| < \delta \text{ for some } i\}$. Then $E_\delta^c \subset S_0$ and $U''(y) > 0$ for $y \in E_\delta$ if δ small enough. Take $q = \min U(y) - 1$ and assume $\|\phi_k\|_2 = 1$. Then

$$\begin{aligned} J_q(\phi_k) &= \int_{y_1}^{y_2} \left(|\phi_k'|^2 + \alpha_k^2 |\phi_k|^2 + \frac{U''(U-q)}{|U - c_k|^2} |\phi_k|^2 \right) dy \\ &\geq \alpha_k^2 + \int_{E_\delta^c} \frac{U''(U-q)}{|U - c_k|^2} |\phi_k|^2\, dy + \int_{E_\delta} \frac{U''(U-q)}{|U - c_k|^2} |\phi_k|^2\, dy \\ &\geq \alpha_k^2 - \sup_{E_\delta^c} \frac{|U''(U-q)|}{(U - c_k)^2} \int_{E_\delta^c} |\phi_k|^2\, dy. \end{aligned}$$

Since ϕ_k converges to $\phi_s \equiv 0$ uniformly on E_δ^c, we have

$$\lim_{k \to \infty} \inf J_q(\phi_k) \geq \alpha_s^2.$$

So for large k, $J_q(\phi_k) \neq 0$, which is a contradiction to Lemma 3.4.

So by Lemma 3.5, there is some z_i such that $\phi_s(z_i) \neq 0$. Then

$$\int_{E_\delta} \frac{U''(U-q)}{|U - c_s|^2} |\phi_s|^2\, dy \geq \int_{|y - z_i| < \delta} \frac{U''}{|U - c_s|^2} |\phi_s|^2\, dy = +\infty$$

since c_s is not an inflection value. By Fatou's lemma,

$$\lim_{k \to \infty} \inf \int_{E_\delta} \frac{U''(U-q)}{|U - c_k|^2} |\phi_k|^2\, dy = +\infty.$$

So from

$$J_q(\phi_k) \geq \int_{E_\delta} \frac{U''(U-q)}{|U - c_k|^2} |\phi_k|^2\, dy - \sup_{E_\delta^c} \frac{U''(U-q)}{|U - c_k|^2},$$

we get $\lim_k \inf J_q(\phi_k) = +\infty$, which is a contradiction to the fact that $J_q(\phi_k) = 0$ (Lemma 3.4). Thus c_s must be an inflection value. This ends the proof of Theorem 3.3. □

To show Theorem 1.7, we need to get some a priori estimate for the sequence of unstable solutions $\{\phi_k\}$ in Definition 1.6. We have the following.

LEMMA 3.7. *For the flow $U(y)$ in class $\mathcal{K}$, if ϕ is the solution to* (3) *with* $\operatorname{Im} c > 0$, *then we have*

$$\int_{y_1}^{y_2} \left(|\phi'|^2 + \alpha^2 |\phi|^2\right) dy < \int_{y_1}^{y_2} K(y) |\phi|^2 dy \tag{27}$$

and

$$\int_{y_1}^{y_2} \left(|\phi''|^2 + 2\alpha^2 |\phi'|^2 + \alpha^4 |\phi|^2\right) dy < \|K\|_\infty \int_{y_1}^{y_2} K(y) |\phi|^2 dy. \tag{28}$$

Proof. Inequality (27) was obtained in [8], but we prove it here for completeness. Denote $c = c_r + ic_i$ $(c_i > 0)$. By Lemma 3.4, for any real q

$$\int_{y_1}^{y_2} \left(|\phi'|^2 + \alpha^2 |\phi|^2 + \frac{U''(U-q)}{|U-c_r|^2 + c_i^2} |\phi|^2\right) dy = 0. \tag{29}$$

Let $q = U_s - 2(U_s - c_r)$. Then

$$\begin{aligned}
\int_{y_1}^{y_2} \left(|\phi'|^2 + \alpha^2 |\phi|^2\right) dy &= \int_{y_1}^{y_2} K(y) \frac{(U-U_s)(U-q)}{|U-c_r|^2 + c_i^2} |\phi|^2 dy \\
&= \int_{y_1}^{y_2} K(y) \frac{(U-U_s)^2 + 2(U-U_s)(U_s-c_r)}{|U-c_r|^2 + c_i^2} |\phi|^2 dy \\
&= \int_{y_1}^{y_2} K(y) \frac{(U-c_r)^2 - (U_s-c_r)^2}{|U-c_r|^2 + c_i^2} |\phi|^2 dy \\
&< \int_{y_1}^{y_2} K(y) |\phi|^2 dy.
\end{aligned}$$

This proves (27).

In (29), let $q = U_s$, we get by (27)

$$\int_{y_1}^{y_2} K(y) \frac{(U-U_s)^2}{|U-c_r|^2 + c_i^2} |\phi|^2 dy = \int_{y_1}^{y_2} \left(|\phi'|^2 + \alpha^2 |\phi|^2\right) dy < \int_{y_1}^{y_2} K(y) |\phi|^2 dy. \tag{30}$$

We shall show that

$$\int_{y_1}^{y_2} \left(|\phi''|^2 + 2\alpha^2 |\phi'|^2 + \alpha^4 |\phi|^2\right) dy - \int_{y_1}^{y_2} \frac{(U'')^2 |\phi|^2}{|U-c_r|^2 + c_i^2} dy = 0, \tag{31}$$

which was first proved in [2]. For completeness we now give the proof of (31). We multiply the Rayleigh equation

$$\left(\frac{d^2}{dy^2} - \alpha^2\right)\phi - \frac{U''}{U-c}\phi = 0$$

by $(\phi^*)''$ and integrate it to get

$$\int_{y_1}^{y_2} \left((\phi^*)'' \left(\phi'' - \alpha^2 \phi\right)\right) = \int_{y_1}^{y_2} \left((\phi^*)'' \frac{U''}{U-c}\phi\right). \tag{32}$$

By integration by parts,

$$LHS \text{ of } (32) = \int_{y_1}^{y_2} \left(|\phi''|^2 + \alpha^2 |\phi'|^2\right) dy.$$

Using the Rayleigh equation for ϕ^*, we have

$$\begin{aligned} RHS \text{ of } (32) &= \int_{y_1}^{y_2} \left(\left(\alpha^2\phi^* + \left(\frac{U''}{U-c}\phi\right)^*\right)\left(\frac{U''}{U-c}\phi\right)\right) dy \\ &= \alpha^2 \int_{y_1}^{y_2} \frac{U''|\phi|^2}{U-c} dy + \int_{y_1}^{y_2} \frac{(U'')^2 |\phi|^2}{|U-c_r|^2 + c_i^2} dy. \end{aligned}$$

So

$$\begin{aligned} \int_{y_1}^{y_2} \left(|\phi''|^2 + \alpha^2 |\phi'|^2\right) dy &= \operatorname{Re} RHS \\ &= \alpha^2 \int_{y_1}^{y_2} \frac{U''(U-c_r)|\phi|^2}{|U-c_r|^2 + c_i^2} dy + \int_{y_1}^{y_2} \frac{(U'')^2 |\phi|^2}{|U-c_r|^2 + c_i^2} dy \\ &= -\alpha^2 \left(\int_{y_1}^{y_2} \left(|\phi'|^2 + \alpha^2 |\phi|^2\right) dy\right) + \int_{y_1}^{y_2} \frac{(U'')^2 |\phi|^2}{|U-c_r|^2 + c_i^2} dy, \end{aligned}$$

by (29) with $q = c_r$. Now (31) follows.

Then inequality (28) follows easily from (5), (30), and (31). □

Remark 3.8 (stability). The inequality (27) was used in [8] to prove that there is no unstable solution to (3) when $\alpha \geq \alpha_{\max}$. Indeed, from (27), if there exists some solution ϕ with $\operatorname{Im} c > 0$, then

$$\begin{aligned} -\alpha^2 &> \frac{\int_{y_1}^{y_2} \left(|\phi'|^2 - K(y)|\phi|^2\right) dy}{\int_{y_1}^{y_2} |\phi|^2 dy} \\ &\geq \inf_{\phi \in \mathbf{H}_0^1(y_1,y_2)} \frac{\int_{y_1}^{y_2} \left(|\phi'|^2 - K(y)|\phi|^2\right) dy}{\int_{y_1}^{y_2} |\phi|^2 dy} = \alpha_{\max}^2. \end{aligned}$$

This proves that the condition in Theorem 1.2 is sharp for instability.

Proof of Theorem 1.7. Given (c_s, α_s, ϕ_s), let $\{(c_k, \alpha_k, \phi_k)\}$ (with $\operatorname{Im} c_k > 0$) be a sequence of solutions to the Rayleigh equation (3), as in Definition 1.6 of the introduction. Here we take $\|\phi_k\|_2 = 1$. By Theorem 3.3, $c_k \to U_s$. From Lemma 3.7, we get

$$\int_{y_1}^{y_2} \left(|\phi_k'|^2 + |\phi_k''|^2\right) dy < \max\left\{\|K\|_\infty^2, 1\right\}.$$

So there is a subsequence $\{\phi_{n_k}\}$ of $\{\phi_k\}$ and $\phi_0 \in \mathbf{H}^2 \cap \mathbf{H}_0^1(y_1, y_2)$ such that

$$\|\phi_{n_k} - \phi_0\|_{C^1} \to 0 \quad \text{and} \quad \|\phi_0\|_2 = 1.$$

Taking limits in

$$\left(\frac{d^2}{dy^2}-\alpha_{n_k}^2\right)\phi_{n_k}-\frac{U''}{U-c_{n_k}}\phi_{n_k}=0,$$

we get

$$-\frac{d^2}{dy^2}\phi_0+\frac{U''}{U-U_s}\phi_0=-\alpha_s^2\phi_0.$$

From the definition of ϕ_s, we have $\phi_s=\phi_0$ and thus the conclusion of Theorem 1.7 follows. □

THEOREM 3.9. *Let $U(y)$ be in class $\mathcal{K}$. Then the set Ξ of all unstable wave numbers is open. Any boundary point α of Ξ must satisfy the condition that $-\alpha^2$ is a negative eigenvalue of $-\frac{d^2}{dy^2}-K(y)$ in $\mathbf{H}^2\cap\mathbf{H}_0^1(y_1,y_2)$.*

Proof. If $\alpha\in\Xi$, then there exists c with $\operatorname{Im}c>0$ such that the Rayleigh equation (3) has some solution ϕ. Let

$$\psi=\left(\frac{d^2}{dy^2}-\alpha^2\right)\phi,\quad B_\alpha\psi:=U\psi-U''\left(\frac{d^2}{dy^2}-\alpha^2\right)^{-1}\psi.$$

Then from (3) we have $B_\alpha\psi=c\psi$. It is easy to see that $\sigma_{ess}(B_\alpha)=[U_{\min},U_{\max}]$. So c is some discrete eigenvalue of B_α. Since B_α is norm continuous in α, for any α' near α, there is also a complex c' in the spectrum of $B_{\alpha'}$. So Ξ is open. From the definition of neutral limiting modes, we know immediately that the boundary points of Ξ are neutral limiting wave numbers. Then the other conclusion in the theorem follows from Theorem 1.7. □

From Theorem 3.9, we know that in order to determine Ξ, we only need to know the instability property near any neutral limiting wave number. This is the basis of our method in the next section for obtaining a sufficient condition for instability.

4. Proof of Theorem 1.2. Let the steady flow $U(y)$ be in the class $\mathcal{K}^+$. To prove Theorem 1.2, we need to study the instability near each neutral limiting wave number. Tollmien [23] heuristically showed that unstable modes exist near a neutral mode for a symmetric flow in class $\mathcal{K}^+$. This was later reconsidered and the asymptotic growth rate was found by C. C. Lin [17]. However, the existence of unstable modes near a neutral mode had still not been rigorously proved. Another approach was recently given in [20] for a monotone flow in class $\mathcal{K}^+$, where the implicit function theorem was invoked to get existence. However, because the differentiability condition was only established on half of a neighborhood, the standard implicit function theorem does not apply. Moreover, the convergence to the neutral eigenfunction in their computation was not specified. Thus, as far as we are aware, a complete proof of Tollmien's argument does not yet exist.

Therefore in this section, we rigorously prove a perturbation result of Tollmien type for flows in class $\mathcal{K}^+$. The existence of an unstable mode is established when the wave number is slightly to the left of a neutral wave number.

THEOREM 4.1. *Suppose $U(y)$ is in class $\mathcal{K}^+$ and (ϕ_s,α_s,U_s) with $\alpha_s>0$ satisfies*

$$-\frac{d^2}{dy^2}\phi_s+\frac{U''}{U-U_s}\phi_s=-\alpha_s^2\phi_s \tag{33}$$

with $\phi_s(y_1) = \phi_s(y_2) = 0$. *Then there exists* $\varepsilon_0 < 0$ *such that if* $\varepsilon_0 < \varepsilon < 0$, *there is a nontrivial solution* ϕ_ε *to the Rayleigh equation*

$$(U - U_s - c(\varepsilon))\left(\frac{d^2}{dy^2} - \alpha(\varepsilon)^2\right)\phi_\varepsilon - U''\phi_\varepsilon = 0$$

with $\phi_\varepsilon(y_1) = \phi_\varepsilon(y_2) = 0$. *Here* $\alpha(\varepsilon) = \sqrt{\varepsilon + \alpha_s^2}$ *is the perturbed wave number and* $U_s + c(\varepsilon)$ *is an unstable eigenvalue with* $\operatorname{Im} c(\varepsilon) > 0$. *Moreover, the function* $c(\varepsilon)$ *is differentiable in* $(\varepsilon_0, 0)$ *and*

$$\lim_{\varepsilon \to 0-} c(\varepsilon) = 0, \tag{34}$$

$$\lim_{\varepsilon \to 0-} c'(\varepsilon) = \frac{\int_{y_1}^{y_2} \phi_s^2(y)\, dy}{i\pi \sum_{k=1}^{l} \left(|U'|^{-1} K\phi_s^2\right)|_{y=a_k} + \mathcal{P}\int_{y_1}^{y_2} \left(K(y)\phi_s^2(y)\right)/(U(y) - U_s)\, dy}, \tag{35}$$

where $a_1, \ldots, a_l$ *are the inflection points such that* $U(a_k) = U_s$, $k = 1, \ldots, l$, *and* $\mathcal{P}\int_{y_1}^{y_2}$ *denotes the Cauchy principal part.*

Remark 4.2. As mentioned in Remark 3.2, the number of points where U takes the value U_s is finite. In formula (35), we have

$$\sum_{k=1}^{l} \left(|U'|^{-1} K\phi_s^2\right)|_{y=a_k} > 0.$$

This is due to the following two facts:

(a) The function ϕ_s must be nonzero at at least one of the points a_k. This is a corollary of Lemma 3.5, where $c = U_s$ and $z_j = a_j$.

(b) We have $U'(a_k) \neq 0$ for each k. Otherwise there exists some k such that $U'(a_k) = 0$. Then it is easy to see that $K(a_k) = \infty$, which is contradictory to our assumption that K is bounded.

Proof of Theorem 4.1. Define $\phi_1(y; c, \varepsilon)$ and $\phi_2(y; c, \varepsilon)$ to be the solutions of

$$-\frac{d^2}{dy^2}\phi + \frac{U''}{U - U_s - c}\phi + \left(\alpha_s^2 + \varepsilon\right)\phi = 0, \tag{36}$$

with $\phi_1(y_1) = 0$, $\phi_1'(y_1) = \phi_s'(y_1)$ and $\phi_2(y_1) = -\frac{1}{\phi_s'(y_1)}$, $\phi_2'(y_1) = 0$. Here $\varepsilon < 0$ and $\operatorname{Im} c > 0$. Then ϕ_1, ϕ_2 are analytic in the upper half-plane as a function of c and ϕ_1, ϕ_2 are independent with Wronskian 1. Now define $I(c, \varepsilon) = \phi_1(y_2; c, \varepsilon)$. The existence of a solution to the Rayleigh equation is equivalent to the existence of a root of I with $\operatorname{Im} c > 0$. It will be proved by a modified Newton method, i.e., by finding a fixed point of

$$c \to c - \frac{I(c, \varepsilon)}{\partial I/\partial c\,|_{(c,\varepsilon)=(0,0)}}\ .$$

Letting

$$N(t, y; \varepsilon, c) = \phi_1(t; \varepsilon, c)\,\phi_2(y; \varepsilon, c) - \phi_2(t; \varepsilon, c)\,\phi_1(y; \varepsilon, c),$$

we will show that

$$\frac{\partial I}{\partial \varepsilon} = \int_{y_1}^{y_2} N\left(y, y_2; \varepsilon, c\right) \phi_1\left(y; c, \varepsilon\right) dy \tag{37}$$

and

$$\frac{\partial I}{\partial c} = \int_{y_1}^{y_2} N\left(y, y_2; \varepsilon, c\right) \frac{U''(y)}{\left(U(y) - U_s - c\right)^2} \phi_1\left(y; c, \varepsilon\right) dy. \tag{38}$$

In order to prove (37) and (38), notice that for (c', ε') close to (c, ε) with $\operatorname{Im} c' > 0$, the function $\phi_1\left(y; c', \varepsilon'\right)$ satisfies

$$\begin{aligned} -\frac{d^2}{dy^2}\phi + \frac{U''}{U - U_s - c}\phi + \left(\alpha_s^2 + \varepsilon\right)\phi \\ = \left[\frac{-U''(y)\left(c' - c\right)}{\left(U(y) - U_s - c\right)\left(U(y) - U_s - c'\right)} - \left(\varepsilon' - \varepsilon\right)\right]\phi. \end{aligned}$$

So

$$\begin{aligned} \phi_1\left(y; c', \varepsilon'\right) = \phi_1\left(y; c, \varepsilon\right) \\ - \int_{y_1}^{y} N\left(t, y; \varepsilon, c\right) \left[\frac{-U''(t)\left(c' - c\right)}{\left(U(t) - U_s - c\right)\left(U(t) - U_s - c'\right)} - \left(\varepsilon' - \varepsilon\right)\right] \phi_1\left(t; c', \varepsilon'\right) dt. \end{aligned}$$

Thus, letting $y = y_2$,

$$\begin{aligned} I\left(c', \varepsilon'\right) = I\left(c, \varepsilon\right) \\ + \int_{y_1}^{y_2} N\left(t, y_2; \varepsilon, c\right) \left[\frac{U''(t)\left(c' - c\right)}{\left(U(t) - U_s - c\right)\left(U(t) - U_s - c'\right)} + \left(\varepsilon' - \varepsilon\right)\right] \phi_1\left(t; c', \varepsilon'\right) dt. \end{aligned}$$

Identities (37) and (38) follow from this identity by letting (c', ε') tend to (c, ε).

Now define the triangle

$$\Delta_{(R,b)} = \left\{c_r + i c_i \mid \ |c_r| < R c_i, 0 < c_i < b\right\}$$

and the Cartesian product

$$E_{(R,b_1,b_2)} = \Delta_{(R,b_1)} \times \left(-b_2, 0\right),$$

where $b_1, b_2 > 0$.

We make the following claims:

(a) For fixed R, $(c, \varepsilon) \in E_{(R,b_1,b_2)}$, $\phi_1\left(y; c, \varepsilon\right)$ uniformly converges to $\phi_s(y)$ in $C^1[y_1, y_2]$ as $c \to 0$, $\varepsilon \to 0-$. That is, for any $\delta > 0$, there exists some $b_0 > 0$ such that

$$\left\|\phi_1\left(y; c, \varepsilon\right) - \phi_s(y)\right\|_{C^1} \leq \delta$$

for $b_1, b_2 < b_0$, $(c, \varepsilon) \in E_{(R,b_1,b_2)}$.

(b) The same conclusion holds true for $\phi_2\left(y; c, \varepsilon\right)$. We denote $\phi_2\left(y; 0, 0\right) = \phi_z(y)$, so that $\phi_z\left(y_2\right) = -\frac{1}{\phi_s'(y_2)}$. Then $\phi_2\left(y; c, \varepsilon\right)$ uniformly converges to $\phi_z(y)$ in $C^1[y_1, y_2]$ for $(c, \varepsilon) \in E_{(R,b_1,b_2)}$, $c \to 0$, $\varepsilon \to 0-$.

Proof of claim (a). Indeed, if it is not true, then there exists $\delta_0 > 0$ and a sequence $\{(c_k, \varepsilon_k)\}_{k=1}^{\infty}$, $(c_k, \varepsilon_k) \to (0,0)$, $|\operatorname{Re} c_k| < R \operatorname{Im} c_k$ such that

$$\|\phi_1(y; c_k, \varepsilon_k) - \phi_s(y)\|_{C^1} \geq \delta_0.$$

Since $|\operatorname{Re} c_k| < R \operatorname{Im} c_k$ and $\operatorname{Im} c_k < b_0$, we have

$$\left|\frac{U''(y)}{U(y) - U_s - c_k}\right| \leq |K(y)| + |K(y)| \left|\frac{c_k}{U(y) - U_s - c_k}\right| \leq |K(y)| \left(1 + \sqrt{R^2+1}\right).$$

Thus

$$\left\|\frac{U''(y)}{U(y) - U_s - c_k}\right\|_{\infty} \leq \|K\|_{\infty} \left(1 + \sqrt{R^2+1}\right). \tag{39}$$

Let $\phi_k = \phi_1(y; c_k, \varepsilon_k)$; then we have uniform bound for $\|\phi_k\|_{C^2}$ because ϕ_k satisfies an ODE (36) with uniformly bounded coefficients and the same initial value. So by the Ascoli–Arzelà lemma, there is a subsequence $\{\phi_{k_i}\}$ and a function $\phi_0 \in C^1[y_1, y_2]$ such that

$$\|\phi_{k_i} - \phi_0\|_{C^1} \to 0$$

as $k_i \to \infty$. Since ϕ_{k_i} satisfies Rayleigh's equation, ϕ_0 satisfies

$$-\frac{d^2}{dy^2}\phi_0 + \frac{U''}{U - U_s}\phi_0 = -\alpha_s^2 \phi_0,$$

with $\phi_0(y_1) = 0, \phi_0'(y_1) = \phi_s'(y_1)$; thus $\phi_0 = \phi_s$. So $\|\phi_{k_i} - \phi_s\|_{C^1} \to 0$, which is a contradiction to our assumption. Claim (b) follows similarly.

In the appendix we prove that

$$\frac{\partial I}{\partial \varepsilon} \to -\frac{1}{\phi_s'(y_2)} \int_{y_1}^{y_2} \phi_s^2(y)\, dy \tag{40}$$

and

$$\frac{\partial I}{\partial c} \to \frac{1}{\phi_s'(y_2)} \left(i\pi \sum_{k=1}^{l} \left(|U'|^{-1} K \phi_s^2 \right)|_{y=a_k} + \mathcal{P} \int_{y_1}^{y_2} \left(K(y) \phi_s^2(y) \right) / \left(U(y) - U_s \right) dy \right) \tag{41}$$

uniformly in $E_{(R,b_1,b_2)}$ as $c \to 0, \varepsilon \to 0-$. Denote these limits by

$$B = -\frac{1}{\phi_s'(y_2)} \int_{y_1}^{y_2} \phi_s^2(y)\, dy,$$

$$C = \frac{1}{\phi_s'(y_2)} \mathcal{P} \int_{y_1}^{y_2} \left(K(y) \phi_s^2(y) \right) / \left(U(y) - U_s \right) dy,$$

$$D = \frac{\pi}{\phi_s'(y_2)} \sum_{k=1}^{l} \left(|U'|^{-1} K \phi_s^2 \right)|_{y=a_k},$$

where a_k $(k = 1, \ldots, l)$ are the inflection points. Denote

$$f(c, \varepsilon) = I(c, \varepsilon) - B\varepsilon - (C + Di)\, c$$

and

$$h(c,\varepsilon) = -\frac{B}{C+iD}\varepsilon - \frac{f(c,\varepsilon)}{C+iD}.$$

Then by the uniform convergence of (40) and (41), for any $\delta_0 > 0$, there exists b_0 so that when $b_1, b_2 < b_0$ we have

$$\left|\frac{\partial f}{\partial c}\right|, \left|\frac{\partial f}{\partial \varepsilon}\right| < \delta_0 \quad \forall (c,\varepsilon) \in E_{(R,b_1,b_2)}. \tag{42}$$

So for any (c,ε), (c',ε') in the convex set $E_{(R,b_1,b_2)}$,

$$|f(c,\varepsilon) - f(c',\varepsilon')| \le \delta_0 (|\varepsilon - \varepsilon'| + |c - c'|). \tag{43}$$

Now in (43) we let $(c',\varepsilon') \to (0,0)$ and notice that

$$\lim_{(c',\varepsilon')\to(0,0)} f(c',\varepsilon') = \lim_{(c',\varepsilon')\to(0,0)} I(c',\varepsilon') = \lim_{(c',\varepsilon')\to(0,0)} \phi_1(y_2; c',\varepsilon') = \phi_s(y_2) = 0,$$

so we obtain

$$|f(c,\varepsilon)| \le \delta_0 (|\varepsilon| + |c|) \ \forall (c,\varepsilon) \in E_{(R,b_1,b_2)}. \tag{44}$$

Note that for fixed ε, a zero of $I(c,\varepsilon)$ is a fixed point of $c \to h(c,\varepsilon)$. Let $R = 4\left|\frac{C}{D}\right|$ if $C \neq 0$ and $R = 1$ if $C = 0$. Notice that by Remark 4.2, $BD < 0$. Denote

$$Q = \sqrt{R^2+1}\frac{-2DB}{C^2+D^2} + 1.$$

Let

$$\delta_0 = \frac{1}{2}\min\left\{\frac{|BC|}{Q(C^2+D^2)}, \frac{-BD}{Q(C^2+D^2)}, 1\right\}\sqrt{C^2+D^2}$$

if $C \neq 0$ and

$$\delta_0 = \frac{1}{2}\min\left\{\frac{-BD}{Q(C^2+D^2)}, 1\right\}\sqrt{C^2+D^2}$$

if $C = 0$. There exists b_0 such that if $b_1, b_2 < b_0$, then (44) and (42) hold. We choose

$$b_2 = \min\left\{\frac{C^2+D^2}{-2DB\sqrt{R^2+1}}, 1\right\} b_0, \qquad b_1 = Qb_2.$$

Fix $\varepsilon \in (-b_2, 0)$ and let

$$b(\varepsilon) = \frac{-2DB}{C^2+D^2}\varepsilon.$$

We will prove that

$$h(\cdot,\varepsilon) : \Delta_{(R,b(\varepsilon))} \to \Delta_{(R,b(\varepsilon))} \text{ is a contraction map,} \tag{45}$$

with contraction ratio no greater than $\frac{1}{2}$ for all $-b_2 < \varepsilon < 0$.

Assuming (45), the theorem follows easily. Indeed, for each $\varepsilon \in (-b_2, 0)$ there exists a unique $c(\varepsilon) \in \Delta_{(R,b(\varepsilon))}$ so that $h(c(\varepsilon), \varepsilon) = c(\varepsilon)$. Since for fixed ε, $h(c, \varepsilon)$ is analytic in $\Delta_{(R,b_1)}$ and uniformly contracting, we know that $c(\varepsilon)$ is the unique fixed point in $\Delta_{(R,b_1)}$ and is differentiable with respect to ε in the interval $(-b_2, 0)$ (see [5, p. 25]). We now let $\varepsilon_0 = -b_2$. Since $c(\varepsilon) \in \Delta_{(R,b(\varepsilon))}$, we have

$$\lim_{\varepsilon \to 0-} c(\varepsilon) = 0.$$

From $I(c(\varepsilon), \varepsilon) = 0$, we obtain

$$c'(\varepsilon) = -\frac{\partial I/\partial \varepsilon}{\partial I/\partial c}.$$

So by (40) and (41), we have

$$\lim_{\varepsilon \to 0-} c'(\varepsilon) = \frac{\int_{y_1}^{y_2} \phi_s^2(y)\, dy}{i\pi \sum_{k=1}^{l} \left(|U'|^{-1} K \phi_s^2 \right) |_{y=a_k} + \mathcal{P} \int_{y_1}^{y_2} \left(K(y) \phi_s^2(y) \right) / \left(U(y) - U_s \right) dy}.$$

This proves (35), and the proof of Theorem 4.1 is complete assuming (45).

Now we prove (45). By our choices of δ_0, b_0, b_1, b_2, and (42), we know

$$\left| \frac{\partial h}{\partial c} \right| = \frac{1}{\sqrt{C^2 + D^2}} \left| \frac{\partial f}{\partial c} \right| \le \frac{1}{\sqrt{C^2 + D^2}} \delta_0 \le \frac{1}{2} \ \forall (c, \varepsilon) \in E_{(R,b_1,b_2)}.$$

Thus h is uniformly contracting with ratio no greater than $\frac{1}{2}$ for each fixed $\varepsilon \in (-b_2, 0)$. We still need to show that $h(c, \varepsilon)$ maps $\Delta_{(R,b(\varepsilon))}$ to itself. If $C \neq 0$, by (44) and the definitions of $b(\varepsilon), Q$, and δ_0, we have

$$\left| \frac{f(c, \varepsilon)}{C + iD} \right| \le \delta_0 \frac{|c| + |\varepsilon|}{\sqrt{C^2 + D^2}} \le \frac{\delta_0}{\sqrt{C^2 + D^2}} \left(1 + \sqrt{R^2 + 1} \frac{-2DB}{C^2 + D^2} \right) |\varepsilon| = \frac{\delta_0 |\varepsilon| Q}{\sqrt{C^2 + D^2}}$$

$$\le \frac{1}{2} \min \left\{ \frac{|BC|}{C^2 + D^2}, \frac{-BD}{C^2 + D^2} \right\} |\varepsilon|. \tag{46}$$

Substituting (46) into

$$\operatorname{Re} h = \frac{-BC}{C^2 + D^2} \varepsilon - \operatorname{Re} \frac{f(c, \varepsilon)}{C + iD},$$

we readily get

$$\frac{1}{2} \frac{|BC|}{C^2 + D^2} |\varepsilon| \le |\operatorname{Re} h| \le 2 \frac{|BC|}{C^2 + D^2} |\varepsilon|. \tag{47}$$

In the same way we get

$$\frac{1}{2} \frac{-BD}{C^2 + D^2} |\varepsilon| \le \operatorname{Im} h \le 2 \frac{-BD}{C^2 + D^2} |\varepsilon| = b(\varepsilon). \tag{48}$$

Combining (47) and (48), we have

$$|\operatorname{Re} h| \le 4 \left| \frac{C}{D} \right| \operatorname{Im} h = R \operatorname{Im} h.$$

So $h \in \Delta_{(R,b(\varepsilon))}$. The proof for the case $C = 0$ is the same. This proves (45), and thus the proof of Theorem 4.1 is complete. □

Proof of Theorem 1.2. Let $-\alpha_m^2 < -\alpha_{m-1}^2 < \cdots < -\alpha_1^2 < 0$ be all the negative eigenvalues of $-\frac{d^2}{dy^2} + K(y)$. Here $\alpha_m = \alpha_{\max}$ as defined by (7). Combining Theorems 3.9 and 4.1, we deduce that if $\alpha \in (0, \alpha_m)$ and $\alpha \neq \alpha_i \, (i = 1, \ldots, m)$, then there exists an unstable mode.

Now we investigate the possibility of an instability at $\alpha = \alpha_i \, (i = 1, \ldots, m)$. For each $\alpha \in (\alpha_i, \alpha_{i+1})$, we know that there exists some unstable eigenvalue $c(\alpha) = c_r(\alpha) + ic_i(\alpha)$ with $c_i > 0$. We claim that

$$\text{as } \alpha \to \alpha_i+, \ c_i(a) \text{ has some lower bound } \delta > 0. \tag{49}$$

Assuming (49), we now show the existence of an unstable eigenvalue at α_i. We take a sequence $\{(c_k, \alpha_k, \phi_k)\}_{k=1}^{\infty}$ with $\alpha_k \to \alpha_i+$ and $\operatorname{Im} c_k \geq \delta > 0$. The function ϕ_k with $\|\phi_k\|_2 = 1$ satisfies the Rayleigh equation

$$-\frac{d^2}{dy^2}\phi_k + \frac{U''}{U - c_k}\phi_k = -\alpha_k^2 \phi_k. \tag{50}$$

By Lemma 3.7, there is an a priori bound for $\|\phi_k\|_{\mathbf{H}^2}$, so there exists some nonzero function $\phi_0 \in \mathbf{H}^2$ such that $\phi_k \to \phi_0$ strongly in $\mathbf{H}^1$. Note that c_k is bounded by (4). Suppose $c_k \to c_0$ with $\operatorname{Im} c_0 \geq \delta$. Now

$$\left\| \frac{U''}{U - c_k} \right\|_{\infty} \leq \frac{\|U''\|_{\infty}}{\delta},$$

so we can pass to the limit in (50) to deduce that ϕ_0 is a weak solution to

$$-\frac{d^2}{dy^2}\phi_0 + \frac{U''}{U - c_0}\phi_0 = -\alpha_i^2 \phi_0.$$

Since $\operatorname{Im} c_0 > 0$, $\frac{U''}{U - c_0}$ is a smooth function. So by elliptic regularity theory, ϕ_0 is a classical solution. Thus at $\alpha = \alpha_i$, we get an unstable eigenvalue c_0.

Proof of (49). If it is not true, then there exists a sequence $\{(c_k, \alpha_k, \phi_k)\}_{k=1}^{\infty}$ of solutions to Rayleigh's equation, with $\alpha_k \to \alpha_i+$ and $\operatorname{Re} c_k \to c_s$, $\operatorname{Im} c_k \to 0+$. By Theorem 1.7, c_s must equal U_s. From the proof of Theorem 1.7, we know that $\phi_k \to \phi_s$ in $C^1[y_1, y_2]$, where ϕ_s is a solution to

$$-\frac{d^2}{dy^2}\phi_s + \frac{U''}{U - U_s}\phi_s = -\alpha_i^2 \phi_s. \tag{51}$$

Multiplying (51) by ϕ_k and subtracting ϕ_s times (50), then integrating from y_1 to y_2, we get

$$\left(\alpha_k^2 - \alpha_i^2\right) \int_{y_1}^{y_2} \phi_s \phi_k dy = -\left(c_k - U_s\right) \int_{y_1}^{y_2} \frac{U'' \phi_s \phi_k}{\left(U - c_k\right)\left(U - U_s\right)} dy.$$

Let

$$A_k = \int_{y_1}^{y_2} \phi_s \phi_k dy, \ B_k = -\int_{y_1}^{y_2} \frac{U'' \phi_s \phi_k}{\left(U - c_k\right)\left(U - U_s\right)} dy.$$

Then

$$\lim_{k\to\infty} A_k = \int_{y_1}^{y_2} |\phi_s|^2 \, dy. \tag{52}$$

In the appendix we will prove

$$\lim_{k\to\infty} B_k = \mathcal{P}\int_{y_1}^{y_2} \frac{K(y)\,\phi_s^2}{(U-U_s)} dy + i\pi \sum_{k=1}^{l} \left(|U'|^{-1} K\phi_s^2 \right) |_{y=a_k}. \tag{53}$$

Now we have

$$\lim_{k\to\infty} \frac{A_k}{B_k} = \frac{\int_{y_1}^{y_2} |\phi_s|^2 \, dy}{\mathcal{P}\int_{y_1}^{y_2} \frac{K(y)\phi_s^2}{(U-U_s)} dy + i\pi \sum_{k=1}^{l} \left(|U'|^{-1} K\phi_s^2 \right) |_{y=a_k}} = a + ib$$

with $b < 0$. Thus if k is large enough,

$$\operatorname{Im} c_k = \left(\alpha_k^2 - \alpha_i^2\right) \operatorname{Im} \frac{A_k}{B_k} < 0,$$

which is a contradiction. So (49) is proved and the proof of Theorem 1.2 is complete. □

We also have the following result about the instability at $\alpha = 0$.

LEMMA 4.3. *If $U(y_1) \neq U(y_2)$, then at $\alpha = 0$ there is also some unstable solution to the Rayleigh equation.*

Proof. Let $\{(c_k, \alpha_k, \phi_k)\}_{k=1}^{\infty}$ be a sequence of unstable solutions with $\alpha_k \to 0+$. It suffices to prove that there is some positive lower bound for $\{\operatorname{Im} c_k\}$. Indeed the existence of an unstable solution at $\alpha = 0$ would follow by the same argument as in the proof of Theorem 1.2.

Assume there is no lower bound. Then $\operatorname{Im} c_k \to 0$, $\operatorname{Re} c_k \to c$. Then ϕ_k converges to a neutral solution $\phi_0 \in \mathbf{H}^2 \cap \mathbf{H}_0^1$ satisfying equation

$$(U-c)\,\phi_0'' - U''\phi_0 = 0 \tag{54}$$

sectionally in each (z_i, z_{i+1}). Here we use the same notation as immediately before Lemma 3.5. We now show that ϕ_0 cannot vanish at any zero $z_1, \ldots, z_{k_c}$ of $U - c$. Indeed, if it is not true, we suppose $\phi_0(z_i) = 0$ and consider (54) in (z_i, z_{i+1}). Then

$$\left((U-c)\,\phi_0' - U'\phi_0\right)(y) \equiv \left((U-c)\,\phi_0' - U'\phi_0\right)(z_i) = 0$$

for all y in the interval (z_i, z_{i+1}). So ϕ_0 and $U - c$ are linearly dependent in (z_i, z_{i+1}). Thus $\phi_0(z_{i+1}) = 0$. Repeating the process, we know that $\phi_0(z_i) = 0$ for all $i = 1, \ldots, k_c$ and there is some constant b such that $U(y) - c = b\phi_0(y)$ for all y in (y_1, y_2). This implies that $U(y_1) = U(y_2) = 0$, which is a contradiction.

Thus ϕ_0 takes a nonzero value at each zero of $U - c$ and ϕ_0 is the limit of unstable eigenfunctions. By the proof of Theorem 3.3 we know that c must equal U_s. Now by the argument in the last part of the proof of Theorem 1.2, we know that there is no perturbation of the neutral mode at $\alpha = 0$ to its right neighborhood. This contradiction shows that the $\operatorname{Im} c_k$ has some positive lower bound. The proof of the lemma is finished. □

Remark 4.4. If $U(y_1) = U(y_2)$, it is possible that at $\alpha = 0$ there is no unstable solution to (3). One such example is $U(y) = \cos 6y$, whose complete spectrum was found in [12] and for which there is no growing mode at $\alpha = 0$.

5. Rotating flows. In this section, we consider the radially symmetric steady flows in an annulus $0 < R_1 \le r \le R_2$. Using polar coordinates (r, θ), we rewrite the vorticity equation (1) as

$$\partial_t \Delta\psi + \frac{1}{r}\frac{\partial\psi}{\partial\theta}\frac{\partial}{\partial r}\Delta\psi - \frac{\partial\psi}{\partial r}\frac{1}{r}\frac{\partial}{\partial\theta}\Delta\psi = 0.$$

Here ψ is the stream function and

$$(u_r, u_\theta) = \left(\frac{1}{r}\frac{\partial\psi}{\partial\theta}, -\frac{\partial\psi}{\partial r}\right), \qquad \omega = -\Delta\psi = \frac{1}{r}\frac{\partial}{\partial r}(ru_\theta) - \frac{1}{r}\frac{\partial u_r}{\partial\theta}$$

are the velocity and vorticity, respectively. And ψ is constant on $r = R_j$ $(j = 1, 2)$. The steady flow is $(u_r, u_\theta) = (0, r\Omega)$, with $\Omega = \Omega(r)$ the steady angular velocity. The linearized equation about this steady flow is

$$\partial_t \Delta\tilde{\psi} + \Omega\frac{\partial}{\partial\theta}\Delta\tilde{\psi} + \left(\frac{\partial}{\partial r}Z\right)\frac{1}{r}\frac{\partial\tilde{\psi}}{\partial\theta} = 0, \tag{55}$$

with $\tilde{\psi}$ constant on $r = R_j$ $(j = 1, 2)$ and the steady vorticity $Z = 2\Omega + r\frac{d\Omega}{dr}$. Taking $\tilde{\psi}(r, \theta, t) = \phi(r)\exp(st + in\theta)$ and letting $D_* = \frac{d}{dr} + \frac{1}{r}$, $D = \frac{d}{dr}$ (following the notation in [6]), we rewrite (55) as

$$(s + in\Omega)\left(D_*D - n^2/r^2\right)\phi - inr^{-1}(rD^2\Omega + 3D\Omega)\phi = 0, \tag{56}$$

with $\phi(R_1) = \phi(R_2) = 0$ and n a positive integer. Letting $c = \frac{s}{in}$, we get the rotating Rayleigh equation (8). Instability would mean that there exists a solution to (8) with $\operatorname{Im} c > 0$. In this section we study the flows such that the function $K(r)$ defined by (9) is positive and bounded, which we still denote by class $\mathcal{K}^+$.

We are interested only in the case when α is a positive integer of the following extended Rayleigh equation:

$$(\Omega - c)\left(D_*D - \alpha^2/r^2\right)\phi - r^{-1}(rD^2\Omega + 3D\Omega)\phi = 0, \tag{57}$$

with $\phi(R_1) = \phi(R_2) = 0$. However, by embedding the original problem into a family of problems (57) depending on a continuous positive parameter α, we can use the same idea as in the shear flow case. For that purpose, first we need to prove the rotating versions of some results used in the shear flow case. We give detailed proofs only when they are really different. First is the extension of Lemma 3.5 to the rotating case. Let $r_1 < r_2 < \cdots < r_{k_c}$ be the zeros of $\Omega(y) - c$ and let S_0 be the complement of the set of points $\{r_i\}$ in the interval $[R_1, R_2]$. Here c is any real number in the range of $\Omega(y)$. Let $r_0 = R_1$ and $r_{k_c+1} = R_2$. Note that for the rotating flows in class $\mathcal{K}^+$, k_c is finite for any c by the same argument as in Remark 3.2.

LEMMA 5.1. *Let ϕ satisfy (57) on S_0 with real $\alpha > 1$ and real c in the range of Ω. We assume ϕ is sectionally continuous on the open intervals (r_j, r_{j+1}), $j = 0, 1, \ldots, k_c$. Then ϕ cannot vanish at both endpoints of any of the intervals (r_j, r_{j+1}) unless it vanishes identically in that interval.*

Proof. The function ϕ satisfies

$$(\Omega - c)\left(D_*D - \alpha^2/r^2\right)\phi - r^{-1}(rD^2\Omega + 3D\Omega)\phi = 0,$$

which is the same as

$$(\Omega - c)D^2\phi - \phi D^2\Omega + (\Omega - c)r^{-1}D\phi - 3r^{-1}\phi D\Omega = (\Omega - c)\frac{\alpha^2}{r^2}\phi.$$

We multiply both sides of the above by r^2 to get

$$(\Omega - c)r^2 D^2\phi - r^2\phi D^2\Omega + (\Omega - c)rD\phi - 3r\phi D\Omega = (\Omega - c)\alpha^2\phi. \tag{58}$$

We have

$$\begin{aligned} LHS &= (\Omega - c)\left[D^2\left(r^2\phi\right) - 4rD\phi - 2\phi\right] - r^2\phi D^2\Omega + (\Omega - c)\, rD\phi - 3r\phi D\Omega \\ &= (\Omega - c)D^2\left(r^2\phi\right) - r^2\phi D^2\Omega - 3\left(D\left(r\phi\right)(\Omega - c) + r\phi D\Omega\right) + \phi(\Omega - c) \\ &= D\left((\Omega - c)D\left(r^2\phi\right) - r^2\phi D\Omega\right) - 3D\left(\left(r\phi\right)(\Omega - c)\right) + \phi(\Omega - c). \end{aligned}$$

So (58) becomes (using $'$ to replace D)

$$\left((\Omega - c)\left(r^2\phi\right)' - r^2\phi\Omega'\right)' - 3\left(\left(r\phi\right)(\Omega - c)\right)' = \left(\alpha^2 - 1\right)\phi(\Omega - c). \tag{59}$$

Suppose ϕ vanishes at r_i and r_{i+1}. Here we mean $\phi\left(r_i+\right)$ for $\phi\left(r_i\right)$ and $\phi\left(r_{i+1}-\right)$ for $\phi\left(r_{i+1}\right)$ when studying (59) in $[r_i, r_{i+1}]$. From the definition of r_i we know that $\Omega - c$ has constant sign in (r_i, r_{i+1}). If $r_i \neq R_1\, (i \neq 0)$, then $\Omega\left(r_i\right) - c = 0$. Let $\tilde{r} \leq r_{i+1}$ be the nearest zero of ϕ in $(r_i, r_{i+1}]$. Since (59) is a real equation, we may assume ϕ is real and nonnegative on the interval $(r_i, \tilde{r})$ and that $\left(r^2\phi\right)'\left(r_i\right) \geq 0$ and $\left(r^2\phi\right)'(\tilde{r}) \leq 0$. Integrating (59) over $\left(r_i, \tilde{r}\right)$, we get

$$\left(\Omega(\tilde{r}) - c\right)\left(r^2\phi\right)'(\tilde{r}) = \left(\alpha^2 - 1\right)\int_{r_i}^{\tilde{r}} \phi(\Omega - c)dr',$$

since ϕ vanishes at the endpoints $r_i, \tilde{r}$.

If $\tilde{r} = r_{i+1,}$ then the left-hand side above must be zero. Hence ϕ is identically zero on (r_i, r_{i+1}). On the other hand, if $\tilde{r} < r_{i+1}$, then $\Omega(\tilde{r}) \neq c$ and

$$\left(r^2\phi\right)'(\tilde{r}) = \left(\alpha^2 - 1\right)\int_{r_i}^{\tilde{r}} \frac{\left(\Omega\left(r'\right) - c\right)}{\left(\Omega(\tilde{r}) - c\right)}\phi\left(r'\right)dr',$$

which could not hold true unless $\phi = 0$ on $[r_i, \tilde{r}]$. But the second order ODE (58) is regular on (r_i, r_{i+1}); thus $\tilde{r}$ could not be a cluster point of nontrivial ϕ. Thus ϕ must be identically zero on $\left(r_i, r_{i+1}\right)$.

If $i = 0$, then we repeat the same argument with the right endpoint of the interval $\left(R_1, r_1\right)$. □

We need Howard's semicircle theorem for the rotating case, which seems not to have been proven in the literature. So we give a proof here.

LEMMA 5.2. *If $\alpha > 1$, then for the extended Rayleigh equation* (57) *to have a solution, c (with $\operatorname{Im} c > 0$) must lie in the semicircle*

$$\left(c_r - \frac{1}{2}\left(\Omega_{\min} + \Omega_{\max}\right)\right)^2 + c_i^2 \leq \left(\frac{1}{2}\left(\Omega_{\min} - \Omega_{\max}\right)\right)^2, \tag{60}$$

where $\Omega_{\min}$ and $\Omega_{\max}$ are the minimum and maximum of $\Omega(r)$ in $[R_1, R_2]$.

Proof. Let ϕ be a solution to (57). As in the proof of the last lemma, ϕ satisfies (59), which we can rewritten as

$$\left(r\left(\phi'(\Omega - c)r - \left((\Omega - c)r\right)'\phi\right)\right)' = \left(\alpha^2 - 1\right)\phi(\Omega - c). \tag{61}$$

The above identity is equivalent to

$$\left(r^3(\Omega-c)^2\left(\frac{\phi}{(\Omega-c)r}\right)'\right)' = \left(\alpha^2-1\right)\phi(\Omega-c). \tag{62}$$

Now $\psi = \frac{\phi}{(\Omega-c)r}$ is a regular function since $\operatorname{Im} c \neq 0$. Then (62) becomes

$$\left(r^3(\Omega-c)^2\psi\right)' = \left(\alpha^2-1\right) r(\Omega-c)^2\psi. \tag{63}$$

Multiplying (63) by ψ^* (conjugate of ψ) and integrating it, we obtain

$$\int_{R_1}^{R_2}(\Omega-c)^2\left(r^3\left|\psi'\right|^2+\left(\alpha^2-1\right)r\left|\psi\right|^2\right)dr = 0. \tag{64}$$

The rest of the proof is the same as in the case of shear flows [14], [6]. We repeat it here for completeness. Let

$$P \equiv r^3\left|\psi'\right|^2+\left(\alpha^2-1\right)r\left|\psi\right|^2.$$

Then (64) becomes

$$\int_{R_1}^{R_2}(\Omega-c)^2 P\ dr = 0. \tag{65}$$

The function P is nonnegative and not identically zero. Comparing the real and imaginary parts of (65), we get

$$\int_{R_1}^{R_2}\left((\Omega-c_r)^2-c_i^2\right)P\ dr = 0 \quad \text{and} \quad 2c_i\int_{R_1}^{R_2}(\Omega-c_r)P\ dr = 0. \tag{66}$$

Observe that

$$\begin{aligned} 0 &\geq \int_{R_1}^{R_2}\left(\Omega-\Omega_{\min}\right)\left(\Omega-\Omega_{\max}\right)P\ dr \\ &= \int_{R_1}^{R_2}\left\{\left(c_r^2+c_i^2\right)-\left(\Omega_{\min}+\Omega_{\max}\right)c_r+\Omega_{\min}\Omega_{\max}\right\}P\ dr, \end{aligned}$$

where (66) is used. So

$$\left(c_r^2+c_i^2\right)-\left(\Omega_{\min}+\Omega_{\max}\right)c_r+\Omega_{\min}\Omega_{\max} \leq 0$$

and the conclusion follows. □

We also have an a priori bound for unstable solutions. The proof of the following lemma is essentially the same as that of Lemma 3.7 in the shear flow case. So we state only the result.

LEMMA 5.3. *Denote* $\omega(r) = rD^2\Omega+3D\Omega$. *For any solution* (α, c, ϕ) *to* (57) *with* α *real positive,* $c = c_r + ic_i\,(c_i > 0)$, *and* $\|\phi\|_2 = 1$, *we have the identities*

$$\int_{R_1}^{R_2} r\left|D\phi\right|^2 dr + \alpha^2\int_{R_1}^{R_2}\frac{1}{r}\left|\phi\right|^2 dr + \int_{R_1}^{R_2}\frac{\omega(r)\left|\phi\right|^2(\Omega-q)}{\left|\Omega-c\right|^2}dr = 0\ \forall q\in\mathbf{R}, \tag{67}$$

$$\int_{R_1}^{R_2}\left(\left|D_*D\phi\right|^2 r^3+2\alpha^2 r\left|D\phi\right|^2 dr+\alpha^4\frac{1}{r}\left|\phi\right|^2\right)dr=\int_{R_1}^{R_2} r\frac{\omega\left(r\right)^2}{\left|\Omega-c\right|^2}\left|\phi\right|^2 dr. \tag{68}$$

For a flow of class $\mathcal{K}^+$, we have the inequalities

$$\int_{R_1}^{R_2}\left(r\left(D\phi\right)^2+\alpha^2\frac{1}{r}\left|\phi\right|^2\right)dr<\int_{R_1}^{R_2} K\left(r\right)\phi^2 dr \tag{69}$$

and

$$\int_{R_1}^{R_2}\left(\left|D_*D\phi\right|^2 r^3+2\alpha^2 r\left|D\phi\right|^2 dr+\alpha^4\frac{1}{r}\left|\phi\right|^2\right)dr<R_2\left\|\omega\right\|_\infty\int_{R_1}^{R_2} K\left(r\right)\phi^2 dr. \tag{70}$$

In particular, we have the a priori estimate $\left\|\phi\right\|_{\mathbf{H}^2}\leq C\left(\Omega\right)$, where $C\left(\Omega\right)$ is some constant depending only on Ω.

Indeed, (67), (68), (69), (70) are the analogues of (29), (31), (27), and (28), respectively. Their proofs are similar to that of the shear flow case.

Remark 5.4. From (69) we see that a necessary condition for instability in the rotating case is

$$\inf_{\phi\in\mathbf{H}_0^1(R_1,R_2)}\frac{\int_{R_1}^{R_2} r\left(D\phi\right)^2 dr-\int_{R_1}^{R_2} K\left(r\right)\phi^2 dr}{\int_{R_1}^{R_2}\frac{1}{r}\phi^2 dr}<-1,$$

since α must be a positive integer.

We now define the neutral limiting modes for the rotating case.

DEFINITION 5.5. *The triple (c_s,α_s,ϕ_s) with c_s real and $\alpha_s>1$ is said to be a neutral limiting mode if it is the limit of growing solutions (c_k,α_k,ϕ_k) (with $\operatorname{Im}c_k>0$) of the extended Rayleigh equation* (57). *The precise notion of convergence of ϕ_k to ϕ_s is made clear in Lemma* 5.6. *Formally (c_s,α_s,ϕ_s) ought to satisfy*

$$\left(\Omega-c_s\right)\left(D_*D-\alpha_s^2/r^2\right)\phi_s-r^{-1}\omega\left(r\right)\phi_s=0. \tag{71}$$

We call c_s the neutral limiting phase speed and α_s the neutral limiting wave number.

The following is the analogue of Lemma 3.6.

LEMMA 5.6. *Let $\{(c_k,\alpha_k,\phi_k)\,(\text{with}\operatorname{Im}c_k>0)\}_{k=1}^{\infty}$ be the solutions to the extended Rayleigh equation* (57) *with $\left\|\phi_k\right\|=1$, and let (c_k,α_k) converge to (c_s,α_s) with $\alpha_s>1$. Then ϕ_k converges uniformly to a function ϕ_s on any compact subset of S_0, ϕ_s'' exists on S_0, and ϕ_s satisfies* (71).

We state the following results about neutral limiting modes without proof. They are the analogues of Theorems 1.7 and 3.9, respectively.

LEMMA 5.7. *If the rotating flow is in class $\mathcal{K}^+$, then for any neutral limiting mode (c_s,α_s,ϕ_s) with $\alpha_s>1$, we must have $c_s=\Omega_s$, and $\phi_s\in\mathbf{H}^2\cap\mathbf{H}_0^1\left(R_1,R_2\right)$ must satisfy*

$$\left(D_*D-\alpha_s^2/r^2\right)\phi_s+r^{-1}K\left(r\right)\phi_s=0. \tag{72}$$

LEMMA 5.8. *Let $\Omega\left(y\right)$ be as in Theorem* 1.3. *Let Ξ be the set of all unstable wave numbers greater than* 1. *Then Ξ is open and any real boundary point α_s of Ξ is either* 1 *or some wave number satisfying* (72) *for some nontrivial ϕ_s in $\mathbf{H}^2\cap\mathbf{H}_0^1\left(R_1,R_2\right)$.*

We also have a perturbation result near neutral modes, the analogue of Theorem 4.1.

THEOREM 5.9. *Suppose* $\Omega(y)$ *is in class* $\mathcal{K}^+$ *and* $(\phi_s, \alpha_s, \Omega_s)$ $(\alpha_s > 1)$ *satisfies*

$$(\Omega - \Omega_s)\left(D_*D - \alpha_s^2/r^2\right)\phi_s - r^{-1}\omega(r)\phi_s = 0, \tag{73}$$

with $\phi_s(R_1) = \phi_s(R_2) = 0$. *Then there exists* $\varepsilon_0 < 0$ *such that if* $\varepsilon_0 < \varepsilon < 0$, *there is a nontrivial solution* ϕ_ε *to the extended Rayleigh equation*

$$(\Omega - \Omega_s - c(\varepsilon))\left(D_*D - \alpha(\varepsilon)^2/r^2\right)\phi_\varepsilon - r^{-1}\omega(r)\phi_\varepsilon = 0,$$

with $\phi_\varepsilon(R_1) = \phi_\varepsilon(R_2) = 0$. *Here* $\alpha(\varepsilon) = \sqrt{\varepsilon + \alpha_s^2}$ *is the perturbed wave number and* $\Omega_s + c(\varepsilon)$ *is an unstable eigenvalue with* $\operatorname{Im} c(\varepsilon) > 0$. *The function* $c(\varepsilon)$ *is differentiable in* $(-\varepsilon_0, 0)$ *and*

$$\lim_{\varepsilon \to 0-} c(\varepsilon) = 0,$$

$$\lim_{\varepsilon \to 0-} c'(\varepsilon) = \frac{\int_{R_1}^{R_2} \frac{1}{r}\phi_s^2(r)\,dr}{i\pi \sum_{k=1}^{l}\left(|\Omega'|^{-1} K\phi_s^2\right)|_{y=r_k} + P\int_{R_1}^{R_2}\left(K(r)\phi_s^2(r)\right)/\left(\Omega(r) - \Omega_s\right)dr}, \tag{74}$$

where $r_1, \ldots, r_l$ *are the points such that* $\Omega(r) = \Omega_s$ *and* $\mathcal{P}\int_{y_1}^{y_2}$ *denotes the Cauchy principal part.*

Proof of Theorem 1.3. If (11) is satisfied, we know that for any $\alpha \in (1, \alpha_{\max})$, there is an unstable solution to the extended Rayleigh equation (57). The proof is essentially the same as that of Theorem 1.2, by using Theorem 5.9 and Lemma 5.8, so we skip it here. If condition (12) is satisfied, then $\alpha_{\max} > 2$, and we get instability at $n = 2$ for the rotating Rayleigh equation (8).

Now we turn to the case when $1 < \alpha_{\max} \le 2$ and $\Omega(R_1) \ne \Omega(R_2)$. We want to show that there exists an unstable mode for $n = 1$. This is the bottom case for rotating flows. Now for each $\alpha \in (1, \alpha_{\max})$, we already have an unstable mode. We shall show that the growth rate $\operatorname{Im} c(\alpha)$ has some positive lower bound when α tends to 1. Assuming this, we can find some unstable mode for $\alpha = 1$ by using the same argument as in the proof of Theorem 1.2.

We now prove that $\operatorname{Im} c(\alpha)$ has some positive lower bound. The argument we use here is similar to that in the proof of Lemma 4.3. Supposing otherwise, we can find a sequence $\{(c_k, \alpha_k, \phi_k)\}_{k=1}^{\infty}$ with $\alpha_k \to 1+$, $\operatorname{Im} c_k \to 0$, $\operatorname{Re} c_k \to c$. The convergence of $\{c_k\}$ is guaranteed by (60), from which we also know that c is in the range of Ω. Because of Lemma 5.3, $\|\phi_k\|_{\mathbf{H}^2}$ is uniformly bounded. So there exists some nonzero ϕ_0 in $\mathbf{H}^2 \cap \mathbf{H}_0^1$ such that a subsequence $\{\phi_{n_k}\}$ converges to it in the $\mathbf{C}^1$ sense. By passing to the limit in the equation

$$(\Omega - c_{n_k})\left(D_*D - \alpha_{n_k}^2/r^2\right)\phi_{n_k} - r^{-1}(rD^2\Omega + 3D\Omega)\phi_{n_k} = 0,$$

we deduce that the function ϕ_0 satisfies

$$(\Omega - c)\left(D_*D - 1/r^2\right)\phi_0 - r^{-1}(rD^2\Omega + 3D\Omega)\phi_0 = 0 \tag{75}$$

sectionally in each (r_i, r_{i+1}). Here $r_1 < r_2 < \cdots < r_{k_c}$ are the zeros of $\Omega(y) - c$ and $r_0 = R_1$, $r_{k_c+1} = R_2$. By (59), ϕ_0 satisfies

$$\left((\Omega - c)\left(r^2\phi_0\right)' - r^2\phi_0\Omega'\right)' - 3\left((r\phi_0)(\Omega - c)\right)' = 0,$$

which is equivalent to

$$\left(r\left(((\Omega-c)r)'\phi_0-\phi_0'(\Omega-c)r\right)\right)'=0. \tag{76}$$

From (76), we deduce that ϕ_0 must be nonzero at each point $r_i\,(i=1,\ldots,k_c)$. Indeed, supposing otherwise, by the same argument as in the proof of Lemma 4.3, we can show that there is some constant b such that $(\Omega(r)-c)r=b\phi_0(r)$ in $[R_1,R_2]$. This implies that $\Omega(R_1)=\Omega(R_2)$, a contradiction. Now by the same argument as in the last part of the proof of Theorem 1.2, we can show that there is no perturbation of the neutral mode at $\alpha=1$ to its right neighborhood. This contradiction shows that $\operatorname{Im}c(\alpha)$ is bounded below and $\operatorname{Im}c>0$. Thus ϕ_0 satisfying (75) is an unstable solution to the rotating Rayleigh equation (8).

Combining this result with Remark 5.4, we deduce that the condition (11) is sharp for instability when $\Omega(R_1)\neq\Omega(R_2)$. This finishes the proof of Theorem 1.3. □

6. Unbounded flows. We now consider the unbounded shear flows. We prove Theorem 1.5(i) only for the flow $U(y)$ defined on $(-\infty,+\infty)$. The proof of Theorem 1.5(i) for the shear flows defined on the half line is similar. The flow with $U(-\infty)=U(+\infty)$ is called a jet and the one with $U(-\infty)\neq U(+\infty)$ is called a shear layer, as in [7].

Proof of Theorem 1.5(i). We divide the proof into several steps.

Step 1. First we observe that for any real c, $U(y)=c$ holds for only a finite number of points. Otherwise, there exists some real c_0 and an infinite sequence $\{y_n\}$ such that $U(y_n)=c_0$ for each n. Then $\{y_n\}$ must be bounded by our condition that $U(\pm\infty)$ are obtained at only a finite number of points. So there exists some y_0 such that a subsequence $\{y_{n_k}\}$ converges to it. Since $U(y)$ is a $\mathbf{C}^2$ function, we deduce that $U(y_0)=c_0$ and $U'(y_0)=U''(y_0)=0$. So y_0 is an inflection point and c_0 equals the inflection value U_s. But then $K(y)$ defined by (5) is unbounded at y_0 since $U'(y_0)=0$. This is a contradiction.

Since $K(y)\to 0$ as $y\to\infty$, it is easy to see that $-\frac{d^2}{dy^2}-K(y)$ is a relatively compact perturbation of $-\frac{d^2}{dy^2}$ defined on $\mathbf{H}^2(\mathbf{R})$. So by Weyl's theorem [21]

$$\sigma_{\mathrm{ess}}\left(-\frac{d^2}{dy^2}-K(y)\right)=\sigma_{\mathrm{ess}}\left(-\frac{d^2}{dy^2}\right)=(0,+\infty).$$

Thus $-\frac{d^2}{dy^2}-K(y)$ has only a discrete set of negative eigenvalues which can accumulate only at 0. Let $-\alpha_0^2<-\alpha_1^2<\cdots<-\alpha_k^2<\cdots<0$ denote all the negative eigenvalues of the operator $-\frac{d^2}{dy^2}-K(y)$ on $\mathbf{H}^2(\mathbf{R})$. We fix some α in $(0,\alpha_0)$ and assume $\alpha\neq\alpha_i$ for each $i\geq 1$. Let $I_n=(-n,n)$ and let L_n denote the operator $-\frac{d^2}{dy^2}-K(y)$ on $\mathbf{H}^2(I_n)\cap\mathbf{H}_0^1(I_n)$. Take n large enough so that all the inflection points of $K(y)$ are included in I_n. Denote by $-\alpha_{0,n}^2$ the lowest eigenvalue of L_n. Then by the result of [3], $-\alpha_{0,n}^2$ converges to $-\alpha_0^2$ as n tends to infinity. So by taking n large enough, α is in $(0,\alpha_{0,n})$. Now since $U(y)|_{I_n}$ is clearly in class $\mathcal{K}^+$, by applying Theorem 1.2 to this truncated flow, we obtain a solution ϕ_n in $\mathbf{H}^2(I_n)\cap\mathbf{H}_0^1(I_n)$ satisfying the Rayleigh equation with c_n $(\operatorname{Im}c_n>0)$, that is,

$$(U-c_n)\left(\frac{d^2}{dy^2}-\alpha^2\right)\phi_n-U''\phi_n=0 \tag{77}$$

in I_n.

Step 2. Now we have a sequence $\{(\phi_n, c_n)\}$ satisfying (77) with $\|\phi_n\|_{\mathbf{L}^2(I_n)} = 1$. By Lemma 3.7, we have $\|\phi_n\|_{\mathbf{H}^2(I_n)} \leq C$ (where C is some constant depending only on $\|K\|_\infty$). We extend ϕ_n to a function in $\mathbf{H}^2(\mathbf{R})$ by setting it to be zero on I_n^c. For convenience, we still use ϕ_n to denote the extended function. Then $\|\phi_n\|_{\mathbf{H}^2(\mathbf{R})} \leq C$. So ϕ_n converges weakly in $\mathbf{H}^2(\mathbf{R})$ to a function $\phi_0 \in \mathbf{H}^2(\mathbf{R})$. We shall show that ϕ_0 is not identically zero.

Let n_0 be sufficiently large such that

$$K(y) < \frac{1}{2}\alpha^2 \quad \text{if} \quad y \in I_{n_0}^c.$$

From (27) we have

$$\int_{I_n} \left(|\phi_n'|^2 + \alpha^2 |\phi_n|^2 \right) dy < \int_{I_n} K(y) |\phi_n|^2 \, dy. \tag{78}$$

For any $n > n_0$, from (78) and $K > 0$ we have

$$\begin{aligned} \int_{I_{n_0}} K(y) |\phi_n|^2 \, dy &> \alpha^2 - \int_{I_{n_0}^c \cap I_n} K(y) |\phi_n|^2 \, dy \\ &> \alpha^2 - \frac{1}{2}\alpha^2 \int_{I_{n_0}^c \cap I_n} |\phi_n|^2 \, dy \\ &> \frac{1}{2}\alpha^2. \end{aligned} \tag{79}$$

Since ϕ_n converges strongly in $\mathbf{H}^1(I_{n_0})$ to ϕ_0, from (79) we get

$$\int_{I_{n_0}} K(y) |\phi_0|^2 \, dy \geq \frac{1}{2}\alpha^2.$$

This shows that ϕ_0 is nontrivial.

Step 3. By Howard's semicircle theorem (see (4)), $\{c_n\}$ is bounded. Supposing c_n to converge to c_0, we shall show that $\operatorname{Im} c_0 > 0$. Otherwise, c_0 is some real number in the range of $U(y)$. From the a priori $\mathbf{H}^2$ bound provided by Lemma 3.7, we deduce that ϕ_n converges to ϕ_0 locally in $\mathbf{C}^1$. Suppose $z_1 < \cdots < z_{k_0}$ are all the points such that $U(z_i) = c_0$. Then by taking limit $k \to \infty$ in (77), we deduce that ϕ_0 satisfies

$$(U - c_0) \left(\frac{d^2}{dy^2} - \alpha^2 \right) \phi_0 - U'' \phi_0 = 0 \tag{80}$$

within each interval $(-\infty, z_1)$, $(z_1, z_2), \ldots, (z_{k_0}, \infty)$. Since $\phi_0 \neq 0$, by the proof of Lemma 3.5, we deduce that $\phi_0(z_i) \neq 0$ for some z_i. Then by the same argument as in the proof of Theorem 3.3, we know that c_0 must be an inflection value of $U(y)$, which can only be U_s. Now

$$\frac{U''}{U - c_0} = -K(y)$$

is a bounded continuous function and ϕ_0 is in $\mathbf{C}^1$. So from (80), ϕ_0 satisfies

$$\left(\frac{d^2}{dy^2} - \alpha^2 \right) \phi_0 + K(y) \phi_0 = 0$$

on $\mathbf{R}$ and $\phi_0 \in \mathbf{H}^2(\mathbf{R})$. Thus $-\alpha^2$ is a negative eigenvalue of

$$-\frac{d^2}{dy^2} - K(y)$$

on $\mathbf{H}^2(\mathbf{R})$, which is a contradiction to our assumption that $\alpha \neq \alpha_i$ for all i. So we must have $\operatorname{Im} c_0 > 0$.

Now $\frac{U''}{U-c_0}$ is a bounded continuous function. By passing to limits in (77), we deduce that $\phi_0 \neq 0$ satisfies the Rayleigh equation (3) on the whole line, with $\operatorname{Im} c_0 > 0$. We thus get instability for each wave number α in $(0, \alpha_0)$ such that $\alpha \neq \alpha_i (i \geq 1)$.

Step 4. Now we investigate the possibility of an instability at $\alpha = \alpha_i (i \geq 1)$. The argument is the same as in the proof of Theorem 1.2 for the case when a wave number equals some neutral limiting wave number. We only sketch it here. For each $\alpha > \alpha_i$, we get instability by the previous steps. The main point is still to show that the growth rate $\operatorname{Im} c(\alpha)$ has some positive lower bound when $\alpha \to \alpha_i +$. Supposing otherwise, we get a sequence $\{(c_k, \alpha_k, \phi_k)\}_{k=1}^{\infty}$ of solutions to Rayleigh's equation (3), with $\alpha_k \to \alpha_i +$ and $\operatorname{Re} c_k \to c_s$, $\operatorname{Im} c_k \to 0 +$. It is not difficult to see that we can extend Theorem 1.7 to the current case, so c_s must equal U_s. By the same argument as in Step 2, we deduce that ϕ_k converges to some $\phi_s \neq 0$ weakly in $\mathbf{H}^2(\mathbf{R})$ and locally in $\mathbf{C}^1$. Now ϕ_s satisfies

$$-\frac{d^2}{dy^2}\phi_s + \frac{U''}{U-U_s}\phi_s = -\alpha_i^2 \phi_s. \tag{81}$$

By Remark 7.2 in the appendix, the analogues of the two limits (52) and (53) still hold true. That is, we have

$$\int_{\mathbf{R}} \phi_k \phi_s \to \int_{\mathbf{R}} |\phi_s|^2 \, dy \tag{82}$$

and

$$\int_{\mathbf{R}} \frac{K(y)\phi_s \phi_k}{U-c_k} dy \to \mathcal{P}\int_{\mathbf{R}} \frac{K(y)\phi_s^2}{(U-U_s)} dy + i\pi \sum_{i=1}^{l} \left(|U'|^{-1} K \phi_s^2 \right)|_{y=a_i}. \tag{83}$$

Here $a_1, \ldots, a_l$ are all the inflection points and $c_k \to U_s$ with $\operatorname{Im} c_k > 0$. Then the rest of the proof just follows what we did in the last part of the proof of Theorem 1.2. So we skip it here. □

Proof of Theorem 1.5(ii). Assume ϕ is odd on the whole line. Fix α in the interval $(\alpha_{2k_0-1}, \alpha_{2k_0-2})$ for some k_0. Using the same notation as in Step 1 of the proof of (i), we get the truncated operator L_n defined on $I_n = (-n, n)$. Let $-\alpha_{0,n}^2 < -\alpha_{1,n}^2 < \cdots < -\alpha_{k,n}^2 < \cdots < 0$ be all the negative eigenvalues of L_n. By the result in [3], L_n has at least $2k_0$ negative eigenvalues when n is large enough and $(-\alpha_{0,n}^2, \ldots, -\alpha_{2k_0-1,n}^2)$ converges to $\left(-\alpha_0^2, \ldots, -\alpha_{2k_0-1}^2\right)$ as n tends to infinity. So for n large enough, α is in the interval $(\alpha_{2k_0-1,n}, \alpha_{2k_0-2,n})$. Since $U(y)|_{I_n}$ is odd, by Theorem 1.4 we obtain an unstable solution ϕ_n satisfying (77) in I_n, with $c_n = i\lambda_n$ $(\lambda_n > 0)$. We get the same a priori bound as in Lemma 3.7 directly as follows. By Lemma 3.5 with $q = 0$, we have

$$\int_{I_n} \left(|\phi_n'|^2 + \alpha^2 |\phi_n|^2 + \frac{U''U}{|U - i\lambda_n|^2} |\phi_n|^2 \right) dy = 0.$$

Thus

$$\int_{I_n}\left(\left|\phi_n'\right|^2+\alpha^2\left|\phi_n\right|^2\right)dy=\int_{I_n}K(y)\frac{U^2}{U^2+\lambda_n^2}\left|\phi_n\right|^2dy$$
$$\leq\int_{I_n}\left|K(y)\right|\left|\phi_n\right|^2dy. \tag{84}$$

Similarly, we get from (31) that

$$\int_{I_n}\left(\left|\phi_n''\right|^2+2\alpha^2\left|\phi_n'\right|^2+\alpha^4\left|\phi\right|^2\right)dy<\int_{I_n}K(y)^2\left|\phi_n\right|^2dy. \tag{85}$$

From (84) and (85), we have $\|\phi_n\|_{\mathbf{H}^2(\mathbf{R})}\leq C\left(\|K\|_\infty\right)$. So ϕ_n converges to some ϕ_0 weakly in $\mathbf{H}^2(\mathbf{R})$. Let λ_n converge to some nonnegative λ_0. By the same argument as in Steps 2 and 3 of the proof of (i), we show that $\phi_0\neq 0$ and $\lambda_0>0$. Then we get an unstable solution ϕ_0 to the Rayleigh equation on the whole line, with the unstable eigenvalue $c=i\lambda_0$. The proof of Theorem 1.5(ii) is thus finished. □

Using the same argument as in the proof of Lemma 4.3, we can show that if $U(-\infty)\neq U(+\infty)$ (the shear layer case), then at $\alpha=0$ there is also an unstable solution to the Rayleigh equation. This coincides with the conclusion in [7], which was deduced from the asymptotic expansion in the long wave limit.

7. Appendix. In this appendix, we prove some asymptotic formulas used in the proof of Theorems 4.1 and 1.2.

LEMMA 7.1. *Assume a sequence of differentiable functions* $\{\psi_k\}_{k=1}^\infty$ *converges in* $C^1[y_1,y_2]$ *to* $\psi_0(y)$. *Let* $c_k=p_k+ib_k\,(b_k>0)$ *converge to* 0. *Denote* $W_k(y)=U(y)-U_s-p_k$. *Then we have following limits:*

$$\lim_{k\to\infty}\int_{y_1}^{y_2}\frac{\psi_k(y)W_k(y)}{W_k(y)^2+b_k^2}dy=\mathcal{P}\int_{y_1}^{y_2}\frac{\psi_0(y)}{U(y)-U_s}dy, \tag{86}$$

$$\lim_{k\to\infty}\int_{y_1}^{y_2}\frac{\psi_k(y)W_k(y)^3}{\left(W_k(y)^2+b_k^2\right)^2}dy=\mathcal{P}\int_{y_1}^{y_2}\frac{\psi_0(y)}{U(y)-U_s}dy, \tag{87}$$

$$\lim_{k\to\infty}\int_{y_1}^{y_2}\frac{\psi_k(y)b_k}{W_k(y)^2+b_k^2}dy=\pi\sum_{i=1}^{l}\left(\left|U'\right|^{-1}\psi_0\right)|_{y=a_i}, \tag{88}$$

$$\lim_{k\to\infty}\int_{y_1}^{y_2}\frac{\psi_k(y)b_k^3}{\left(W_k(y)^2+b_k^2\right)^2}dy=\frac{1}{2}\pi\sum_{i=1}^{l}\left(\left|U'\right|^{-1}\psi_0\right)|_{y=a_i}, \tag{89}$$

$$\lim_{k\to\infty}\int_{y_1}^{y_2}\frac{\psi_k(y)b_k^2W_k}{\left(W_k(y)^2+b_k^2\right)^2}dy=\lim_{k\to\infty}\int_{y_1}^{y_2}\frac{\psi_k(y)b_k^3}{\left(W_k(y)^2+b_k^2\right)^2}dy=0. \tag{90}$$

Here $a_1,\ldots,a_l$ *are all the points such that* $U(y)=U_s$, *and we assume* $U'(y)\neq 0$ *at each* a_k.

Proof. Let $\|\psi_k\|_{\mathbf{C}^1} \leq M$ (independent of k). For $\delta_0 > 0$, we can find $r_0 > 0$ such that $|U'(y)| \geq \delta_0$ for all y in the set $E_{r_0} = \cup_{i=1}^{l} I(a_i; r_0)$. Here $I(a_i; r_0) = (a_i - r_0, a_i + r_0)$. Taking k large enough, then we know there are exactly l points such that $W_k(y) = 0$, one in each $I(a_i; r_0)$, which we denote by $a_i^{(k)}$.

Proof of (86). By the definition of Cauchy principal part, for any $\varepsilon > 0$, there exists some $r_1 > 0$ such that if $0 < r < r_1$, then

$$\left| \int_{E_r^c} \frac{\psi_0(y)}{U(y) - U_s} dy - \mathcal{P} \int_{y_1}^{y_2} \frac{\psi_0(y)}{U(y) - U_s} dy \right| < \frac{\varepsilon}{3}. \tag{91}$$

Now

$$\frac{\psi_k(y) W_k(y)}{W_k(y)^2 + b_k^2} \to \frac{\psi_0(y)}{U(y) - U_s}$$

in E_r^c uniformly as $k \to \infty$. So if k is large enough, then

$$\left| \int_{E_r^c} \frac{\psi_k(y) b_k}{W_k(y)^2 + b_k^2} dy - \int_{E_r^c} \frac{\psi_0(y)}{U(y) - U_s} dy \right| < \frac{\varepsilon}{3}. \tag{92}$$

Let $r < \min\{r_0, r_1\}$. We estimate the integral on each $I(a_i; r)$. Suppose U is increasing on $I(a_1; r)$. Let

$$t = W_k(y), \quad t_1^k = W_k(a_1 + r), \quad t_0^k = W_k(a_1 - r),$$

$$\tilde{\psi}_k(t) = \psi_k(W_k(t)) \frac{1}{U'(W_k(t))}.$$

Then

$$\int_{I(a_1;r)} \frac{\psi_k(y) W_k(y)}{W_k(y)^2 + b_k^2} dy$$

$$= \int_{t_0^k}^{t_1^k} \frac{\tilde{\psi}_k(t)\, t}{t^2 + b_k^2} dt = \tilde{\psi}_k(0) \int_{t_0^k}^{t_1^k} \frac{t}{t^2 + b_k^2} dt + \left(\int_{t_0^k}^{t_1^k} \frac{\left(\tilde{\psi}_k(t) - \tilde{\psi}_k(0)\right) t}{t^2 + b_k^2} dt \right) = I + II.$$

We have

$$\left| t_1^k - t_0^k \right| \leq \|U'\|_\infty 2r,$$

and $|\tilde{\psi}_k'(t)| \leq M'$ (independent of k) in $I(a_1; r)$. So

$$\|II\| \leq M' \left| t_1^k - t_0^k \right| \leq M' \|U'\|_\infty 2r.$$

And

$$I = \tilde{\psi}_k(0) \frac{1}{2} \ln \frac{t_1^k + b_k^2}{t_0^k + b_k^2}$$

tends to zero as k tends to infinity for any fixed $r > 0$. Thus by choosing r small enough and then letting k large, we have

$$\left| \int_{E_r} \frac{\psi_k(y) W_k(y)}{W_k(y)^2 + b_k^2} dy \right| < \frac{\varepsilon}{3}.$$

Combining with (91) and (92), we have for k large enough

$$\left| \int_{y_1}^{y_2} \frac{\psi_k(y) W_k(y)}{W_k(y)^2 + b_k^2} dy - \mathcal{P} \int_{y_1}^{y_2} \frac{\psi_0(y)}{U(y) - U_s} dy \right| < \varepsilon.$$

This ends the proof of (86).

We prove (87) and (90) in the same way.

Proof of (88). For any fixed r,

$$\lim_{k\to\infty} \int_{E_r^c} \frac{\psi_k(y) b_k}{W_k(y)^2 + b_k^2} dy = 0.$$

So we only need to consider the integral on each small interval $I(a_i; r)$. Using the same notation as above, we have

$$\begin{aligned} &\int_{I(a_1;r)} \frac{\psi_k(y) b_k}{W_k(y)^2 + b_k^2} dy \\ &= \int_{t_0^k}^{t_1^k} \frac{\tilde{\psi}_k(t) b_k}{t^2 + b_k^2} dt = \tilde{\psi}_k(0) \int_{t_0^k}^{t_1^k} \frac{b_k}{t^2 + b_k^2} dt + \left(\int_{t_0^k}^{t_1^k} \frac{\left(\tilde{\psi}_k(t) - \tilde{\psi}_k(0)\right) b_k}{t^2 + b_k^2} dt \right) = I + II. \end{aligned}$$

Then it is easy to see

$$|II| \le \frac{1}{2} M' \left| t_1^k - t_0^k \right| \le M' \|U'\|_\infty r.$$

Since

$$\lim_{k\to\infty} \tilde{\psi}_k(0) = \lim_{k\to\infty} \psi_k\left(a_1^{(k)}\right) \frac{1}{U'\left(a_1^{(k)}\right)} = \psi_0(a_1) \frac{1}{U'(a_1)},$$

$$\lim_{k\to\infty} t_1^k = U(a_1 + r) - U(a_1) > 0, \quad \lim_{k\to\infty} t_0^k = U(a_1 - r) - U(a_1) < 0$$

so we have

$$\lim_{k\to\infty} I = \lim_{k\to\infty} \tilde{\psi}_k(0) \int_{\frac{t_0^k}{b_k}}^{\frac{t_0^k}{b_k}} \frac{1}{t^2 + 1} dt = \psi_0(a_1) \frac{1}{U'(a_1)} \int_{-\infty}^{+\infty} \frac{1}{t^2 + 1} dt = \pi \psi_0(a_1) \frac{1}{U'(a_1)}.$$

Summing the contributions from each $I(a_i; r)$, we deduce (88). The proof of (89) is the same by noticing that

$$\int_{\mathbf{R}} \frac{1}{(1 + t^2)^2} dt = \frac{1}{2}\pi. \qquad \square$$

Now (53) follows directly from the above lemma. Indeed letting $\psi_k = K(y) \phi_s \phi_k$, $c_k = p_k + i b_k$, we can write B_k in (53) as

$$B_k = \int_{y_1}^{y_2} \frac{\psi_k(y) W_k(y)}{W_k(y)^2 + b_k^2} dy + i \int_{y_1}^{y_2} \frac{\psi_k(y) b_k}{W_k(y)^2 + b_k^2} dy.$$

Since

$$\psi_k \to \psi_0 = K(y) \phi_s^2 \quad \text{in } C^1[y_1, y_2] \text{ as } k \to \infty,$$

by (86) and (88) we have

$$\lim_{k\to\infty} B_k = \mathcal{P}\int_{y_1}^{y_2} \frac{K(y)\,\phi_s^2}{(U-U_s)}dy + i\pi\sum_{k=1}^{l}\left(\left|U'\right|^{-1} K\phi_s^2\right)|_{y=a_k}.$$

Remark 7.2. *The limit* (53) *still holds for the case* $[y_1, y_2] = (-\infty, +\infty)$ *or* $(0, +\infty)$ *under the assumption that* $\{\phi_k\}_{k=1}^{\infty}$ *converges weakly in* $\mathbf{L}^2(y_1, y_2)$ *and locally in* C^1 *to* $\phi_0(y)$. *To see it, we notice that for fixed* $r_0 > 0$,

$$\lim_{k\to\infty}\int_{\mathbf{R}/E_{r_0}} \frac{K(y)\,\phi_s\phi_k W_k(y)}{W_k(y)^2 + b_k^2}dy = \int_{\mathbf{R}/E_{r_0}} \frac{K(y)\,\phi_s^2}{U(y) - U_s}dy$$

by weak convergence of ϕ_k. *We can deal with the integral on each small interval* $I(a_i; r_0)$ *in the same way as in the proof of Lemma* 7.1, *noticing that the* $\mathbf{C}^1$ *norm of* ϕ_k *is locally uniformly bounded.*

To prove (41) we need the following lemma.

LEMMA 7.3. *Assume a sequence of differentiable functions* $\{\Gamma_k\}_{k=1}^{\infty}$ *converges in* $C^1[y_1, y_2]$ *to* $\Gamma_0(y)$. *Let* $c_k = p_k + ib_k$ *converge to* 0, *where* $b_k > 0$ *and* $|p_k| \le Rb_k$. *Then we have*

(93)
$$\lim_{k\to\infty} -\int_{y_1}^{y_2} \frac{U''(y)\,\Gamma_k(y)}{(U - U_s - c_k)^2}dy = \mathcal{P}\int_{y_1}^{y_2} \frac{K(y)\,\Gamma_0(y)}{U(y) - U_s}dy + i\pi\sum_{i=1}^{l}\left(\left|U'\right|^{-1} K\phi_0\right)|_{y=a_i}.$$

Proof. Denote $W_k(y) = U(y) - U_s - p_k$. We have

$$\begin{aligned}
&\int_{y_1}^{y_2} \frac{-U''(y)\,\Gamma_k(y)}{(U - U_s - c_k)^2}dy\\
&= \int_{y_1}^{y_2} \frac{K(y)\,\Gamma_k(y)\,(U(y) - U_s)}{(W_k(y) - ib_k)^2}dy\\
&= \int_{y_1}^{y_2} \frac{K(y)\,\Gamma_k(y)\,(W_k + p_k)\left(W_k^2 + 2ib_kW_k - b_k^2\right)}{(W_k^2 + b_k^2)^2}dy\\
&= \int_{y_1}^{y_2} \frac{K(y)\,\Gamma_k(y)\,W_k^3}{(W_k^2 + b_k^2)^2}dy + 2i\int_{y_1}^{y_2} \frac{K(y)\,\Gamma_k(y)\,b_kW_k^2}{(W_k^2 + b_k^2)^2}dy - \int_{y_1}^{y_2} \frac{K(y)\,\Gamma_k(y)\,W_kb_k^2}{(W_k^2 + b_k^2)^2}dy\\
&\qquad + \int_{y_1}^{y_2} \frac{K(y)\,\Gamma_k(y)\,p_k\left(W_k^2 - b_k^2\right)}{(W_k^2 + b_k^2)^2}dy + 2i\int_{y_1}^{y_2} \frac{K(y)\,\Gamma_k(y)\,p_kb_kW_k}{(W_k^2 + b_k^2)^2}dy\\
&= I + II + III + IV + V\ .
\end{aligned}$$

Now we estimate each term separately. By (87) in Lemma 7.1, we have

$$\lim_{k\to\infty} I = \mathcal{P}\int_{y_1}^{y_2} \frac{K(y)\,\Gamma_0(y)}{U(y) - U_s}dy.$$

By (88) and (89), we have

$$\begin{aligned}\lim_{k\to\infty} II &= 2i\left(\lim_{k\to\infty}\int_{y_1}^{y_2}\frac{K(y)\Gamma_k(y)b_k}{W_k^2+b_k^2}dy-\lim_{k\to\infty}\int_{y_1}^{y_2}\frac{K(y)\Gamma_k(y)b_k^3}{\left(W_k^2+b_k^2\right)^2}dy\right)\\ &= 2i\left(\pi\sum_{i=1}^{l}\left(|U'|^{-1}K\phi_0\right)|_{y=a_i}-\frac{1}{2}\pi\sum_{i=1}^{l}\left(|U'|^{-1}K\phi_0\right)|_{y=a_i}\right)\\ &= i\pi\sum_{i=1}^{l}\left(|U'|^{-1}K\phi_0\right)|_{y=a_i}.\end{aligned}$$

By (90), $\lim_{k\to\infty} III = 0$. Notice that

$$\begin{aligned}IV &= \frac{p_k}{b_k}\int_{y_1}^{y_2}\frac{K(y)\Gamma_k(y)b_k\left(W_k^2+b_k^2-2b_k^2\right)}{\left(W_k^2+b_k^2\right)^2}dy\\ &= \frac{p_k}{b_k}\left(\int_{y_1}^{y_2}\frac{K(y)\Gamma_k(y)b_k}{W_k^2+b_k^2}dy-2\int_{y_1}^{y_2}\frac{K(y)\Gamma_k(y)b_k^3}{\left(W_k^2+b_k^2\right)^2}dy\right)\\ &= \frac{p_k}{b_k}VI.\end{aligned}$$

By (88) and (89), we have

$$\begin{aligned}\lim_{k\to\infty} VI &= \lim_{k\to\infty}\int_{y_1}^{y_2}\frac{K(y)\Gamma_k(y)b_k}{W_k^2+b_k^2}dy-2\lim_{k\to\infty}\int_{y_1}^{y_2}\frac{K(y)\Gamma_k(y)b_k^3}{\left(W_k^2+b_k^2\right)^2}dy\\ &= \pi\sum_{i=1}^{l}\left(|U'|^{-1}K\ \Gamma_0\right)|_{y=a_i}-2\frac{1}{2}\pi\sum_{i=1}^{l}\left(|U'|^{-1}K\ \Gamma_0\right)|_{y=a_i}\\ &= 0.\end{aligned}$$

Combining it with the fact that $|p_k| \le Rb_k$, we have $\lim_{k\to\infty} IV = 0$. Now for the last term, we have

$$V = 2i\frac{p_k}{b_k}\int\frac{K(y)\Gamma_k(y)b_k^2W_k}{\left(W_k^2+b_k^2\right)^2}dy = 2i\frac{p_k}{b_k}VII.$$

By (90), $\lim_{k\to\infty} VII = 0$. Thus we also have $\lim_{k\to\infty} V = 0$ since $|p_k| \le Rb_k$.

Combining the five terms above, we get (93). □

Proof of (41). We must show that (41) holds uniformly in $E_{(R,b_1,b_2)}$. Supposing otherwise, we can find some $\delta_0 > 0$ and a sequence (c_k, ε_k) in E_{R,b_1,b_2} with $\max\left\{b_1^k, b_2^k\right\}$ tending to 0 such that

$$\left|\frac{\partial I}{\partial c}(c_k,\varepsilon_k)-(C+iD)\right| > \delta_0,$$

where

$$C+iD = \frac{1}{\phi_s'(y_2)}\left(i\pi\sum_{k=1}^{l}\left(|U'|^{-1}K\phi_s^2\right)|_{a_k}+\mathcal{P}\int_{y_1}^{y_2}\left(K(y)\phi_s^2(y)\right)/\left(U(y)-U_s\right)dy\right).$$

But

$$\frac{\partial I}{\partial c}(c_k, \varepsilon_k) = \int_{y_1}^{y_2} \frac{-U''(y)\,\Gamma_k(y)}{(U - U_s - c_k)^2} dy,$$

where

$$\Gamma_k(y) = -N(y, y_2; \varepsilon_k, c_k)\,\phi_1(y; c_k, \varepsilon_k).$$

Since Γ_k converges in $\mathbf{C}^1$ to

$$-N(y, y_2; 0, 0)\,\phi_1(y; 0, 0) = \frac{1}{\phi_s'(y_2)}\phi_s^2(y),$$

Lemma 7.3 implies

$$\begin{aligned}
&\lim_{k\to\infty} \frac{\partial I}{\partial c}(c_k, \varepsilon_k) \\
&= \frac{1}{\phi_s'(y_2)} \left(i\pi \sum_{k=1}^{l} \left(|U'|^{-1} K \phi_s^2 \right)|_{a_k} + \mathcal{P} \int_{y_1}^{y_2} \left(K(y)\,\phi_s^2(y) \right) / \left(U(y) - U_s \right) dy \right) \\
&= C + iD,
\end{aligned}$$

which is a contradiction. Thus the uniform convergence of (41) is proved.

Acknowledgments. The author is truly indebted to his advisor, Walter Strauss, for continuous encouragement, guidance, and many helpful discussions. He especially thanks Louis Howard, whose very helpful comments and discussions have improved this paper. The author also thanks Yan Guo and Claude Bardos for useful comments and discussions. Finally, he would like to thank the referee for his comments.

REFERENCES

[1] C. Bardos, Y. Guo, and W. Strauss, *Stable and unstable ideal plane flows*, Chinese Ann. Math., 23B (2002), pp. 149–164.

[2] M. B. Banerjee, R. G. Shandil, K. S. Shirkot, and D. Sharma, *Importance of Tollmien's counter example*, Stud. Appl. Math., 105 (2000), pp. 191–202.

[3] P. B. Bailey, W. N. Everitt, J. Weidmann, and A. Zettl, *Regular approximations of singular Sturm–Liouville problems*, Results Math., 23 (1993), pp. 3–22.

[4] L. Belenkaya, S. Friedlander, and V. Yudovich, *The unstable spectrum of oscillating shear flows*, SIAM J. Appl. Math., 59 (1999), pp. 1701–1715.

[5] S. N. Chow and J. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.

[6] P. G. Drazin and W. H. Reid, *Hydrodynamic Stability*, Cambridge Monogr. Mech. Appl. Math., Cambridge University Press, Cambridge, UK, 1981.

[7] P. G. Drazin and L. N. Howard, *The instability to long waves of unbounded parallel inviscid flow,* J. Fluid Mech., 14 (1962), pp. 257–283.

[8] P. G. Drazin and L. N. Howard, *Hydrodynamical stability of parallel flow of inviscid fluid,* in Advances in Applied Mechanics, Vol. 7, G. Kuerti, ed., Academic Press, New York, 1966, pp. 1–89.

[9] L. D. Faddeev, *On the theory of the instability of a stationary plane-parallel flows of an ideal fluid*, Zap. Naučn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI), 21 (1971), pp. 164–172.

[10] R. Fjørtoft, *Application of integral theorems in deriving criteria of stability of laminar flow and for baroclinic circular vortex*, Geofys. Publ. Norske Vid.-Akad. Oslo, 17 (1950), pp. 1–52.

[11] S. Friedlander, W. Strauss, and M. M. Vishik, *Nonlinear instability in an ideal fluid,* Ann. Inst. H. Poincaré Anal. Non Linéaire, 14 (1997), pp. 187–204.

[12] S. Friedlander and L. N. Howard, *Instability in parallel flow revisited*, Stud. Appl. Math., 101 (1998), pp. 1–21.
[13] E. Grenier, *On the nonlinear instability of Euler and Prandtl equations*, Comm. Pure. Appl. Math., 53 (2000), pp. 1067–1091.
[14] L. N. Howard, *Note on a paper of John W. Miles,* J. Fluid Mech., 10 (1961), pp. 509–512.
[15] L. N. Howard, *The number of unstable modes in hydrodynamic stability problems,* J. Mécanique, 3 (1964), p. 433.
[16] Lord Rayleigh, *On the stability or instability of certain fluid motions*, Proc. London Math. Soc., 9 (1880), pp. 57–70.
[17] C. C. Lin, *The Theory of Hydrodynamic Stability*, Cambridge University Press, Cambridge, UK, 1955.
[18] Z. Lin, *Instability of periodic BGK waves,* Math. Res. Lett., 8 (2001), pp. 521–534.
[19] Z. Lin, *Some stability and instability criteria for ideal plane flows*, Comm. Math. Phys., submitted.
[20] C. Marchioro and M. Pulvirenti, *Mathematical Theory of Incompressible Nonviscous Fluids*, Springer-Verlag, New York, 1994.
[21] M. Reed and B. Simon, *Methods of Modern Mathematical Physics,* Vol. 4, Academic Press, New York, 1978.
[22] D. H. Sattinger, *On the Rayleigh problem in hydrodynamic stability,* SIAM J. Appl. Math., 15 (1967), pp. 419–425.
[23] W. Tollmien, *Ein Allgemeines Kriterium der Instabitität laminarer Geschwindigkeitsverteilungen*, Nachr. Ges. Wiss. Göttingen Math. Phys., 50 (1935), pp. 79–114.

SIAM J. MATH. ANAL.
Vol. 35, No. 2, pp. 357–375

ASYMPTOTIC BEHAVIOR TO DISSIPATIVE QUASI-GEOSTROPHIC FLOWS*

MARIA E. SCHONBEK† AND TOMAS P. SCHONBEK‡

Abstract. We consider the long time behavior of solutions of dissipative quasi-geostrophic (DQG) flows with subcritical powers. The flow under consideration is described by the nonlinear scalar equation

$$\frac{\partial\theta}{\partial t} + u\cdot\nabla\theta + \kappa(-\triangle)^{\alpha}\theta = f, \quad \theta|_{t=0} = \theta_0.$$

Rates of decay are obtained for both the solutions and higher derivatives in different Sobolev spaces.

Key words. quasi-geostrophic, decay, Fourier splitting

AMS subject classifications. 35Q35, 76B03

DOI. 10.1137/S0036141002409362

1. Introduction. In this paper we are concerned with the long time behavior of the solutions to a special case of surface two-dimensional dissipative quasi-geostrophic (DQG) flows with subcritical powers α:

$$\frac{\partial\theta}{\partial t} + u\cdot\nabla\theta + \kappa(-\triangle)^{\alpha}\theta = f, \qquad \theta|_{t=0} = \theta_0. \tag{1.1}$$

Here $\alpha \in (0,1]$, $\kappa > 0$, $\theta(t)$ is a real function of two space variables $x \in \mathbb{R}^2$ and a time variable t. The function $\theta(t) = \theta(x,t)$ represents the potential temperature. The fluid velocity u is determined from θ by a stream function ψ,

$$(u_1, u_2) = \left(-\frac{\partial\psi}{\partial x_2}, \frac{\partial\psi}{\partial x_1}\right), \tag{1.2}$$

where the function ψ satisfies

$$(-\triangle)^{\frac{1}{2}}\psi = -\theta.$$

Equation (1.1) is obtained when dissipative mechanisms are incorporated into the inviscid two-dimensional quasi-geostrophic (2DQG) equation. The 2DQG equation is derived from the general quasi-geostrophic (GQG) equations by reduction to the special case of solutions with constant potential vorticity in the interior and constant buoyancy frequency [3]. For information on the GQG equations we refer the reader to [8]. The fractional power $\alpha = 1/2$ is perhaps the most interesting one since it corresponds to a fundamental model of quasi-geostrophic equations; see [4] and [8]. As pointed out in [4], "Dimensionally the 2DQG equation with $\alpha = 1/2$ is the analogue of the 3D Navier–Stokes equations."

*Received by the editors June 10, 2002; accepted for publication (in revised form) February 21, 2003; published electronically July 18, 2003.
http://www.siam.org/journals/sima/35-2/40936.html
†Department of Mathematics, University of California at Santa Cruz, Santa Cruz, CA 95064 (schonbek@math.ucsc.edu).
‡Department of Mathematics, Florida Atlantic University, Boca Raton, FL 33431 (schonbek@fau.edu).

Two main problems will be considered. In the first, the power α will range in the interval $(\frac{1}{2}, 1]$. In this case when $\alpha \in (1/2, 1]$ it is known that the solutions are smooth on the torus; see [4]. In [16] Wu establishes regularity of solutions for certain type of data and forcing functions. Here we obtain smooth solutions in $\mathbb{R}^2$ by establishing uniform bounds in the H^m norms for solutions with appropriate data and forcing term. Interest will then be focused on the analysis of the asymptotic behavior of the energy of derivatives of all orders.

To establish decay in $H^m(\mathbb{R}^n)$ spaces the main tool will be the Fourier splitting method [11], [12]. This technique and others were used to treat solutions to parabolic conservation laws and Navier–Stokes equations. What makes the approach different here is that unlike the case of parabolic conservation laws and Navier–Stokes equations the dissipative mechanism is not given by a straightforward Laplacian but by a fractional power of the Laplacian, and new estimates are necessary. Before even addressing questions of decay new estimates are necessary to establish uniform bounds for the derivatives.

Some of the proofs presented in this paper consider only the case when $\alpha \in (1/2, 1]$; these proofs could be extended to the case $\alpha \in (0, 1]$, provided there was an a priori bound (possibly time dependent) of the derivatives of the solutions in the space L^2. In particular the estimate obtained by Wu in [16] could be used once a uniform bound on the $W^{1,\infty}$ norm of the velocity u is established.

The second question we address is the decay of the solutions in L^p. Given the decay in L^2 obtained in [4], the new H^m decay obtained in the first part of the paper will immediately yield, via a Gagliardo–Nirenberg inequality, decay in all L^p spaces with $p \geq 2$. Decay rate in L^p had already been obtained in [16], for $p > 1$. The problem now is to improve this decay by imposing conditions on the initial data which insure the decay of the L^1 norm of the solutions. Two cases are considered. First, the weak solution will be analyzed when $\alpha = 1/2$ and decay will be shown in L^1. Second, decay is established for solutions in $W^{q,p}$, $p \geq 2$ and $q \geq 1$, in the case where $\alpha \in (1/2, 1]$.

1.1. Notation and preliminaries. The Fourier transform of $v \in \mathcal{S}(\mathbb{R}^2)$ is defined by $\widehat{v}(\xi) = (2\pi)^{-1} \int_{\mathbb{R}^2} e^{-ix\cdot\xi} v(x)\, dx$. It is then extended as usual to $\mathcal{S}'$. Given a multi-index $\gamma = (\gamma_1, \gamma_2)$ and $m = |\gamma| = \gamma_1 + \gamma_2$, we denote

$$\partial^\gamma = \frac{\partial^{|\gamma|}}{\partial_{x_1}^{\gamma_1} \partial_{x_2}^{\gamma_2}}$$

and

$$D^m = \sum_{|\alpha|=m} \partial^\gamma.$$

If k is a nonnegative integer, $W^{k,p}(\mathbb{R}^2)$ will be, as is standard, the Sobolev space consisting of functions in $L^p(\mathbb{R}^2)$ whose generalized derivatives up to order k belong to $L^p(\mathbb{R}^2)$. As usual, when $p = 2$, then $W^{k,2}(\mathbb{R}^2) = H^k(\mathbb{R}^2)$, where (also as usual) the space H^s is defined for all $s \in \mathbb{R}$ as the space of all $f \in \mathcal{S}'$ such that $(1+|\xi|^2)^{s/2} \hat{f}(\xi) \in L^2$.

Following Constantin and Wu [4], we denote by

$$\Lambda = (-\triangle)^{\frac{1}{2}} \tag{1.3}$$

the operator defined by $\widehat{\Lambda f}(\xi) = |\xi|\hat{f}(\xi)$. More generally, if $s \geq 0$, we define Λ^s by

$$\widehat{\Lambda^s f}(\xi) = |\xi|^s \hat{f}(\xi).$$

Clearly $\Lambda^s f$ is well defined (and in L^2) if $f \in H^s$. More generally, one can define the domain of Λ^s as consisting of all elements $f \in \mathcal{S}'$ such that $\hat{f}$ is a function (i.e., locally integrable); it is then clear that the definition given above defines $\Lambda^s f$ as a tempered distribution.

We denote by $\mathcal{R}_1, \mathcal{R}_2$ the Riesz transforms in $\mathbb{R}^2$; i.e., $\widehat{\mathcal{R}_j f}(\xi) = -i(\xi_j/|\xi|)\hat{f}(\xi)$. The operator $\mathcal{R}^\perp$, taking scalar-valued functions to vector-valued functions, is defined by

$$\mathcal{R}^\perp f = (-\partial_{x_2}\Lambda^{-1} f, \partial_{x_1}\Lambda^{-1} f) = (-\mathcal{R}_2 f, \mathcal{R}_1 f). \tag{1.4}$$

The relation between u and θ in (1.1) can then briefly be stated as $u = \mathcal{R}^\perp\theta$.

If F is a function defined on $\mathbb{R}^2 \times [0,\infty)$, we define for $t \geq 0$ the function $F(t)$ on $\mathbb{R}^2$ by $F(t)(x) = F(x,t)$. For such F, the Fourier transform (and inverse Fourier transform) is always with respect to the space variables; thus

$$\hat{F}(\xi,t) = \widehat{F(t)}(\xi)$$

for all $t \geq 0$. The letters C, C_0, C_1, etc., will denote generic positive constants, which may vary from line to line during computations.

2. Uniform estimates. In this section we suppose $\alpha \in (1/2, 1]$. We show that $\Lambda^\beta\theta$ decays in the L^2 norm for $\beta \geq 0$; in particular we establish the uniform boundedness of the solution θ in H^m if the initial datum $\theta_0 \in H^m$. Our results in Theorem 2.4 can easily be adapted to the torus and as such extend those of Constantin and Wu [4, Theorem 2.1]. The decay we obtain in this section is not optimal but is needed to obtain the optimal rate of decay in the next section. In the last part of this section we establish uniform estimates on the L^∞ norms of the solutions. These estimates are obtained by bounding the L^1 norm of $\hat{\theta}$. We will need to use Theorem 3.1 from [4] and state it here for ease of reference.

THEOREM 2.1. *Let $\alpha \in (0,1]$ and $\theta_0 \in L^1 \cap L^2$. Assume that $f \in L^1([0,\infty); L^2)$, satisfying*

$$\|f(t)\|_2 \leq C_0(1+t)^{-\frac{1}{\alpha}-1}, \quad |\hat{f}(\xi,t)| \leq C_0|\xi|\alpha \tag{2.1}$$

for some constant C_0. Then there exists a weak solution θ of the 2DQG equation

$$\frac{\partial\theta}{\partial t} + u\cdot\nabla\theta + \kappa(-\triangle)^\alpha\theta = f, \quad \theta|_{t=0} = \theta_0 \tag{2.2}$$

such that

$$\|\theta(\cdot,t)\|_{L^2(\mathbb{R}^2)} \leq C(t+1)^{-\frac{1}{2\alpha}}, \tag{2.3}$$

where C is a constant depending on the L^1 and L^2 norms of θ_0, on the $L^1(L^2)$ norm of f, and on C_0.

We also need the following Sobolev type estimate.

LEMMA 2.2. *Let $2 < p < \infty$ and let $\sigma = 1 - \frac{2}{p}$. There exists a constant $C \geq 0$ such that if $f \in \mathcal{S}'$ is such that $\hat{f}$ is a function, then*

$$\|f\|_p \leq C\|\Lambda^\sigma f\|_2.$$

Proof. Since $\hat{f}$ is a function, we have $\hat{f}(\xi) = |\xi|^{-\sigma}|\xi|^{\sigma}\hat{f}(\xi)$. Taking the inverse Fourier transform, we get $f = I_\sigma(\Lambda^\sigma f)$, where I_σ is the Riesz potential of order σ. It is well known (cf. [13, Chapter V, Theorem 1]) that I_σ is bounded from $L^2(\mathbb{R}^2)$ to $L^p(\mathbb{R}^2)$ if $\frac{1}{p} = \frac{1}{2} - \frac{\sigma}{2}$. The lemma follows. □

Next, a simple observation connecting the L^2 norms of the temperature and the velocity (or transport term) that will be used repeatedly.

Remark 2.3. Let $1 < p < \infty$. There exists a constant C_p depending only on p such that

$$\|\Lambda^\beta u(t)\|_p \le C_p \|\Lambda^\beta \theta(t)\|_p \tag{2.4}$$

for all $\beta \ge 0$, $t \ge 0$. If $p = 2$, this inequality can be strengthened to

$$\|\Lambda^\beta u(t)\|_2 = \|\Lambda^\beta \theta(t)\|_2. \tag{2.5}$$

In fact, (2.4) is immediate from the fact that $u = \mathcal{R}^\perp\theta$, the fact that the Riesz transforms commute with Λ^β, and the boundedness of the Riesz transforms in L^p. Concerning (2.5), it suffices to observe that

$$\widehat{\Lambda^\beta u}(\xi, t) = \frac{i}{|\xi|}(\xi_2, \xi_1)|\xi|^\beta \hat{\theta}(\xi, t)$$

and the norm equality follows.

We are ready to state and prove the main result of this section. This first theorem gives a uniform bound for the derivatives of the solution $\theta(t)$ of the 2DQG and, for a sufficiently fast decaying f, an auxiliary rate of decay that will be improved in the next section.

THEOREM 2.4. *Let $\alpha \in (1/2, 1]$, $\beta \ge \alpha$, and assume q satisfies $2/(2\alpha - 1) < q < \infty$. Suppose $\theta_0 \in L^1 \cap L^2$, $\Lambda^\beta\theta_0 \in L^2$, $f \in L^1([0,\infty] : L^q \cap L^2)$ satisfies* (2.1) *and $\Lambda^{\beta-\alpha} f \in L^2((0,\infty), L^2)$. If θ is a solution to* (1.1) *with initial datum θ_0, then*

$$\|\Lambda^\beta\theta(t)\|_{L^2} \le C_0(1+t)^{-\frac{1}{2\alpha}} + C_1\left(\int_0^t \|\Lambda^{\beta-\alpha} f(s)\|_2^2\, ds\right)^{1/2} \tag{2.6}$$

for $t \ge 0$, where C_0, C_1 are constants depending only on norms of the initial datum and f. In particular, if $f = 0$, then

$$\|\Lambda^\beta\theta(t)\|_{L^2} \le C_0(1+t)^{-\frac{1}{2\alpha}} \tag{2.7}$$

for all $t \ge 0$.

Remark 2.5. In [4, Theorem 2.1] the authors assume, in case $\beta < 1$, that $q = 2/(1-\beta)$. This choice is consistent with our more general one, since it is also assumed in [4] that $\beta + 2\alpha > 2$, which implies $2/(1-\beta) > 2/(2\alpha - 1)$. The assumption $\theta_0 \in L^1 \cap L^2$ (as well as f satisfying (2.1)) is needed to apply Theorem 2.1.

Proof. The first part of the proof we present is formal. At the end of the proof we give a sketch on how to make the arguments rigorous. To obtain (2.6) multiply both sides of (1.1) by $\Lambda^{2\beta}\theta(t)$ and integrate in space:

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^2} |\Lambda^\beta\theta(t)|^2\, dx + \kappa \int_{\mathbb{R}^2} |\Lambda^{\alpha+\beta}\theta(t)|^2\, dx \\ &\quad = -\int_{\mathbb{R}^2} (u\cdot\nabla\theta)\Lambda^{2\beta}\theta\, dx + \int_{\mathbb{R}^2} f\Lambda^{2\beta}\theta.\, dx. \end{aligned} \tag{2.8}$$

We estimate the second term on the right-hand side of the last equation by

$$\int_{\mathbb{R}^2} f\Lambda^{2\beta}\theta\,dx \le \frac{\kappa}{8}\int_{\mathbb{R}^2}|\Lambda^{\alpha+\beta}\theta(t)|^2\,dx + \frac{2}{\kappa}\int_{\mathbb{R}^2}|\Lambda^{\beta-\alpha}f|^2\,dx. \tag{2.9}$$

Estimating the first term will take a little longer. We claim that there exists a constant $C(\kappa,\theta_0,f)$, depending only on the initial datum θ_0, the $L^1(0,\infty,L^q)$ norm of the external force f, and κ such that

$$\left|\int_{\mathbb{R}^2}(u\cdot\nabla\theta)\Lambda^{2\beta}\theta\,dx\right| \le \frac{\kappa}{8}\|\Lambda^{\alpha+\beta}\theta\|_2^2 + C(\theta_0,f,\kappa)\|\Lambda^{s+1-(2/p)}\theta\|_2^2, \tag{2.10}$$

where $s=\beta-\alpha+1$, p is determined by $\frac{1}{p}+\frac{1}{q}=\frac{1}{2}$, and q is as in the statement of the theorem. The meaning of s,p,q will not change for the remainder of this proof. To establish the claim, we begin observing that because $\operatorname{div} u=0$ we can write

$$u\cdot\nabla\theta = \operatorname{div}(u\theta) - \theta\operatorname{div}u = \operatorname{div}(u\theta);$$

thus, by Plancherel, Hölder, and again Plancherel,

$$\begin{aligned}\left|\int_{\mathbb{R}^2}(u\cdot\nabla\theta)\Lambda^{2\beta}\theta\,dx\right| &= \left|\int_{\mathbb{R}^2}(\xi_1\widehat{\theta u_1}(\xi)+\xi_1\widehat{\theta u_2}(\xi))|\xi|^{2\beta}\hat\theta(\xi)\,d\xi\right| \\ &\le \sum_{i=1}^{2}\int_{\mathbb{R}^2}|\xi|^{\beta-\alpha+1}|\widehat{\theta u_i}(\xi)||\xi|^{\alpha+\beta}|\hat\theta(\xi)|\,d\xi \\ &\le \sum_{i=1}^{2}\|\Lambda^{\beta-\alpha+1}(\theta u_i)\|_2\|\Lambda^{\alpha+\beta}\theta\|_2;\end{aligned}$$

hence

$$\left|\int_{\mathbb{R}^2}(u\cdot\nabla\theta)\Lambda^{2\beta}\theta\,dx\right| \le \frac{\kappa}{8}\|\Lambda^{\alpha+\beta}\theta\|_2^2 + \frac{2}{\kappa}\sum_{i=1}^{2}\|\Lambda^s(\theta u_i)\|_2^2. \tag{2.11}$$

We estimate $\|\Lambda^s(\theta u_i)\|_2$ by the calculus inequality, getting

$$\|\Lambda^s(\theta u_i)\|_2 \le C\left(\|u_i\|_q\|\Lambda^s\theta\|_p + \|\theta\|_q\|\Lambda^s u_i\|_p\right)$$

for $i=1,2$. This inequality follows easily by combining Hölder's inequality with the Gagliardo–Nirenberg and Young inequalities; see also inequality (3.1.59) on page 74 of [14]. Since $u_i=\pm\mathcal{R}_j\theta$ $(i,j\in\{1,2\},\ i\ne j)$ and the Riesz transforms commute with Λ and are bounded in L^p, L^q (notice that $2<p,q<\infty$), we have $\|\Lambda^s u_i\|_p \le C\|\Lambda^s\theta\|_p$ and $\|u_i\|_q \le C\|\theta\|_q$ for $i=1,2$. Applying this to the previous estimate, we get

$$\|\Lambda^s(\theta u_i)\|_2 \le C\|\theta\|_q\|\Lambda^s\theta\|_p \tag{2.12}$$

for $i=1,2$. To continue, we estimate $\|\theta\|_q$ by the following maximum principle:

$$\|\theta\|_{L^q} \le \|\theta_0\|_{L^q} + \int_0^t\|f(\tau)\|_{L^q}\,d\tau. \tag{2.13}$$

For details on this inequality and its proof we refer the reader to [10], [1], but we briefly describe the main idea, as given by Wu [15]. Specifically, (2.13) follows by multiplying both sides of (1.1) by $q|\theta|^{q-2}\theta$ and integrating with respect to x to get

$$\frac{d}{dt}\|\theta\|_{L^q}^q \le q\left(\int|\theta|^{q-2}\theta f\,dx - \int|\theta|^{q-2}\theta(u\cdot\nabla\theta)\,dx - \int|\theta|^{q-2}\theta\kappa(-\triangle)^\alpha\theta)\,dx\right).$$

One sees that the second integral on the right is zero. The last integral on the right can be shown to be positive [10], [15]. Thus

$$\frac{d}{dt}\|\theta\|_{L^q}^q \le q\int |\theta|^{q-2}\theta f\,dx \le q\|f\|_{L^q}\|\theta\|_{L^q}^{q-1},$$

and (2.13) follows. Because $f \in L^1(0,\infty;L^q)$, we proved that

$$\|\Lambda^s(\theta u_i)\|_2 \le C(\theta_0, f)\|\Lambda^s\theta\|_p$$

for $i = 1, 2$, where $C_0(\theta_0, f)$ is independent of t, depends only on θ_0, f. By Lemma 2.2,

$$\|\Lambda^s(\theta u_i)\|_2 \le C(\theta_0, f)\|\Lambda^{s+1-(2/p)}\theta\|_2$$

for $i = 1, 2$. Using this in (2.11), (2.10) follows, establishing our claim, with $C(\kappa, \theta_0, f) = \frac{4}{\kappa}C(\theta_0, f)^2$. Combining (2.8), (2.9), and (2.10) yields

$$\frac{1}{2}\frac{d}{dt}\|\Lambda^\beta\theta(t)\|_2^2 + \frac{3\kappa}{4}\|\Lambda^{\alpha+\beta}\theta(t)\|_2^2 \le C_0\|\Lambda^\gamma\theta\|_2^2 + \frac{2}{\kappa}\|\Lambda^{\beta-\alpha}f\|_2^2, \tag{2.14}$$

where $C_0 = C(\kappa, \theta_0, f)$ and we introduced $\gamma = s+1-\frac{2}{p} = \beta-\alpha+2(1-\frac{1}{p})$. To continue estimating, let $B_M = \{\xi : |\xi|^2 \le M\}$, with $M > 0$ to be determined appropriately below. The choice $2/(2\alpha-1) < q < \infty$ implies $\frac{1}{2} > \frac{1}{p} = \frac{1}{2} - \frac{1}{q} > 1-\alpha$, and hence $\frac{1}{p} + \alpha - 1 > 0$. Thus $\gamma = \alpha+\beta-2(\frac{1}{p}+\alpha-1) < \alpha+\beta$ and

$$\begin{aligned}\|\Lambda^\gamma\theta(t)\|_2^2 dx &= \int_{B_M} |\xi|^{2\gamma}|\hat\theta(t)|^2\,d\xi + \int_{B_M^c} |\xi|^{2\gamma}|\hat\theta(t)|^2\,d\xi \\ &\le M^{2\gamma}\|\theta(t)\|_2^2 + M^{-4(\frac{1}{p}+\alpha-1)}\|\Lambda^{\alpha+\beta}\theta(t)\|_2^2.\end{aligned}$$

Selecting M large enough to satisfy $M^{-4(\frac{1}{p}+\alpha-1)} < \kappa/(4C_0)$, it follows that

$$C_0\|\Lambda^\gamma\theta(t)\|_2^2 dx \le \frac{\kappa}{4}\|\Lambda^{\alpha+\beta}\theta(t)\|_2^2 + C_0M^{2\gamma}\|\theta(t)\|_2^2. \tag{2.15}$$

Next,

$$\begin{aligned}\|\Lambda^{\alpha+\beta}\theta\|_2^2 &\ge \int_{B_M^c} |\xi|^{2(\alpha+\beta)}|\hat\theta|^2\,d\xi \ge M^{2\alpha}\int_{B_M^c} |\xi|^{2\beta}|\hat\theta|^2\,d\xi \\ &= M^2\alpha\|\Lambda^\beta\theta\|_2^2 - M^{2\alpha}\int_{B_M} |\xi|^{2\beta}|\hat\theta|^2\,d\xi,\end{aligned}$$

implying

$$\|\Lambda^{\alpha+\beta}\theta\|_2^2 \ge M^2\alpha\|\Lambda^\beta\theta\|_2^2 - M^{2(\alpha+\beta)}\|\theta(t)\|_2^2. \tag{2.16}$$

By Theorem 3.1 in [4] (stated as Theorem 2.1 in this article), $\|\theta(t)\|_2$ decays at the rate of $(1+t)^{-1/\alpha}$. Using this estimate in (2.15) and (2.16), and then returning to (2.14), we get

$$\frac{d}{dt}\|\Lambda^\beta\theta(t)\|_2^2 + \kappa M^{2\alpha}\|\Lambda^\beta\theta(t)\|_2^2 \le \tilde C_0 M^c(1+t)^{-1/a} + \frac{2}{\kappa}\|\Lambda^{\beta-\alpha}f(t)\|_2^2, \tag{2.17}$$

where $\tilde{C}_0$ is a new constant depending only on f, θ_0, κ and $c = \max(2\gamma, 2\alpha + 2\beta)$. For convenience, let $\nu = \kappa M^{2\alpha}$. Multiplying both sides of (2.17) by $e^{\nu t}$ and integrating in time we see that

$$\|\Lambda^\beta \theta(t)\|_2^2 \le e^{-\nu t}\|\Lambda^\beta \theta_0\|_2^2 + \tilde{C}_0 M^c \int_0^t e^{-\nu(t-s)}(s+1)^{-\frac{1}{\alpha}}\,ds + \frac{2}{\kappa}\int_0^t e^{-\nu(t-s)}\|\Lambda^{\beta-\alpha} f(s)\|_2^2\,ds.$$

The desired estimate (2.6) now follows, since

$$\int_0^t e^{-\nu(t-s)}(1+s)^{-\frac{1}{\alpha}}\,ds \le C(1+t)^{-\frac{1}{\alpha}}, \tag{2.18}$$

$$\int_0^t e^{-\nu(t-s)}H(s)\,ds \le \int_0^t H(s)\,ds \tag{2.19}$$

for all $t \ge 0$, some C.

This completes the formal part of the proof. To make the above arguments rigorous, apply the same proof to the "retarded mollifications θ_n," which are solutions of the sequence of approximate equations

$$\frac{\partial \theta_n}{\partial t} + u_n \cdot \nabla \theta_n + \kappa(-\triangle)^\alpha \theta_n = f, \tag{2.20}$$

where $u_n = \Psi_{\delta_n}(\theta_n)$ is obtained from θ_n by

$$\Psi_{\delta_n}(\theta_n) = \int_0^t \phi(\tau)\mathcal{R}^\perp \theta_n(t - \delta_n \tau)\,d\tau, \tag{2.21}$$

and $\mathcal{R}^\perp$ is defined by (1.4).

The function ϕ is smooth and has support in $[1,2]$, and $\int_0^\infty \phi(t)\,dt = 1$. This construction is similar to the one used by Caffarelli, Kohn, and Nirenberg in [2] for solutions to the Navier–Stokes equations. It is easy to see that for each n the values of u_n depend only on the values of θ_n in $[t - 2\delta_n, t - \delta_n]$. As stated in [4] the θ_n converge to a weak solution θ and strongly in L^2 almost everywhere in t. Since the bounds for the $\Lambda^\beta \theta_n$ are independent of n it follows that they hold for the limiting solution θ.

This concludes the proof of the theorem. □

Remark 2.6. In proving Theorem 2.4 we estimated (see (2.19))

$$\int_0^t e^{-\nu(t-s)}\|\Lambda^{\beta-\alpha} f(s)\|_2^2\,ds \le \int_0^t \|\Lambda^{\beta-\alpha} f(s)\|_2^2\,ds$$

to get the second term on the right-hand side of (2.6). The assumption $f = 0$ then causes the L^2 norm of $\Lambda^\beta \theta(t)$ to decay in time. However, by (2.18), it follows that we have decay of this norm as long as $\|\Lambda^{\beta-\alpha} f\|_2$ decays fast enough. For example, if

$$\|\Lambda^{\beta-\alpha} f(t)\|_2 \le C(1+t)^{-\delta}$$

for some $\delta > 0$, then (2.7) can be replaced by

$$\|\Lambda^\beta \theta(t)\|_{L^2} \le C_0(1+t)^{-\min(\frac{1}{2\alpha}, \delta)}, \tag{2.22}$$

where C_0 only depends on f and the initial datum θ_0.

The remainder of this section deals with obtaining L^∞ bounds of the solution—more precisely, L^1 bounds of the Fourier transform of the solution. If the hypotheses of Theorem 2.4 are satisfied with $\beta > 1$, it is clear that $\hat{\theta}(t) \in L^1$ and $\|\hat{\theta}(t)\|_1$ is uniformly bounded in t. In fact, $\theta \in L^2 \cap \tilde{H}^\beta = H^\beta$, and hence

$$\int_{\mathbb{R}^2} |\hat{\theta}(\xi)|\, d\xi \le C \left(\int_{\mathbb{R}^2} (1+|\xi|^2)^\beta |\hat{\theta}(\xi)|^2\, d\xi \right)^{1/2}$$

with

$$C = \left(\int_{\mathbb{R}^2} (1+|\xi|^2)^{-\beta} |\, d\xi \right)^{1/2} < \infty.$$

In the next lemma, we show that we also have $\hat{\theta}(t) \in L^1$, with a uniformly bounded L^1 norm, if $\beta = 1$.

The next lemma gives an a priori bound of the L^1 norm of $\hat{\theta}(t)$. It then suffices to establish a local existence theorem to obtain a global uniform bound.

LEMMA 2.7 (a priori bound). *Assume the hypothesis of Theorem* 2.4 *with* $\beta \ge 1$. *If* $\beta = 1$, *assume also that* $\hat{\theta}_0 \in L^1$ *and that* $\hat{f} \in L^1(0,\infty,L^1)$. *It follows that there exists* $C \ge 0$ *such that*

$$\|\hat{\theta}(t)\|_1 \le C$$

for all $t \ge 0$

Remark 2.8. The hypothesis on f in case $\beta = 1$ can be considerably weakened, but the proof becomes somewhat more involved.

Proof. Since we only want an a priori bound the proof is formal. The case $\beta > 1$ was dealt with in the remarks preceding this lemma; we assume from now on that $\beta = 1$. By Theorem 2.4, there exists $C \ge 0$ such that

$$\|\nabla\theta(t)\|_2 = \|\Lambda\theta(t)\|_2 \le C$$

for all $t \ge 0$. An easy calculation yields

$$\hat{\theta} = e^{-\kappa|\xi|^{2\alpha}t}\hat{\theta_0} - \int_0^t e^{-\kappa|\xi|^{2\alpha}(t-s)} \widehat{u\cdot\nabla\theta}\, ds + H(t),$$

where

$$H(t) = \int_0^t e^{-\kappa|\xi|^{2\alpha}(t-s)} \hat{f}(s)\, ds.$$

By the additional hypothesis on f, it is obvious that $H(t)$ is uniformly bounded in the L^1 norm. Hence

$$\|\hat{\theta}(t)\|_1 \le \|\hat{\theta_0}\|_1 + \int_0^t \|e^{-\kappa|\xi|^{2\alpha}(t-s)} \widehat{u\cdot\nabla\theta}\|_1\, ds + C, \tag{2.23}$$

where C is chosen so that $\|H(t)\|_1 \le C$ for all $t \ge 0$. Since the first term of the right-hand side of (2.23) is bounded by hypothesis, we need to bound only the second

term. For this purpose, we split it into two parts for an appropriate value of $\epsilon > 0$ as follows.

$$\int_0^t \|e^{-\kappa|\xi|^{2\alpha}(t-s)}\widehat{u\cdot\nabla\theta}\|_1\,ds = I + II,$$

where, if $t \geq \epsilon$,

$$I = \int_0^{t-\epsilon} \|\widehat{u\cdot\nabla\theta}\|_1\,ds,$$
$$II = \int_{t-\epsilon}^t \|e^{-\kappa|\xi|^{2\alpha}(t-s)}\widehat{u\cdot\nabla\theta}\|_1\,ds;$$

if $0 \leq t < \epsilon$, then $I = 0$ and $II = \int_0^t \|\widehat{u\cdot\nabla\theta}\|_1\,ds$. We begin bounding II, assuming $t > \epsilon$.

$$II \leq \int_{t-\epsilon}^t \|e^{-\kappa|\xi|^{2\alpha}(t-s)}\|_2\|\widehat{u\cdot\nabla\theta}\|_2\,ds \leq C\int_{t-\epsilon}^t \frac{1}{(t-s)^{\frac{1}{2\alpha}}}\|\nabla\theta\|_2\|u\|_\infty\,ds$$
$$\leq C\sup_{t\geq 0}\|\nabla\theta(t)\|_2 \sup_{0\leq s\leq t}\|\hat\theta(s)\|_1\epsilon^{1-\frac{1}{2\alpha}},$$

where we used the fact that $\|u(t)\|_\infty \leq C\|\hat u(t)\|_1 \leq C\|\hat\theta(t)\|_1$, since the components of $\hat u$ are obtained multiplying $\hat\theta$ by functions of absolute value 1. Since $\|\nabla\theta(t)\|_2$ is bounded in t, we can select $\epsilon > 0$ so that

$$II \leq \frac{1}{2}\sup_{0\leq s\leq t}\|\hat\theta(s)\|_1 \tag{2.24}$$

for all $t \geq \epsilon$. Assuming now $t < \epsilon$, we estimate essentially the same way to get

$$II \leq C\sup_{t\geq 0}\|\nabla\theta(t)\|_2 \sup_{0\leq s\leq t}\|\hat\theta(s)\|_1\int_0^t (t-s)^{-\frac{1}{2\alpha}}\,ds \leq C\sup_{0\leq s\leq t}\|\hat\theta(s)\|_1\epsilon^{1-\frac{1}{2\alpha}}.$$

Decreasing the size of $\epsilon > 0$ if necessary, we can assume that (2.24) also holds for $0 < t < \epsilon$.

To bound I, we use the fact that $\|u(s)\|_2 = \|\theta(s)\|_2 \leq C(1+s)^{-1/2\alpha} \leq C$, $\|\nabla\theta(s)\|_2 \leq C$ for all $s \geq 0$ (some constant C). We assume $t \geq \epsilon$ (otherwise $I = 0$).

$$I \leq \int_0^{t-\epsilon} \|e^{-\kappa|\xi|^{2\alpha}(t-s)}\|_1\|\widehat{u\cdot\nabla\theta}\|_\infty\,ds \leq C\int_0^{t-\epsilon}\frac{1}{(t-s)^{\frac{1}{\alpha}}}\|\widehat{u\cdot\nabla\theta}\|_\infty\,ds$$
$$\leq C\int_0^{t-\epsilon}\frac{1}{(t-s)^{\frac{1}{\alpha}}}\|u\|_2\|\nabla\theta\|_2\,ds \leq C\int_0^{t-\epsilon}\frac{1}{(t-s)^{\frac{1}{\alpha}}}\,ds.$$

The last integral is bounded by $C\epsilon^{1-\frac{1}{\alpha}}$ if $\alpha < 1$, by $C\log(1/\epsilon)$ if $\alpha = 1$, and in either case by a constant since ϵ has been fixed. In other words I is bounded independently of t; using this and (2.24) in (2.23), we get

$$\|\hat\theta(t)\|_1 \leq C + \frac{1}{2}\sup_{0\leq s\leq t}\|\hat\theta(s)\|_1$$

for all $t \geq 0$, C independent of t. The lemma follows. □

COROLLARY 2.9. *Under the hypotheses of Lemma* 2.7 *one has that* $\|\theta(t)\|_\infty$, $\|\hat{u}(t)\|_1$, *and* $\|u(t)\|_\infty$ *are uniformly bounded in* t.

Proof. Since the components of u are Riesz transforms of θ (hence, the components of $\hat{u}$ differ from $\hat{\theta}$ by factors of absolute value 1), it is immediate from Lemma 2.7 that $\|\hat{u}(t)\|_1$ is uniformly bounded in time. The uniform bound on the L^∞ norms now follows. □

The next lemma gives the local existence for solutions with data θ_0, where $\theta_0 \in H^1$ and $\hat{\theta_0} \in L^1$

LEMMA 2.10. *Let* $\theta_0 \in H^1$ *and* $\hat{\theta_0} \in L^1$, *and let* f *satisfy the hypothesis of Theorem* 2.4. *Let* $\alpha \in (1/2, 1]$, $\beta = 1$. *Then there exists* $T > 0$ *and a solution* θ *of* (1.1) *such that* $\theta \in L^\infty([0,T] : H^1)$ *and* $\hat{\theta} \in L^\infty([0,T] : L^1)$.

Proof. The proof follows by a straightforward application of the contraction mapping theorem to the sequence of solutions of the equations

$$\frac{\partial \theta_n}{\partial t} + (-\mathcal{R}_2\theta_{n-1}, \mathcal{R}_1\theta_{n-1}) \cdot \nabla\theta_n + \kappa(-\triangle)^\alpha \theta_n = f,$$
$$\theta|_{t=0} = \theta_0. \qquad \square$$

THEOREM 2.11. *Under the conditions of Theorem* 2.4 *there exists a global solution* $\theta \in L^\infty([0,\infty) : H^1)$ *such that* $\hat{\theta} \in L^\infty([0,\infty) : L^1)$.

Proof. Combine the two last lemmas. □

3. H^m and fractional derivative decay. In this section we improve the decay of the derivatives of order β of the solution θ of (1.1), assuming the external force $f = 0$. The decay established in the last section is not optimal but does provide the stepping stone to obtain the optimal decay, that is, a decay rate which coincides with that of the underlying linear part. The main tool used is the Fourier splitting method (see [11], [12]). The solutions considered here are supposed to be smooth. The assumption that the external force is zero is not essential. The same results can be obtained when $f \neq 0$, provided $\|\Lambda^{\beta-\alpha} f\|_2$ decays sufficiently fast (see Remark 2.6 above and Corollary 3.4 at the end of this section). The proof is the same as the one presented below, with the addition of a term that decays sufficiently fast by hypothesis.

We assume throughout this section that $\alpha \in (\frac{1}{2}, 1]$, $m \geq \alpha$, and θ is the solution of (1.1) (with $f = 0$ until further notice) such that $\theta_0 = \theta(0)$ satisfies $\theta_0 \in L^1(\mathbb{R}^2) \cap H^m(\mathbb{R}^2)$. The hypotheses of Theorem 2.4 are thus satisfied for any $\beta \in [\alpha, m]$. The numbers p, q are as in the previous section: $\frac{1}{p} + \frac{1}{q} = \frac{1}{2}$, $0 < \frac{1}{q} < \alpha - \frac{1}{2}$.

Before improving the rate of decay of the derivatives of θ, we state some of the immediate consequences of Theorem 2.4.

COROLLARY 3.1. *Under the assumptions mentioned above, the following estimates hold for* $t \geq 0$:

$$\|\theta(t)\|_{H^m} \leq C(1+t)^{-\frac{1}{2\alpha}}, \tag{3.1}$$

$$\|u(t)\|_{H^m} \leq C(1+t)^{-\frac{1}{2\alpha}}, \tag{3.2}$$

with C *a constant depending only on norms of the initial datum; moreover, if* $m \geq 1$, *with* r *any exponent in* $[2, \infty)$, *then*

$$\|\theta(t)\|_r \leq C_r(1+t)^{-\frac{1}{2\alpha}}, \quad 0 \leq \gamma \leq \beta - 1, \tag{3.3}$$

$$\|\Lambda^\gamma u(t)\|_r \leq C_r(1+t)^{-\frac{1}{2\alpha}}, \quad 0 \leq \gamma \leq \beta - 1, \tag{3.4}$$

with C_r a constant depending only on norms of the initial datum and r.

Proof. Since $H^m = \{g \in L^2 : \Lambda^m g \in L^2\}$, inequality (3.1) is immediate from Theorem 2.4; inequality (3.2) follows then from Remark 2.3. Inequalities (3.3), (3.4) follow from these and Sobolev's theorem. □

The next theorem will give the optimal rate of decay for the derivatives in the sense that it coincides with the decay rate of the underlying linear part.

THEOREM 3.2. *Assume θ is a solution of* (2.2) *with data $\theta_0 \in L^1 \cap H^m$. Then*

$$\|\Lambda^\beta \theta(t)\|_{L^2} \le C(t+1)^{-\frac{\beta+1}{2\alpha}}, \tag{3.5}$$

where C is a constant which depends only on the norms of the initial datum.

Proof. The proof is based on an appropriately modified Fourier splitting method, combined with the preliminary estimates of the last section. We will assume $\alpha < 1$; refer to [5] for the case $\alpha = 1$.

Assume $\alpha \le \beta \le m$. We return to the derivation of inequality (2.14) in the proof of Theorem 2.4, recalling that $C(\kappa, \theta_0, f) = \frac{4}{\kappa} C(\theta_0, f)^2$ and $C(\theta_0, f)$ was a bound for $\|\theta(t)\|_q$ given by the maximum principle. If we forego this bound, we obtain directly, for some constant C_1, all $t \ge 0$,

$$\frac{1}{2}\frac{d}{dt}\|\Lambda^\beta\theta(t)\|_2^2 + \frac{3\kappa}{4}\|\Lambda^{\alpha+\beta}\theta(t)\|_2^2 \le C_1\|\theta(t)\|_q^2\|\Lambda^{\beta-\alpha+1-\frac{2}{p}}\theta(t)\|_2^2. \tag{3.6}$$

Because $\beta < \beta - a + 1 - \frac{2}{p} < \alpha + \beta$, with $\delta > 0$ such that $\beta - a + 1 - \frac{2}{p} = (1-\delta)\beta + \delta(\alpha + \beta)$, we have

$$\|\Lambda^{\beta-\alpha+1-\frac{2}{p}}\theta(t)\|_2^2 \le \|\Lambda^\beta\theta(t)\|_2^{2(1-\delta)}\|\Lambda^{\alpha+\beta}\theta(t)\|_2^{2\delta} \le \frac{\kappa}{4C_0}\|\Lambda^\beta\theta(t)\|_2^2 + C_2\|\Lambda^{\alpha+\beta}\theta(t)\|_2^2,$$

where we take C_0 so that $C_1\|\theta(t)\|_q^2 \le C_0$ for all $t \ge 0$, C_2 being determined by this choice. Inequality (3.6) can be modified to

$$\frac{1}{2}\frac{d}{dt}\|\Lambda^\beta\theta(t)\|_2^2 + \frac{\kappa}{2}\|\Lambda^{\alpha+\beta}\theta(t)\|_2^2 \le C_2\|\theta(t)\|_q^2\|\Lambda^\beta\theta(t)\|_2^2. \tag{3.7}$$

For $t \ge 0$ set

$$S(t) = \left\{\xi : |\xi|^{2\alpha} \le \frac{\mu}{\kappa(t+1)}\right\},$$

where μ is chosen so that $\mu > \frac{\beta+1}{\alpha} + 1$. Then

$$\begin{aligned}
\|\Lambda^{\alpha+\beta}\theta(t)\|_2^2 &= \int_{\mathbb{R}^2} |\xi|^{2(\alpha+\beta)}|\theta(t)|^2\,d\xi \\
&\ge \frac{2\mu}{3\kappa(t+1)}\int_{S(t)^c} |\xi|^{2\beta}|\theta(t)|^2\,d\xi \\
&= \frac{2\mu}{3\kappa(t+1)}\left(\|\Lambda^\beta\theta(t)\|_2^2 - \int_{S(t)} |\xi|^{2\beta}|\theta(t)|^2\,d\xi\right) \\
&\ge \frac{2\mu}{3\kappa(t+1)}\left(\|\Lambda^\beta\theta(t)\|_2^2 - \left(\frac{2\mu}{3\kappa(t+1)}\right)^{\frac{\beta}{\alpha}}\|\theta(t)\|_2^2\right);
\end{aligned}$$

estimating $\|\theta(t)\|_2$ by const$\cdot(1+t)^{-1/(2\alpha)}$ and putting it into (3.7) we get, after we estimate the factor $\|\theta(t)|_q^2$ in (3.7) by $C(t+1)^{-\frac{1}{\alpha}}$ (Corollary 3.1),

$$\frac{1}{2}\frac{d}{dt}\|\Lambda^\beta\theta(t)\|_2^2+\frac{\mu}{2(t+1)}\|\Lambda^\beta\theta(t)\|_2^2\le C(t+1)^{-\frac{1}{\alpha}}\|\Lambda^\beta\theta\|_2^2+C(t+1)^{-\frac{\beta+1}{\alpha}-1}. \tag{3.8}$$

Assume proved, for some λ, $0<\lambda<(\beta+1)/\alpha$, some $C\ge 0$, and all $t\ge 0$, that

$$\|\Lambda^\beta\theta\|_2^2\le C(t+1)^{-\lambda}. \tag{3.9}$$

Then using this in (3.8) we obtain, after multiplying by the integrating factor $2(t+1)^\mu$,

$$\frac{d}{dt}\left((t+1)^\mu\|\Lambda^\beta\theta(t)\|_2^2\right)\le C(t+1)^{\mu-\frac{1}{\alpha}-\lambda}+C(t+1)^{\mu-\frac{\beta+1}{\alpha}-1}.$$

Integrating from 0 to t, and then dividing by $(t+1)^{-\mu}$,

$$\|\Lambda^\beta\theta(t)\|_2^2\le\|\Lambda^\beta\theta(0)\|_2^2+C)(t+1)^{-\mu}+C(t+1)^{1-\frac{1}{\alpha}-\lambda}+C(t+1)^{-\frac{\beta+1}{\alpha}}.$$

It follows that in (3.9) we can replace λ by $\min(\lambda+\frac{1}{\alpha}-1,\frac{\beta+1}{\alpha})$. Since $\frac{1}{\alpha}-1>0$, we are done. □

COROLLARY 3.3. *Under the conditions of the last theorem it follows that the solutions to 2DQG equations have the decay in L^p*

$$\|D^j u\|_p\le C_p(t+1)^{-\frac{1}{\alpha}[\frac{j+2}{2}-\frac{1}{p}]}.$$

Proof. Use the estimates in Theorem 3.2 and [4] combined with a Gagliardo–Nirenberg inequality.

$$\|D^j u\|_p\le C_p\|u\|_2^{1-a}\|D^{j+1}\|_2^a, \tag{3.10}$$

where $a=1-\frac{2}{j+1}\frac{1}{p}$. Thus

$$\|D^j u\|_p\le C_p(t+1)^{-[(1-a)\frac{1}{2\alpha}+a\frac{j+2}{2\alpha}]}. \tag{3.11}$$

Replacing a with its definition gives the expected decay. □

In the case that $f\ne 0$ we can obtain the same results of Theorem 3.2, provided f decays at the appropriate rate. We state this more precisely in the following corollary.

COROLLARY 3.4. *Under the conditions of Theorem* 3.2*, suppose f satisfies* (2.1) *and*

$$\|\Lambda^{\beta-\alpha}f(\cdot,t)\|_2^2\le C(1+t)^{-\frac{\beta+1}{2}-1}. \tag{3.12}$$

If θ is a solution to (2.2) *with data θ_0, then*

$$\|\Lambda^\beta\theta(t)\|_{L^2}\le Ct^{-\frac{\beta+1}{2\alpha}}, \tag{3.13}$$

where C is constant which depends only on the L^2 norm of the data and f.

Proof. The proof follows the same steps of the last theorem. □

4. L^1 and improved L^p decay. In this section we consider the decay in L^p spaces for $p\in[1,\infty]$. New conditions on the data will be necessary to insure decay of the solutions in the L^p norms when $p\in[1,2)$, mainly that a Riesz potential of the data lies in the corresponding L^p space.

We first consider the L^1 decay of the solutions for the special case when $\alpha=1/2$. In the more general case when $\alpha\in(\frac{1}{2},1)$ the L^1 decay for derivatives of higher order will be obtained. The case of $\alpha=1$ is the easiest since the linear part is the heat equation.

4.1. Linear asymptotics. Let $\alpha \in (0,1]$, $\kappa > 0$. We consider the linear equation

$$\frac{\partial \theta}{\partial t} + \kappa(-\triangle)^{\alpha}\theta = 0 \tag{4.1}$$

in $\mathbb{R}^2 \times \mathbb{R}$; the solution $\theta = \theta(x,t)$ is a function of a space variable $x \in \mathbb{R}^2$ and a time variable $t \geq 0$. Without loss of generality, we assume $\kappa = 1$.

The function G_α will be defined for $\alpha \in (0,1]$ by

$$\hat{G}_\alpha(\xi,t) = e^{-|\xi|^{2\alpha}t}.$$

The solution θ of (4.1) with initial datum θ_0 is then given by

$$\theta(t) = e^{t\Lambda^{2\alpha}}\theta_0 = G_a(t) * \theta_0.$$

We recall once again that if $0 < \beta < 2$, the Riesz potential I_β is defined in the Fourier variables by

$$\widehat{(I_\beta w)}(\xi) = \frac{\widehat{w}(\xi)}{|\xi|^\beta}.$$

Then we can write

$$\partial^\gamma \theta(t) = (\partial^\gamma \Lambda^\beta G_a)(t) * (I_\beta \theta_0). \tag{4.2}$$

By a standard change of variables, since $n = 2$, it follows that

$$(\partial^\gamma \Lambda^\beta G_\alpha)(x,t) = t^{-(\frac{\beta}{2\alpha}+\frac{|\gamma|}{2\alpha}+\frac{1}{\alpha})}(\partial^\gamma \Lambda^\beta G_\alpha)(t^{-\frac{1}{2\alpha}}x, 1). \tag{4.3}$$

Hence, by the Hausdorff–Young inequality,

$$\|\partial^\gamma \theta(t)\|_p \leq t^{-(\frac{\beta}{2\alpha}+\frac{|\gamma|}{2\alpha}+\frac{1}{\alpha}[1-\frac{1}{p}])}\|\partial^\gamma \Lambda^\beta G_\alpha(1)\|_p \|I_\beta \theta_0\|_1 \tag{4.4}$$

for all $t \geq 0$, $1 \leq p \leq \infty$. Thus, in order to establish the L^p decay of $\partial^\gamma\theta(t)$ it will suffice to prove that $\partial^\gamma \Lambda^\beta G_\alpha(1)$ is in L^p. We do this in the next lemma.

LEMMA 4.1. *Assume $\alpha \geq \frac{1}{2}$ and let $p \in [1,\infty]$. Then $\partial^\gamma \Lambda^\beta G_\alpha(1) \in L^p$ for all $\beta \geq 0$ and all multi-indices $\gamma = (\gamma_1, \gamma_2)$.*

Proof. Since

$$\widehat{\partial^\gamma \Lambda^\beta G_\alpha}(1)(\xi) = \xi^\gamma |\xi|^\beta e^{-|\xi|^{2\alpha}}$$

is integrable, it follows that $\partial^\gamma \Lambda^\beta G_\alpha(1) \in L^\infty$ (for all $\alpha > 0$). All that remains to be proved is that $\partial^\gamma \Lambda^\beta G_\alpha(1) \in L^1$.

We consider two cases: $\alpha > \frac{1}{2}$ and $\alpha = \frac{1}{2}$. Assume first that $\alpha > \frac{1}{2}$. It is not hard to see that

$$\begin{aligned}\left|\Delta\left(\widehat{\partial^\gamma \Lambda^\beta G_\alpha}\right)(\xi,1)\right| &= \left|\Delta\left(\xi^\gamma |\xi|^\beta e^{-|\xi|^{2\alpha}}\right)\right| \\ &\leq C(1+|\xi|^N)|\xi|^{|\gamma|+\beta+2\alpha-2}e^{-|\xi|^{2\alpha}}\end{aligned}$$

for some constants $C, N \geq 0$, all $\xi \in \mathbb{R}^2$. It follows that $|\Delta(\widehat{\partial^\gamma \Lambda^\beta G_\alpha})(\xi,1)|^2$ near 0 behaves like $|\xi|^{2|\gamma|+2\beta+2\alpha-4}$, which is integrable since, because $\alpha > \frac{1}{2}$, $2|\gamma|+2\beta+4\alpha-$

$4 \geq 4\alpha - 4 > -2$. It follows that $\Delta(\widehat{\partial^\gamma \Lambda^\beta G_\alpha})(1)$ is in L^2; hence so is $|x|^2 \partial^\gamma \Lambda^\beta G_\alpha(1)$. It being clear that $G_\alpha(1) \in L^2$, it follows that $(1+|x|^2)\partial^\gamma \Lambda^\beta G_\alpha(1)$ is in L^2; since $(1+|x|^2)^{-1}$ is in L^2, the proof that $\partial^\gamma \Lambda^\beta G_\alpha(1)$ is in L^1 is complete.

Assume now that $\alpha = \frac{1}{2}$. Then

$$\begin{aligned}\Lambda^\beta G_{\frac{1}{2}}(x,1) &= \frac{1}{2\pi}\int_{\mathbb{R}^2} |\xi|^\beta e^{ix\cdot\xi} e^{-|\xi|}\, d\xi \\ &= \frac{1}{2\pi}\int_0^{2\pi}\int_0^\infty r^{\beta+1} e^{-r(1-ix\sin\theta)}\, dr\, d\theta \\ &= \frac{\Gamma(\beta+2)}{2\pi}\int_0^{2\pi} \frac{d\theta}{(1-ix\sin\theta)^{\beta+2}}.\end{aligned}$$

From the last expression it is immediate that $\partial^\gamma \Lambda^\beta G_\alpha(1) \in L^1(\mathbb{R}^2)$ if $\beta + |\gamma| > 0$. It remains to be seen whether the same is true if $\beta = 0$, $\gamma = 0$. However, in this case the integral is easily computed by residues; one has

$$G_{\frac{1}{2}}(x,1) = \frac{1}{2\pi}\int_0^{2\pi} \frac{d\theta}{(1-ix\sin\theta)^2} = \frac{1}{4(1+|x|^2)^{3/2}}.$$

The last expression is clearly integrable over $\mathbb{R}^2$. □

Remark 4.2. The last lemma is valid for $\alpha \in (0, 1/2)$, provided $\beta + |\gamma| \geq 1$. In fact, essentially the same proof as for the case $\alpha > \frac{1}{2}$ applies. The only relation α, β, γ had to satisfy for the argument to be valid was $2|\gamma| + 2\beta + 4\alpha - 4 > -2$, which clearly holds if $\alpha > 0$ and $\beta + |\gamma| \geq 1$.

The results in the remainder of this section are based on the ideas described in [5] to study the L^1 decay for solutions to viscous conservation laws.

THEOREM 4.3. *Let $\alpha \in (0,1]$, let $0 < \beta$, and assume that $I_\beta \theta_0 \in L^1(\mathbb{R}^2)$. Let $\gamma = (\gamma_1, \gamma_2)$ be a multi-index; assume $|\gamma| + \beta \geq 1$ if $\alpha < \frac{1}{2}$. Set*

$$A = \lim_{|\xi|\to 0} \frac{\widehat{\theta}_0(\xi)}{|\xi|^\beta} = \int_{\mathbb{R}^2} (I_\beta \theta_0)(x)\, dx. \tag{4.5}$$

Let $\theta(t) = e^{-t\Lambda^{2\alpha}}\theta_0 = G_a(t) * \theta_0$ be the solution of (4.1) *with initial datum θ_0. Then, for $1 \leq p \leq \infty$,*

$$\|\partial^\gamma \theta(t)|_p \leq C t^{-\frac{\beta}{2\alpha} - \frac{|\gamma|}{2\alpha} - (1-\frac{1}{p})\frac{1}{\alpha}} \|I_\beta \theta_0\|_1 \tag{4.6}$$

for all $t > 0$ and $C = C(\beta, \gamma)$ independent of t and θ_0. Moreover,

$$t^{\frac{\beta}{2\alpha} + \frac{|\gamma|}{2\alpha} + (1-\frac{1}{p})\frac{1}{\alpha}} \|\partial^\gamma \theta(t) - A\partial^\gamma \Lambda^\beta G_\alpha(t)\|_p \to 0 \tag{4.7}$$

as $t \to \infty$.

Proof. By Lemma 4.1 (see also Remark 4.2) we have $\partial^\gamma \Lambda^\beta G_\alpha \in L^p$. Writing

$$\partial^\gamma \theta(t) = (\partial^\gamma \Lambda^\beta G_\alpha(t)) * I_\beta \theta_0,$$

in view of Lemma 4.1, (4.6) is immediate from (4.4) $C = \|\partial^\gamma \Lambda^\beta G_\alpha(1)\|_p$.

For the proof of (4.7), we can write

$$\begin{aligned}&\left|\partial^\gamma \theta(x,t) - A\partial^\gamma \Lambda^\beta G_\alpha(x,t)\right| \\ &\leq \int_{\mathbb{R}^2} \left|(\partial^\gamma \Lambda^\beta G_\alpha)(x-y,t) - (\partial^\gamma \Lambda^\beta G_\alpha)(x,t)\right| |I_\beta \theta_0(y)|\, dy \\ &\leq \left(\int_{\mathbb{R}^2} |I_\beta \theta_0(y)|\, dy\right)^{\frac{p-1}{p}} \left(\int_{\mathbb{R}^2} \left|(\partial^\gamma \Lambda^\beta G_\alpha)(x-y,t) - (\partial^\gamma \Lambda^\beta G_\alpha)(x,t)\right|^p |I_\beta \theta_0(y)|\, dy\right)^{1/p}.\end{aligned}$$

Raising to the power p, integrating with respect to x, and changing the variables by $z = xt^{-\frac{1}{2\alpha}}$, combined with the self-similar form of G_α (see (4.3)), leads to the following expression:

$$\begin{aligned}
&\|\partial^\gamma\theta(t) - A\partial^\gamma\Lambda^\beta G_\alpha(t)\|_p^p \\
&\le \|I_\beta\theta_0\|_1^{p-1}\int_{\mathbb{R}^2\times\mathbb{R}^2}\left|(\partial^\gamma\Lambda^\beta G_\alpha)(x-y,t) - (\partial^\gamma\Lambda^\beta G_\alpha)(x,t)\right|^p |I_\beta\theta_0(y)|\ dxdy \\
&= t^{-\frac{p}{2\alpha}(\beta+|\gamma|+2(1-\frac{1}{p}))}\|I_\beta\theta_0\|_1^{p-1} \\
&\quad\times\int_{\mathbb{R}^2\times\mathbb{R}^2}\left|(\partial^\gamma\Lambda^\beta G_\alpha)(z-t^{-\frac{1}{2}}y,1) - (\partial^\gamma\Lambda^\beta G_\alpha)(z,1)\right|^p |I_\beta\theta_0(y)|\ dxdy.
\end{aligned}$$

To complete the proof of (4.7) we need to show only that

$$\lim_{t\to\infty}\int_{\mathbb{R}^2\times\mathbb{R}^2}\left|(\partial^\gamma\Lambda^\beta G_\alpha)(z-t^{-\frac{1}{2}}y,1) - (\partial^\gamma\Lambda^\beta G_\alpha)(z,1)\right|^p |I_\beta\theta_0(y)|\ dxdy = 0. \tag{4.8}$$

The Fourier transforms of all the derivatives of the function $\partial^\gamma\Lambda^\beta G(1)$ are in L^1; it follows that this function is infinitely many times differentiable, with all derivatives bounded. Thus the integrand in (4.8) converges uniformly to 0 over compact subsets of $\mathbb{R}^2\times\mathbb{R}^2$. Moreover, by Lemma 4.1, the function (and its derivatives) are in L^p. By this L^P-integrability, and the integrability of $I_\beta\theta_0$, one can find for each $\epsilon > 0$ a compact subset K_ϵ of $\mathbb{R}^2\times\mathbb{R}^2$ such that

$$\int_{(\mathbb{R}^2\times\mathbb{R}^2)\setminus K_\epsilon}\left|(\partial^\gamma\Lambda^\beta G_\alpha)(z-t^{-\frac{1}{2}}y,1) - (\partial^\gamma\Lambda^\beta G_\alpha)(z,1)\right|^p |I_\beta\theta_0(y)|\ dxdy < \epsilon.$$

This and the aforementioned uniform convergence prove (4.8). This completes the proof of the theorem. ∎

Remark 4.4. In the case $\alpha = 1$ we recall that in [7], [6] Miyakawa obtained the L^1 decay of $e^{t\Delta}u_0$ provided the $|x|^\beta$-momentum of the data is bounded. The assumption on the Riesz potential is weaker than the one assumed by Miyakawa; see [5].

We also obtain the following as an immediate corollary to Lemma 4.1.

COROLLARY 4.5. *Let $\alpha \in (0,1]$, let $0 < \beta$, and assume that $I_\beta\theta_0 \in L^p(\mathbb{R}^2)$, $1 \le p \le \infty$. Let $\gamma = (\gamma_1,\gamma_2)$ be a multi-index; assume $|\gamma| + \beta \ge 1$ if $\alpha < \frac{1}{2}$. Then there exists a constant $C \ge 0$ such that*

$$\|\partial^\gamma e^{-t\Lambda^{2\alpha}}\theta_0\|_p \le Ct^{-\frac{\beta}{2\alpha}-\frac{|\gamma|}{2\alpha}}\|I_\beta\theta_0\|_p \tag{4.9}$$

for all $t > 0$.

Proof. By (4.2), (4.3), and the Hausdorff–Young inequality,

$$\begin{aligned}
\|\partial^\gamma e^{-t\Lambda^{2\alpha}}\theta_0\|_p &\le t^{-\frac{\beta}{2\alpha}-\frac{|\gamma|}{2\alpha}-\frac{1}{\alpha}}\|\partial^\gamma\Lambda^\beta G_\alpha(t^{-\frac{1}{2\alpha}}\cdot,1)\|_1\|I_\beta\theta_0(t)\|_p \\
&= t^{-\frac{\beta}{2\alpha}-\frac{|\gamma|}{2\alpha}}\|\partial^\gamma\Lambda^\beta G_\alpha(1)\|_1\|I_\beta\theta_0(t)\|_p,
\end{aligned}$$

and the result follows from Lemma 4.1. ∎

4.2. Nonlinear asymptotics. The next step is to use the results from the last section to get the decay of the solutions to the geostrophic equations in L^1 and with that improve the decay of the solutions in L^p. The decay will be obtained by estimating the solutions via their integral representation. We note that the decay

below might not be optimal. So as to be able to include the critical case $\alpha = \frac{1}{2}$, we recall the following result due to Constantin, Cordoba, and Wu [1].

THEOREM 4.6. *There exists a constant c_∞ such that for any $\theta_0 \in H^2(\mathbb{R}^2)$ with $\|\theta_0\|_{H^2} \le c_\infty$, the equation*

$$\theta_t + u \cdot \nabla\theta + \Lambda\theta = 0$$

has a unique global solution θ with initial datum θ_0 satisfying

$$\|\theta(t)\|_{H^2} \le \|\theta_0\|_{H^2}$$

for all $t \ge 0$.

Combining this theorem with Theorem 2.1 and using the Gagliardo–Nirenberg inequalities, one obtains for this solution θ, $u = \mathcal{R}^\perp\theta$

$$\|\theta(t)\|_\infty \le C\|\theta\|_2^{\frac{1}{2}}\|\Lambda^2\theta\|_2^{\frac{1}{2}} \le C\|\theta_0\|_{H_2}^{\frac{1}{2}}(1+t)^{-\frac{1}{2}}, \tag{4.10}$$

$$\|u(t)\|_\infty \le C\|u\|_2^{\frac{1}{2}}\|\Lambda^2 u\|_2^{\frac{1}{2}} \le C\|\theta_0\|_{H_2}^{\frac{1}{2}}(1+t)^{-\frac{1}{2}}, \tag{4.11}$$

and by Hölder,

$$\|\nabla\theta(t)\|_2 \le \|\theta(t)\|_2^{\frac{1}{2}}\|\Lambda^2\theta(t)\|_2^{\frac{1}{2}} \le \|\theta_0\|_{H_2}^{\frac{1}{2}}(1+t)^{-\frac{1}{2}}. \tag{4.12}$$

We assume θ is this solution in case $\alpha = \frac{1}{2}$.

THEOREM 4.7. *Let $\beta > 0$, assume that $I_\beta\theta_0 \in L^1(\mathbb{R}^2)$, and let θ be the solution of the homogeneous DQG equation with initial datum θ_0.*

(i) *Assume $\frac{1}{2} \le \alpha < 1$. Then*

$$\|\theta(t)\|_1 \le Ct^{-\nu}$$

for all $t > 0$, some constant C, where

$$\nu = \begin{cases} \min(\beta, \frac{1}{2}) & \text{if} \quad \alpha = \frac{1}{2}, \\ \min(\frac{\beta}{2\alpha}, \frac{1}{2\alpha}) & \text{if} \quad \frac{1}{2} < \alpha < 1. \end{cases}$$

(ii) *Assume $\alpha = 1$. Then*

$$\|\theta(t)\|_1 \le \begin{cases} Ct^{-\frac{\beta}{2}} & \text{if} \quad \beta < 1, \\ Ct^{-\frac{1}{2}}\log(t+1) & \text{if} \quad \beta \ge 1 \end{cases}$$

for some constant C.

Proof. Write the solution by its integral representation,

$$\theta(t) = G_\alpha(t) * \theta_0 + \int_0^t G_\alpha(s) * (u \cdot \nabla\theta)(t-s)\, ds = G_\alpha(t) * \theta_0 + I(t). \tag{4.13}$$

From the last section it follows that

$$\|G_\alpha(t) * \theta_0\|_1 \le Ct^{-\frac{\beta}{2\alpha}}\|I_\beta\theta_0\|_1, \tag{4.14}$$

and the theorem reduces to proving that

$$I(t) = \int_0^t \|G_\alpha(s) * (u \cdot \nabla\theta)(t-s)\|_1\, ds \le C(1+t)^{-\nu}$$

is appropriately bounded. By Hausdorff–Young, Hölder, and the fact that $u\nabla\theta = \operatorname{div}(u\theta)$,

$$I(t) = \int_0^{t/2} \|G_\alpha(s) * (u\cdot\nabla\theta)(t-s)\|_1\,ds + \int_{t/2}^t \|\nabla G_\alpha(s) * (u\theta)(t-s)\|_1\,ds$$
$$\le \int_0^{t/2} \|G_\alpha(s)\|_1\|u(t-s)\|_2\|\nabla\theta(t-s)\|_2\,ds + \int_{t/2}^t \|\nabla G_\alpha(s)\|_1\|u(t-s)\|_2\|\theta(t-s)\|_2\,ds$$
$$= J(t) + K(t).$$

By the results of the last section we have

$$\|G_\alpha(s)\|_1 = \|G_\alpha(1)\|_1 = C, \tag{4.15}$$

$$\|\nabla G_\alpha(s)\|_1 = s^{-\frac{1}{2\alpha}}\|\nabla G_\alpha(1)\|_1 = Cs^{-\frac{1}{2\alpha}}, \tag{4.16}$$

with C a constant depending only on α. We also have

$$\|u(t-s)\|_2\|\nabla\theta(t-s)\|_2 \le \begin{cases} (1+t)^{-3/2} & \text{if} \quad \alpha = \frac{1}{2}, \\ (1+t)^{-\frac{3}{2\alpha}} & \text{if} \quad \frac{1}{2} < \alpha \le 1. \end{cases}$$

The estimate for $\alpha = \frac{1}{2}$ comes from Theorem 2.1 and (4.12), the one for $\alpha > \frac{1}{2}$ from Theorem 3.2. Using this and (4.15) we get, with $\mu = \frac{3}{2}$ if $\alpha = \frac{1}{2}$, $\mu = \frac{3}{2\alpha}$, otherwise

$$J(t) \le C\int_0^{t/2} (1+t-s)^{-\mu}\,ds \le C(1+t)^{1-\mu},$$

since in all cases $\mu > 1$. Note that $\mu - 1 = \frac{1}{2}$ when $\mu = \frac{3}{2}$, and for all other μ's it follows that $\mu - 1 \ge \frac{1}{2\alpha}$, so $\mu - 1 \ge \nu$ in all cases. Thus

$$J(t) \le C(1+t)^{-\nu}, \tag{4.17}$$

and we are done with the estimate for J. To estimate $K(t)$ we use (4.16) and the fact that $\|u(t-s)\|_2\|\theta(t-s)\|_2 \le C(1+t-s)^{-\frac{1}{\alpha}}$ to get

$$K(t) \le C\int_{t/2}^t s^{-\frac{1}{2\alpha}}(1+t-s)^{-\frac{1}{\alpha}}\,ds \le \begin{cases} Ct^{-\frac{1}{2\alpha}} & \text{if} \quad \frac{1}{2} \le \alpha < 1, \\ Ct^{-\frac{1}{2}}\log t & \text{if} \quad \alpha = 1. \end{cases} \tag{4.18}$$

The conclusion of the theorem follows now from (4.13) and (4.14), using (4.17) and (4.18) to bound $I(t) = J(t) + K(t)$. □

Derivatives of the solution θ can be similarly bounded, at least if $\alpha > \frac{1}{2}$. We have the following theorem.

THEOREM 4.8. *Let $\beta > 0$. Assume $I_\beta\theta_0 \in L^1(\mathbb{R}^2)$, $\theta_0 \in H^m$ for some $m \ge 1$, and let γ be a multi-index, $|\gamma| \le m-1$. Then*

$$\|\partial^\gamma\theta(t)\|_1 \le \begin{cases} Ct^{-\min(\frac{\beta+|\gamma|}{2\alpha}, \frac{|\gamma|+1}{2\alpha})}, & \frac{1}{2} < \alpha < 1, \\ Ct^{-\min(\frac{\beta+|\gamma|}{2\alpha}, \frac{|\gamma|+1}{2\alpha})}\log(t+1), & \alpha = 1. \end{cases}$$

Proof. Proceeding as in the proof of Theorem 4.7, we get

$$\|\partial^\gamma\theta(t)\|_1 \le \|\partial^\gamma G_\alpha(t) * \theta_0\|_1 + J_\gamma(t) + K_\gamma(t),$$

where the terms on the right-hand side now have the following interpretations and bounds:

$$\|\partial^\gamma G_\alpha(t) * \theta_0\|_1 \le Ct^{-\frac{\beta+|\gamma|}{2\alpha}}$$

by (4.6). For the second term, using the estimates in Theorem 3.2 we obtain first

$$\begin{aligned}\|\partial^\gamma(u\cdot\nabla\theta)(t-s)\|_1 &\le \sum_{|\gamma_1|+|\gamma_2|=|\gamma|+1} c_{\gamma_1,\gamma_2}\|\partial^{\gamma_1}u(t-s)\|_2\|\partial^{\gamma_2}\theta(t-s)\|_2\\ &\le C(1+t-s)^{-\frac{|\gamma|+3}{2\alpha}}\end{aligned} \tag{4.19}$$

(the coefficients c_{γ_1,γ_2} coming from Leibniz's formula); hence (since $3-2\alpha\ge 1$)

$$J_\gamma(t)=\int_0^{t/2}\|G_\alpha(s)\|_1\|\partial^\gamma(u\cdot\nabla\theta)(t-s)\|_1\,ds \le C(t+1)^{-\frac{|\gamma|+3-2\alpha}{2\alpha}} \le C(t+1)^{-\frac{|\gamma|+1}{2\alpha}}.$$

For the third term we use the fact that $\|\nabla\partial^\gamma G_\alpha(t)\| = Ct^{-\frac{|\gamma|+1}{2\alpha}}$ and, as in Theorem 4.7, that

$$\|u(t-s)\theta(t-s)\|_1 \le (1+t-s)^{-\frac{1}{\alpha}}$$

to get

$$K_\gamma(t)=\int_{t/2}^t\|\nabla\partial^\gamma G_\alpha(s)\|_1\|(u\theta)(t-s)\|_1\,ds \le \begin{cases} Ct^{-\frac{|\gamma|+1}{2\alpha}} & \text{if} \quad \frac12<\alpha<1,\\ Ct^{-\frac{|\gamma|+1}{2}}\log(t+1) & \text{if} \quad \alpha=1.\end{cases}$$

The theorem follows. □

Finally, we see that the solution θ is asymptotically equivalent to the self-similar solution of the linear equation, at least if $\beta<1$. For a given $\beta>0$, (4.3) shows that the self-similar solution $\partial^\gamma\Lambda^\beta G_\alpha$ of the linear equation decays in the L^1 norm at the rate of $t^{-\frac{\beta+|\gamma|}{2\alpha}}$ as $t\to\infty$. Theorem 4.8 shows that the derivative ∂^γ of the solution of the nonlinear equation (with datum θ_0 satisfying $I_\beta\theta_0\in L^1$) decays (at least) at the same rate if $\beta<1$. By asymptotic equivalence, we mean that the difference of θ and the self-similar solution of the linear equation decays at a better rate.

THEOREM 4.9. *Assume $0<\beta<1$ and that the hypotheses of Theorem 4.8 hold. Then, with*

$$A=A_\beta=\int_{\mathbb{R}^2}(I_\beta\theta_0)(x)\,dx,$$

one has

$$\lim_{t\to\infty} t^{\frac{\beta+|\gamma|}{2\alpha}}\left\|\partial^\gamma\theta(t)-A\partial^\gamma\Lambda^\beta G_\alpha(t)\right\|_1=0.$$

Proof. We have

$$\partial^\gamma\theta(t)-A\partial^\gamma\Lambda^\beta G_\alpha(t)=\partial^\gamma(G_a(t)*\theta_0)-A\partial^\gamma\Lambda^\beta G_\alpha(t)+\int_0^t\partial^\gamma G_\alpha(s)*(u\cdot\nabla\theta)(t-s)\,ds,$$

and by Theorem 4.3 it suffices to prove that

$$\lim_{t\to\infty} t^{\frac{\beta+|\gamma|}{2\alpha}}H(t)=0, \tag{4.20}$$

where

$$H(t) = \int_0^t \|\partial^\gamma G_\alpha(s) * (u \cdot \nabla\theta)(t-s)\|_1 \, ds.$$

This is, however, immediate from the proof of Theorem 4.8. In fact, we have $H \le J_\gamma + K_\gamma$, where J_γ, K_γ are as in the proof of Theorem 4.8. It follows that $H(t)$ decays at the rate of either $t^{-\frac{|\gamma|+1}{2\alpha}}$ $(\alpha < 1)$ or $t^{-\frac{|\gamma|+1}{2\alpha}} \log t$ $(\alpha = 1)$; in either case (4.20) holds because $\beta < 1$. □

COROLLARY 4.10. *Under the conditions of Theorem* 4.9 *if* $|\gamma| = 0$*, the conclusion of the theorem is also valid for the case* $\alpha = \frac{1}{2}$.

Proof. The proof follows the same lines as that of Theorem 4.9. □

Acknowledgments. The authors would like to express their thanks to the anonymous referees for many very helpful and thoughtful comments and suggestions.

REFERENCES

[1] P. CONSTANTIN, D. CORDOBA, AND J. WU, *On the critical dissipative quasi-geostrophic equation*, Indiana Univ. Math. J., 50 (2001), pp. 97–107.
[2] L. CAFFARELLI, R. KOHN, AND L. NIRENBERG, *Partial regularity of suitable weak solutions of the Navier-Stokes equations*, Comm. Pure Appl. Math., 35 (1982), pp. 771–831.
[3] P. CONSTANTIN, A. MAJDA, AND E. TABAK, *Formation of strong fronts in the* $2D$*-quasi-geostrophic thermal active scalar*, Nonlinearity, 7 (1994), pp. 1495–1533.
[4] P. CONSTANTIN AND J. WU, *Behavior of solutions of* $2D$ *quasi-geostrophic equations*, SIAM J. Math. Anal., 30 (1999), pp. 937–948.
[5] G. KARCH AND M. E. SCHONBEK, *On zero mass solutions to viscous conservation laws*, Comm. Partial Differential Equations, 27 (2002), pp. 2071–2100.
[6] T. MIYAKAWA, *Application of Hardy space techniques to the time-decay problem for incompressible Navier-Stokes flows in* $\mathbb{R}^2$, Funkcial. Ekvac., 41 (1998), pp. 383–434.
[7] T. MIYAKAWA, *Hardy spaces of solenoidal vector fields, with applications to the Navier-Stokes equations*, Kyushu J. Math., 50 (1996), pp. 1–64.
[8] J. PEDLOSKY, *Geophysical Fluid Dynamics*, Springer-Verlag, New York, 1987.
[9] M. PINSKY, *Introduction to Fourier Analysis and Wavelets*, Brooks/Cole, Pacific Grove, CA, 2002.
[10] S. RESNICK, *Dynamical Problems in Nonlinear Advective Partial Differential Equations*, Ph.D. Thesis, University of Chicago, 1995.
[11] M. E. SCHONBEK, *Decay of solutions to parabolic conservation laws*, Comm. Partial Differential Equations, 5 (1980), pp. 449–473.
[12] M. E. SCHONBEK, L^2 *decay of weak solutions of the Navier-Stokes equations*, Arch. Ration. Mech. Anal., 88 (1985), pp. 209–222.
[13] E. M. STEIN, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
[14] M. TAYLOR, *Pseudo Differential Operators and Nonlinear PDE*, Birkhäuser Boston, Cambridge, MA, 1991.
[15] J. WU, *Private communication.*
[16] J. WU, *Dissipative quasi-geostrophic equations with* L^p *data*, Electron. J. Differential Equations, 2001 (2001), pp. 1–13.

SIAM J. MATH. ANAL.
Vol. 35, No. 2, pp. 376–388

ON THE EXISTENCE OF A WEAK SOLUTION TO A TWO-DIMENSIONAL FREE-BOUNDARY PROBLEM WITH A NONLINEAR FLUX CONDITION*

E. CASELLA†

Abstract. The main result of this paper is a global existence theorem for a two-dimensional free-boundary problem with a nonlinear boundary condition in suitable Sobolev spaces. The existence result is proved by using some a priori estimates and the Schauder fixed point theorem.

Key words. Stefan problem, free-boundary problem, nonlinear flux

AMS subject classifications. 80A22, 35R35

DOI. 10.1137/S0036141002410411

1. Introduction. In the present paper we consider a free-boundary problem concerning the evolution of a heated viscous incompressible fluid with a nonlinear boundary condition on the temperature. Problems of this kind have been studied by several authors; see, for instance, Fasano and Primicerio [11], Cannon and DiBenedetto [4, 5], Damlamian and Kenmochi [9], and Visintin [16].

We consider a coupling between the Navier–Stokes equations and the Stefan equation. Some papers have been devoted to the coupling of these equations; see Cannon and DiBenedetto [5], Cannon, DiBenedetto, and Knightly [6], and Wang [18]. One of the main novelties of this paper is the analysis of a simplified approximate method to couple the two equations. We propose an approximate method (see section 2) that does not impose a vanishing velocity in the solid part. This particular model, analyzed in Casella and Giangi [7], is a simplification of the exact one introduced by Cannon, DiBenedetto, and Knightly [6] and by Wang [18]. Its interest relies on the fact that the existence of weak solutions can be proved in a more straightforward way. Furthermore, this model (coupled with finite differences or finite elements) provides some numerical methods which are very simple to implement and which give results in good agreement with experimental data. In particular, Rady and Mohanty [15] have applied it to the melting of gallium and soldification of tin in a square cavity; Brent, Voller, and Reid [2] and Giangi and Stella [12] to a melting problem; and Giangi, Stella, and Kowalewski [13] to a freezing water problem.

In the present paper we apply this model to a classical problem with nonlinear flux. We consider a fluid that can take a phase transition in a bounded open set $\Omega \subset \mathbb{R}^2$. We suppose that the boundary $\Gamma = \partial\Omega$ is a smooth one-dimensional manifold and that Ω is locally situated on one side of Γ. We consider a homogeneous Dirichlet condition for the velocity field. Concerning the temperature, we impose a Dirichlet condition on part of the boundary and a suitable nonlinear flux condition on the rest. In particular, the flux may depend in a nonlinear way on the temperature. In this case, it is possible to study the physical problem of irradiation with the so-called Stefan–Boltzmann law. This may be used to describe, for example, the electron flux

*Received by the editors June 26, 2002; accepted for publication February 21, 2003; published electronically July 18, 2003.
http://www.siam.org/journals/sima/35-2/41041.html
†Dip. di Matematica, Università di Brescia, Facoltà di Ingegneria, Via Valotti 9, 25133 Brescia, Italy (casella@ing.unibs.it).

in a rarefied gas. The model equation, which we have in mind, is

$$\frac{\partial \theta}{\partial n} = \eta(\theta^4 - \theta_0^4),$$

where θ and θ_0 are, respectively, the temperature of the boundary and that of the source, while η is a physical parameter. The main result of this paper is a global existence theorem in suitable Sobolev spaces for the Stefan problem with convection and nonlinear flux. In particular, existence is proved by considering a sequence of approximating problems, for which a priori estimates are obtained. The limit then provides a weak solution for the original problem.

The paper is organized as follows. In section 2 we describe the mathematical model. In section 3 we define some notation. In section 4 we give the notion of weak solution, and we state and prove the main result.

2. Formulation of the problem. Suppose that in a region Ω of $\mathbb{R}^2$ a Newtonian and incompressible fluid undergoes a change of phase at a fixed temperature. In general, there are in the fluid convective motions, originated by a body force $\mathbf{f}$ which depends on the temperature θ of the fluid.

In what follows, Q denotes the space-time cylinder $\Omega \times (0,T)$, where $0 < T \in \mathbb{R}$ and Ω is a bounded open set in $\mathbb{R}^2$ of Lipschitz class, filled by the fluid.

We define $\Omega(0) := \Omega \times \{0\}$ and assume that Γ_2 is an open subset of Γ of positive one-dimensional Hausdorff measure, and we set $\Sigma := \Gamma_2 \times (0,T)$ and $\Gamma_1 := \Gamma \setminus \Gamma_2$.

The phase-transition temperature is denoted by 0, and the following temperature-phase rule is assumed:

$$\theta \geq 0 \quad \text{in the liquid region} \quad \text{and} \quad \theta \leq 0 \quad \text{in the solid region.}$$

Let M_1 be a real positive constant and let α be the maximal monotone graph defined by

$$\alpha(\theta) = \begin{cases} \{\theta + M_1\} & \text{if } \theta > 0, \\ [-M_1, M_1] & \text{if } \theta = 0, \\ \{\theta - M_1\} & \text{if } \theta < 0. \end{cases} \tag{2.1}$$

The following constitutive relation between the enthalpy w and the temperature θ of the fluid holds:

$$w \in \alpha(\theta).$$

The term β is a function which depends on the temperature:

$$\beta(\theta) = \begin{cases} C\,\theta & \text{if } \theta < 0, \\ 0 & \text{if } \theta \geq 0, \end{cases} \tag{2.2}$$

where C is a *large* real positive constant.

By assuming the Boussinesq approximation, the system of partial differential equations we are going to study is

$$\frac{\partial w}{\partial t} - k\Delta\theta + \mathbf{v} \cdot \nabla\theta = 0 \quad \text{in } Q, \tag{2.3}$$

$$\frac{\partial \mathbf{v}}{\partial t} - \nu\Delta\mathbf{v} + (\mathbf{v} \cdot \nabla)\mathbf{v} + \nabla p + \beta(\theta)\mathbf{v} = \mathbf{f}(\theta) \quad \text{in } Q, \tag{2.4}$$

$$\nabla \cdot \mathbf{v} = 0 \quad \text{in } Q, \tag{2.5}$$

$$w \in \alpha(\theta) \quad \text{a.e. in } Q,$$

where k and ν are given real positive constants and p is the pressure of the fluid. The previous system is supplemented by the following initial and boundary conditions:

$$w(\mathbf{x},0) = w_0(\mathbf{x}) \quad \text{in } \Omega(0), \tag{2.6}$$

$$\mathbf{v}(\mathbf{x},0) = \mathbf{v}_0(x) \quad \text{in } \Omega(0), \tag{2.7}$$

$$\theta = 0 \quad \text{on } \Gamma_1 \times (0,T), \tag{2.8}$$

$$\frac{\partial \theta}{\partial \mathbf{n}} = -g(\theta) \quad \text{on } \Gamma_2 \times (0,T),$$

$$\mathbf{v} = \mathbf{0} \quad \text{on } \Gamma \times (0,T). \tag{2.9}$$

We observe that in the solid region the term $\beta(\theta)$ has the effect of a constraint: when the temperature θ decreases, $\beta(\theta)$ increases and the solution $\mathbf{v}$ of the velocity equation becomes small. On the other hand, the definition of β implies that, where θ is larger than 0, the velocity equation is equivalent to the Navier–Stokes system.

3. Notation and function spaces. In this section we provide a brief description of the function spaces that shall be used in what follows. We shall need the following Banach spaces:

$$\mathcal{W} := \{u \in H^1(\Omega) : u = 0 \text{ on } \Gamma_1\},$$

$$H := \{\mathbf{v} \in L^2(\Omega) : \mathbf{v} \cdot \mathbf{n} = 0 \ \text{ on } \Gamma \quad \text{and} \quad \operatorname{div} \mathbf{v} = 0\},$$

$$V := \{\mathbf{v} \in H^1(\Omega) : \mathbf{v} = \mathbf{0} \ \text{ on } \Gamma \quad \text{and} \quad \operatorname{div} \mathbf{v} = 0\}.$$

By recalling the generalized Poincaré inequality, the spaces $\mathcal{W}$ and V are equipped with the norms

$$\|u\|_{\mathcal{W}} := \|\nabla u\|_{L^2(\Omega)} \quad \text{and} \quad \|\mathbf{v}\|_V := \|\nabla \mathbf{v}\|_{L^2(\Omega)}.$$

Let us consider the trilinear and continuous form $b : V \times V \times V \to R$ defined as

$$b(\mathbf{u},\mathbf{v},\mathbf{w}) := \sum_{i,j=1}^{2} \int_\Omega u_i \, (D_i v_j) \, w_j \, dx,$$

and let $B : V \times V \to V'$ be the continuous operator given by

$$\langle B(\mathbf{u},\mathbf{v}), \mathbf{w} \rangle_{V',V} = b(\mathbf{u},\mathbf{v},\mathbf{w}) \quad \forall \mathbf{w} \in V.$$

Finally, we denote by γ_0 the trace operator

$$\gamma_0 : L^2(0,T;\mathcal{W}) \to L^2(0,T;H^{\frac{1}{2}}(\Gamma_2))$$

and by G the operator defined by

$$G : \{z \in L^2(0,T;\mathcal{W}) : \ g(x,t,\gamma_0 z(x,t)) \in L^2(\Sigma)\} \longrightarrow L^2(0,T;\mathcal{W}')$$

$$\langle G(z), s \rangle_{L^2(0,T;\mathcal{W}'),L^2(0,T;\mathcal{W})} = \int_0^T \int_{\Gamma_2} g(\gamma_0 z) \, \gamma_0 s \, d\sigma \, dt.$$

The interpolation theory provides the following compactness result (the Lions–Aubin theorem), which will be fundamental in proving the main proposition of this paper.

THEOREM 3.1. *For $i = 0, 1$ let $1 < p_i < +\infty$. Let B, B_0, B_1 be Banach spaces, and let B_0, B_1 be reflexive and such that $B_0 \subset B \subset B_1$. Let the inclusion $B_0 \hookrightarrow B$ be compact and the inclusion $B \hookrightarrow B_1$ be continuous.*

Then the embedding $L^{p_0}(0,T;B_0) \cap W^{1,p_1}(0,T;B_1) \hookrightarrow L^{p_0}(0,T;B)$ is compact.

In what follows, as usual, we shall denote by $c_i, i \in \mathbb{N}$, some real positive constants.

4. An existence result. The aim of this section is to prove the existence of a weak solution to the problem described by (2.3)–(2.9).

DEFINITION 4.1. *Let*

$$w_0 \in L^2(\Omega),\ \mathbf{v}_0 \in H,$$

$$\mathbf{f} : R \to \mathbb{R}^2 \text{ a Lipschitz function} : \mathbf{f}(0) = \mathbf{0}.$$

Let g be a function defined on $\Sigma \times \mathbb{R}$, which satisfies some further properties that shall be introduced later. A weak solution *of* (2.3)–(2.9) *is a pair $(w, \mathbf{v})$ such that*

$$w \in L^2(Q) : \theta = \alpha^{-1}(w) \in L^2(0,T;\mathcal{W}), \quad g(x,t,\gamma_0[\theta(x,t)]) \in L^2(\Sigma),$$

$$\mathbf{v} \in L^2(0,T;V) \cap L^\infty(0,T;H),$$

which satisfies

$$\begin{aligned}
&\int_Q \left(-w\frac{\partial \varphi}{\partial t} + k\,\nabla\theta\cdot\nabla\varphi + (\mathbf{v}\cdot\nabla\theta)\varphi \right) dx\,dt \\
&= -\int_0^T \int_{\Gamma_2} g(\gamma_0\theta)\varphi\,d\sigma\,dt + \int_\Omega w^0\varphi(\cdot\,,0)\,dx \\
&\forall\,\varphi \in C^\infty(\overline{Q}) \quad \textit{such that} \quad \varphi = 0 \quad \textit{on} \quad \Gamma_1 \quad \textit{and} \quad \varphi(x,T) = 0 \quad \forall x \in \Omega;
\end{aligned}$$

$$\begin{aligned}
&\int_Q \left(-\mathbf{v}\cdot\frac{\partial \boldsymbol{\psi}}{\partial t} + \nu\,\nabla\mathbf{v} : \nabla\boldsymbol{\psi} + \{(\mathbf{v}\cdot\nabla)\mathbf{v}\}\cdot\boldsymbol{\psi} + \beta(\theta)\mathbf{v}\cdot\boldsymbol{\psi} \right) dx\,dt \\
&= \int_Q \mathbf{f}(\theta)\cdot\boldsymbol{\psi}\,dx\,dt + \int_\Omega \mathbf{v}^0\cdot\boldsymbol{\psi}(\cdot\,,0)\,dx \\
&\forall\,\boldsymbol{\psi} \in C^\infty(\overline{Q}) \quad \textit{such that} \quad \boldsymbol{\psi} = \mathbf{0}\ \textit{on}\ \Gamma,\ \operatorname{div}\boldsymbol{\psi} = 0\ \textit{and}\ \boldsymbol{\psi}(x,T) = 0 \quad \forall x \in \Omega.
\end{aligned}$$

THEOREM 4.2. *Assume that $-M_1 \le w_0 \le M_1$ and that for each $(x,t) \in \Sigma$ the real function*

$$\xi \longrightarrow g(x,t,\xi)$$

is nondecreasing monotone and such that $g(x,t,0) = 0$. Furthermore, let $g(x,t,\xi)$ belong to $L^2(\Sigma)$ for each $\xi \in \mathbb{R}$. Then there exists a weak solution for system (2.3)–(2.9) *such that θ and $w \in L^\infty(Q)$. In particular,*

$$-M_1 \le w \le M_1 \quad \textit{a.e. in } Q.$$

Proof. We prove the previous result via approximation, a priori estimates, and a limit procedure.

4.1. A collection of approximate problems. As a first step we approximate the maximal monotone graph α defined in (2.1).

Let $\{\alpha_m\}_{m\in\mathbb{N}}$ be a sequence of odd smooth functions such that

1. for suitable ζ_0, ζ_1, and c, real positive constants independent of m,

$$\zeta_0 \le \alpha_m'(u) \le \zeta_1 \text{ for } |u| > \frac{2}{m} \quad \text{and} \quad \zeta_0 \le \alpha_m'(u) \le cm \text{ for } |u| \le \frac{2}{m}; \tag{4.1}$$

2. $\{\delta_m\}_{m\in\mathbb{N}} := \{\alpha_m^{-1}\}_{m\in\mathbb{N}}$ is a sequence of odd functions, strongly converges to $\delta(u) := \alpha^{-1}(u)$ uniformly on compact sets, and is such that

$$\delta_m(M_1) > 0, \quad 0 < \frac{1}{cm} \le \delta_m' \le \zeta_0^{-1} \quad \forall m \in \mathbb{N}. \tag{4.2}$$

We refer to [4] for more details on the construction and regularity of the sequences $\{\alpha_m\}_{m\in\mathbb{N}}$ and $\{\delta_m\}_{m\in\mathbb{N}}$. Now let g_m be a sequence in $C^1(\overline{\Sigma}\times\mathbb{R})$ such that

1. for each $(x,t)\in\Sigma$ and $m\in\mathbb{N}$ the real map $\xi \mapsto g_m(x,t,\xi)$ is nondecreasing monotone;
2. the sequence $g_m(x,t,\xi)$ converges to $g(x,t,\xi)$ pointwise on $\Sigma\ \forall\,\xi\in\mathbb{R}$;
3. there exists a function $F\in L^2(\Sigma)$ such that

$$\forall \xi\in\mathbb{R} \quad |g_m(x,t,\xi) - g(x,t,\xi)| \le F(x,t) \quad \text{on } \Sigma;$$

4. the sequence g_m satisfies the inequalities

$$g_m(x,t,\delta_m(-M_1)) \le 0 \quad \text{and} \quad g_m(x,t,\delta_m(M_1)) \ge 0 \quad \text{on } \Sigma.$$

Also let $-M_1 \le w_{0,m} \le M_1$ be a sequence in $L^2(\Omega)$ and let $\mathbf{v}_{0,m}$ be a sequence in V such that

$$w_{0,m} \to w_0 \text{ strongly in } L^2(\Omega) \quad \text{and} \quad \mathbf{v}_{0,m} \to \mathbf{v}_0 \text{ strongly in } H.$$

We now consider the following auxiliary problem. Find $(w_m, \mathbf{v}_m)$ in $L^2(Q)\times L^2(0,T;V) \cap L^\infty(0,T;H) : \theta_m := \delta_m(w_m) \in L^2(0,T;\mathcal{W})$ and

$$\begin{aligned}
&\int_Q \left(-w_m\frac{\partial\varphi}{\partial t} + (\mathbf{v}_m\cdot\nabla\theta_m)\varphi + k\nabla\theta_m\cdot\nabla\varphi\right) dx\,dt \\
&= -\int_0^T\int_{\Gamma_2} g_m(\theta_m)\,\varphi\,d\sigma\,dt + \int_\Omega w_{0,m}\,\varphi(x,0)\,dx \\
&\forall\,\varphi\in C^\infty(\overline{Q}), \quad \varphi = 0 \quad \text{on } \Gamma_1, \quad \varphi(x,T) = 0 \ \ \forall x\in\Omega;
\end{aligned}$$

$$\begin{aligned}
&\int_Q -\mathbf{v}_m\cdot\frac{\partial\boldsymbol{\psi}}{\partial t} + \{(\mathbf{v}_m\cdot\nabla)\mathbf{v}_m\}\cdot\boldsymbol{\psi} + \nu\nabla\mathbf{v}_m : \nabla\boldsymbol{\psi} + \beta(\theta_m)\mathbf{v}_m\cdot\boldsymbol{\psi}\,dx\,dt \\
&= \int_Q \mathbf{f}(\theta_m)\cdot\boldsymbol{\psi}\,dx\,dt + \int_\Omega \mathbf{v}_{0,m}\cdot\boldsymbol{\psi}(x,0)\,dx \\
&\forall\,\boldsymbol{\psi}\in C^\infty(\overline{Q}), \quad \boldsymbol{\psi} = 0 \quad \text{on } \Gamma, \quad \operatorname{div}\boldsymbol{\psi} = 0, \quad \boldsymbol{\psi}(x,T) = 0 \ \ \forall x\in\Omega.
\end{aligned}$$

We employ a Galerkin procedure to solve the previous problem. We consider in $L^2(\Omega)$ the basis $\{z_i\}$ of complete orthonormal polynomials. We also introduce the orthonormal basis $\{\mathbf{w}_i\}$ of H generated by the Stokes problems

$$\begin{cases} -\Delta\mathbf{w}_i + \nabla p = \lambda_i\mathbf{w}_i \text{ in } \Omega, \\ \operatorname{div}\ \mathbf{w}_i = 0 \text{ in } \Omega, \\ \mathbf{w}_i = 0 \text{ on } \Gamma, \end{cases}$$

where p is a scalar function representing a pressure.

The initial data $w_{0,m}$ and $\mathbf{v}_{0,m}$ are represented as

$$w_{0,m}(x) = \sum_{i=1}^{\infty} c^0_{m,i}\, z_i(x), \quad \mathbf{v}_{0,m}(x) = \sum_{i=1}^{\infty} d^0_{m,i}\, \mathbf{w}_i(x), \quad \text{where } c^0_{m,i}, d^0_{m,i} \in \mathbb{R}.$$

For any fixed $l \in \mathbb{N}$, we denote by P_l the $L^2(\Omega)$-projection over the span generated by $z_1, \dots, z_l$, and we set

$$\mathbf{v}^*_l(x,t) = \sum_{i=1}^{l} d^{l,*}_i(t)\mathbf{w}_i(x), \quad d^{l,*}_i(t) \in C^1[0,T] \qquad \text{for } i = 1, \dots, l.$$

We wish to find $w_{m,l}(x,t) = \sum_{i=1}^{l} c^l_{m,i}(t) z_i(x)$, with the real functions $c^l_{m,i} \in C^1[0,T]$ for $i = 1, \dots, l$, such that $\theta_{m,l} := \delta_m(w_{m,l})$ and the Stefan problem

$$\text{(4.3)} \qquad \begin{cases} \dfrac{\partial w_{m,l}}{\partial t} - \nu\Delta\theta_{m,l} + \mathbf{v}^*_l \cdot \nabla\theta_{m,l} = 0 \quad \text{in } Q, \\ \theta_{m,l} = 0 \quad \text{on } \Gamma_1 \times (0,T), \\ \dfrac{\partial \theta_{m,l}}{\partial \mathbf{n}} = -g_m(\theta_{m,l}) \quad \text{on } \Gamma_2 \times (0,T), \\ w_{m,l}(x,0) = P_l(w_{0,m}) \quad \text{in } \Omega \end{cases}$$

with prescribed convection $\mathbf{v}^*_l$ is satisfied in the sense of the projection over the span generated by $z_1, \dots, z_l$. The previous equations form a linear differential system for the real functions $c^l_{m,i}(t)$. With a standard argument, one can easily prove that this system has a maximal solution on the time interval $(0,T)$.

We now denote by π_l the H-projection onto the linear space of $\mathbf{w}_1, \dots, \mathbf{w}_l$, and we employ the function $w_{m,l}$ thus obtained to construct $\mathbf{v}_{m,l} = \sum_{i=1}^{l} d^l_{m,i}(t)\mathbf{w}_i$, solution, in the sense of the projection over the span generated by $\mathbf{w}_1, \dots, \mathbf{w}_l$, of the system

$$\text{(4.4)} \qquad \begin{cases} \dfrac{\partial \mathbf{v}_{m,l}}{\partial t} - \nu\Delta\mathbf{v}_{m,l} + (\mathbf{v}_{m,l} \cdot \nabla)\mathbf{v}_{m,l} + \beta(\delta_m(w_{m,l}))\mathbf{v}_{m,l} = \mathbf{f}(\delta_m(w_{m,l})) \text{ in } Q, \\ \text{div } \mathbf{v}_{m,l} = 0 \quad \text{in } Q, \\ \mathbf{v}_{m,l} = \mathbf{0} \quad \text{on } \Gamma, \\ \mathbf{v}_{m,l}(x,0) = \pi_l(\mathbf{v}_{0,m}) \quad \text{in } \Omega. \end{cases}$$

It can be shown easily that there exists a unique solution of the previous nonlinear differential system for the real functions $d^l_{m,i}(t)$ on the interval $(0,T)$. For $l \in \mathbb{N}$, let B be a ball in $L^\infty(0,T)^l$ with a large enough radius and let $F_l : B \to F_l(B)$ be the function defined as

$$F_l(d^*_1(t), \dots, d^*_l(t)) = (d^l_{m,1}(t), \dots, d^l_{m,l}(t)).$$

By using the Schauder fixed point theorem, we prove that, for each l belonging to $\mathbb{N}$, the map F_l has a fixed point. To check that F_l satisfies all the hypotheses of the Schauder fixed point theorem, some a priori estimates are required.

4.2. Some a priori estimates. We now set $z_{m,l} := (w_{m,l}-M_1)^+-(w_{m,l}+M_1)^-$ and multiply the first equation in (4.3) by $z_{m,l}$. We get that

$$\int_\Omega \frac{\partial w_{m,l}}{\partial t} z_{m,l}\,dx + \int_\Omega \nabla\theta_{m,l}\cdot\nabla z_{m,l}\,dx + \int_\Omega (\mathbf{v}_l^*\cdot\nabla\theta_{m,l})\ z_{m,l}\,dx + \int_{\Gamma_2} g_m(\theta_{m,l})\ z_{m,l}\,d\sigma = 0. \tag{4.5}$$

We claim that

$$\int_\Omega \nabla\theta_{m,l}\cdot\nabla z_{m,l}\,dx \geq 0, \tag{4.6}$$

$$\int_\Omega (\mathbf{v}_l^*\cdot\nabla\theta_{m,l})\ z_{m,l}\,dx = 0, \tag{4.7}$$

$$\int_{\Gamma_2} g_m(\theta_{m,l})\ z_{m,l}\,dx \geq 0. \tag{4.8}$$

The first statement is trivial. Concerning (4.7), by the definition of $z_{m,l}$ and by integrating by parts we get

$$\begin{aligned}
&\int_{\{w_{m,l}<-M_1\}\cup\{w_{m,l}>M_1\}} (\mathbf{v}_l^*\cdot\nabla\theta_{m,l}) z_{m,l}\,dx \\
&\qquad = -\int_{\{w_{m,l}<-M_1\}\cup\{w_{m,l}>M_1\}} (\mathbf{v}_l^*\cdot\nabla z_{m,l})\theta_{m,l}\,dx \\
&\qquad = -\int_{\{w_{m,l}<-M_1\}\cup\{w_{m,l}>M_1\}} (\mathbf{v}_l^*\cdot\nabla w_{m,l})\delta_m(w_{m,l})\,dx \\
&\qquad = -\int_{\{w_{m,l}<-M_1\}\cup\{w_{m,l}>M_1\}} \mathbf{v}_l^*\cdot\nabla\left(\int_{-M_1}^{w_{m,l}}\delta_m(s)ds\right)dx.
\end{aligned}$$

Observing that $\int_{-M_1}^{M_1}\delta_m(s)ds = 0$ and integrating by parts again, we obtain (4.7). We now prove (4.8). Clearly,

$$\int_{\Gamma_2} g_m(\theta_{m,l})\ z_{m,l}\,dx\,dt = \int_{\{w_{m,l}>M_1\}} g_m(\theta_{m,l})\ z_{m,l}\,dx + \int_{\{w_{m,l}<-M_1\}} g_m(\theta_{m,l})\ z_{m,l}\,dx.$$

Since for $w_{m,l} > M_1$ we have $g_m(\theta_{m,l}) \geq g_m(\delta_m(M_1)) \geq 0$ and for $w_{m,l} < -M_1$ we have $g_m(\theta_{m,l}) \leq g_m(\delta_m(-M_1)) \leq 0$, it follows that

$$\int_{\{w_{m,l}>M_1\}} g_m(\theta_{m,l})\ z_{m,l}\,dx \geq 0 \quad\text{and}\quad \int_{\{w_{m,l}<-M_1\}} g_m(\theta_{m,l})\ z_{m,l}\,dx \geq 0.$$

By using (4.6), (4.7), and (4.8), from (4.5) we obtain

$$\frac{1}{2}\frac{d}{dt}\|z_{m,l}\|^2_{L^2(\Omega)} \le 0.$$

By integrating the previous inequality in time from 0 to t, and by observing that $z_{m,l}(0) = 0$, it follows that $z_{m,l}(t) = 0\ \forall t \in [0,T]$. Consequently,

$$-M_1 \le w_{m,l} \le M_1 \ \text{ in } \overline{Q}. \tag{4.9}$$

We now take $w_{m,l}$ as a test function in the first equation of system (4.3). After integrating by parts, we get

$$\int_\Omega \frac{\partial w_{m,l}}{\partial t} w_{m,l}\,dx + \int_\Omega \nabla\theta_{m,l}\cdot\nabla w_{m,l}\,dx = -\int_{\Gamma_2} g_m(\theta_{m,l})\ w_{m,l}\,dx.$$

Since δ_m is a sequence of smooth nondecreasing monotone functions, (4.9) implies

$$\delta_m(-M_1) \le \theta_{m,l} \le \delta_m(M_1) \ \text{ in } \overline{Q}.$$

Hence,

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|w_{m,l}\|^2_{L^2(\Omega)} + \int_\Omega \alpha_m^{'}|\nabla\theta_{m,l}|^2\,dx &\le \int_{\Gamma_2} |g_m(\gamma_0\theta_{m,l})|\ |\gamma_0 w_{m,l}|\,dx \\ &\le M_1 \max(\|g_m(\delta_m(-M_1))\|_{L^1(\Gamma_2)}, \|g_m(\delta_m(M_1))\|_{L^1(\Gamma_2)}).\end{aligned} \tag{4.10}$$

By using (4.1), (4.2), and the monotonicity of the functions g_m, it follows that, for m sufficiently large,

$$\max(\|g_m(\delta_m(-M_1))\|_{L^1(\Gamma_2)}, \|g_m(\delta_m(M_1))\|_{L^1(\Gamma_2)}) \le c_1. \tag{4.11}$$

By (4.1), (4.10), and (4.11), we also get

$$\frac{1}{2}\frac{d}{dt}\|w_{m,l}\|^2_{L^2(\Omega)} + \zeta_0\int_\Omega |\nabla\theta_{m,l}|^2\,dx \le c_2.$$

Hence, since the sequence $w_{m,l}$ is bounded in $L^\infty(Q)$, by integrating in time the previous inequality it follows that there exists a real positive constant c_3 such that

$$\|\theta_{m,l}\|_{L^2(0,T;\mathcal{W})} \le c_3. \tag{4.12}$$

We now multiply the first equation in (4.4) by $\mathbf{v}_{m,l}$. By standard calculations and using (4.12), we obtain that

$$\|\mathbf{v}_{m,l}\|_{L^\infty(0,T;L^2(\Omega))} \le c_4 \quad \text{and} \quad \|\mathbf{v}_{m,l}\|_{L^2(0,T;V)} \le c_4. \tag{4.13}$$

4.3. Application of the Schauder fixed point theorem. To prove that the map F_l satisfies all the hypotheses of the Schauder fixed point theorem, we have to check that F_l is well-defined and compact. Let us consider the ball B of radius c_4, where c_4 is the constant appearing in (4.13), and let $\{d_1^*(t),\dots,d_l^*(t)\}$ in B. Since the first inequality in (4.13) implies that

$$\left(\sum_{i=1}^{l}\{d^l_{m,i}(t)\}^2\right)^{\frac{1}{2}} \le c_4 \quad \forall t \in [0,T],$$

it follows that $F_l(d_1^*(t),\dots,d_l^*(t)) \in B$. Hence, F_l maps the ball B into itself. Moreover, as we shall prove, the sequence $\{\frac{\partial \mathbf{v}_{m,l}}{\partial t}\}_{l\in\mathbb{N}}$ is uniformly bounded in $L^2(Q)$. Consequently, $F_l(d_1^*(t),\dots,d_l^*(t)) \in H^1(0,T)^l \subset L^\infty(0,t)^l$. Since $H^1(0,T)^l$ is compactly embedded in $L^\infty(0,t)^l$, the map F_l is compact.

We now show that there exists a constant $C(m,\, \|\mathbf{v}_{0,m}\|_V)$ depending on m and on the norm of $\mathbf{v}_{0,m}$ in V but which is independent of l such that

$$\left\| \frac{\partial \mathbf{v}_{m,l}}{\partial t} \right\|_{L^2(Q)} \le C(m,\, \|\mathbf{v}_{0,m}\|_V).$$

Let A_s be the Stokes operator. By multiplying the first equation in (4.4) by $\lambda_i \mathbf{w}_i$, we obtain

$$\int_\Omega \left(\lambda_i \frac{\partial \mathbf{v}_{m,l}}{\partial t} \mathbf{w}_i + \beta(\theta_{m,l})\, \mathbf{v}_{m,l}\, A_s \mathbf{w}_i \right) dx + \nu \langle A_s \mathbf{w}_i,\, \mathbf{v}_{m,l} \rangle$$
$$+ b(\mathbf{v}_{m,l}, \mathbf{v}_{m,l}, A_s \mathbf{w}_i) = \int_\Omega \mathbf{f}(\theta_{m,l}) A_s \mathbf{w}_i \, dx.$$

It follows that

$$\frac{1}{2}\frac{d}{dt}\|\mathbf{v}_{m,l}\|_V^2 + \frac{\nu}{2}\|A_s \mathbf{v}_{m,l}\|_{L^2(\Omega)}^2 \le \sigma_{m,l}\|\mathbf{v}_{m,l}\|_V^2 + c_5 \|\theta_{m,l}\|_{L^2(\Omega)}^2, \tag{4.14}$$

where $\sigma_{m,l} = c_6(\|\mathbf{v}_{m,l}\|_{L^2(\Omega)}^2 \|\mathbf{v}_{m,l}\|_V^2 + \|\theta_{m,l}\|_{\mathcal{W}}^2)$. Since the sequence $\{\theta_{m,l}\}_{l\in\mathbb{N}}$ is bounded in $L^\infty(0,T;L^2(\Omega))$ and $\int_0^T \sigma_{m,l}\, dt < +\infty$, by the Gronwall lemma the sequence $\{\mathbf{v}_{m,l}\}_{l\in\mathbb{N}}$ is also bounded in $L^\infty(0,T;V)$. Back to (4.14), we obtain that

$$\nu \|A_s \mathbf{v}_{m,l}\|_{L^2(Q)}^2 \le \|\mathbf{v}_{m,l}(0)\|_V^2 + 2\int_0^T \left(\sigma_{m,l}\, \|\mathbf{v}_{m,l}(t)\|_V^2 + c_5 \|\theta_{m,l}\|_{L^2(\Omega)}^2 \right) dt,$$

which yields that $\{\mathbf{v}_{m,l}\}_{l\in\mathbb{N}}$ is bounded in $L^2(0,T;H^2(\Omega))$. Since

$$\frac{\partial \mathbf{v}_{m,l}}{\partial t} = -\nu A_s \mathbf{v}_{m,l} - B(\mathbf{v}_{m,l}, \mathbf{v}_{m,l}) - [\beta(\theta_{m,l}) + f(\theta_{m,l})]\, \mathbf{v}_{m,l}$$

and in two dimensions the inequality

$$\|B(\mathbf{u},\, \mathbf{u})\|_{L^2(\Omega)}^2 \le c_7 \|\mathbf{u}\|_V^2\, \|\mathbf{u}\|_{L^2(\Omega)} \|A_s \mathbf{u}\|_{L^2(\Omega)} \quad \forall\, \mathbf{u} \in V \cap H^2(\Omega)$$

holds, the sequence $\{\frac{\partial \mathbf{v}_{m,l}}{\partial t}\}_{l\in\mathbb{N}}$ is bounded in $L^2(Q)$ by a constant depending on m but not on l.

Hence, we conclude that F_l has a fixed point $\mathbf{v}_{m,l} \in \operatorname{span}\langle \mathbf{w}_1, \dots, \mathbf{w}_l \rangle$. This point can be written as

$$\mathbf{v}_{m,l} = \sum_{i=1}^{l} \tilde{d}_{m,i}(t) \mathbf{w}_i.$$

Let

$$\varphi_s = \sum_{i=1}^{s} d_i(t) z_i(x), \quad \boldsymbol{\psi}_s = \sum_{i=1}^{s} \nu_i(t) \mathbf{w}_i(x) \quad d_i, \nu_i \in C^1[0,T], \quad d_i(T) = \nu_i(T) = 0.$$

Then $\theta_{m,l}$ and $\mathbf{v}_{m,l}$ satisfy for $l > s$

$$\begin{aligned}\int_Q \alpha_m^{'} \frac{\partial \theta_{m,l}}{\partial t}\varphi_s + \nabla\theta_{m,l}\cdot\nabla\varphi_s + \{\mathbf{v}_{m,l}\cdot\nabla\theta_{m,l}\}\varphi_s dx\,dt \\ = -\int_0^T \int_{\Gamma_2} g_m(\theta_{m,l})\,\varphi_s\,d\sigma\,dt;\end{aligned} \tag{4.15}$$

$$\begin{aligned}\int_Q \frac{\partial \mathbf{v}_{m,l}}{\partial t}\boldsymbol{\psi}_s + \nu\nabla\mathbf{v}_{m,l} : \nabla\boldsymbol{\psi}_s + \{(\mathbf{v}_{m,l}\cdot\nabla)\mathbf{v}_{m,l}\}\boldsymbol{\psi}_s + \beta(\theta_{m,l})\,\mathbf{v}_{m,l}\,\boldsymbol{\psi}_s\,dx\,dt \\ = \int_Q \mathbf{f}(\theta_{m,l})\boldsymbol{\psi}_s\,dx\,dt.\end{aligned} \tag{4.16}$$

4.4. Passage to the limit for $l \to \infty$. The preceding a priori estimates show that the two sequences $\{\theta_{m,l}\}_{l\in\mathbb{N}}$ and $\{\mathbf{v}_{m,l}\}_{l\in\mathbb{N}}$ are weakly compact, respectively, in $L^\infty(Q)\cap L^2(0,T;\mathcal{W})$ and $L^\infty(0,T;H)\cap L^2(0,T;V)$. Therefore, two subsequences can be selected and relabeled with l in such a way that for $m \in \mathbb{N}$ fixed, as $l \to +\infty$

$$w_{m,l} \longrightarrow w_m \text{ weakly star in } L^\infty(Q),$$

$$\theta_{m,l} \longrightarrow \theta_m \text{ weakly in } L^2(0,T;\mathcal{W}) \quad \text{and} \quad \text{weakly star in } L^\infty(Q),$$

$$\mathbf{v}_{m,l} \longrightarrow \mathbf{v}_m \text{ weakly in } L^2(0,T;V) \quad \text{and} \quad \text{weakly star in } L^\infty(0,T;H).$$

By the continuity of the trace operator γ_0, we have

$$\|G_m(\theta_{m,l})\|_{L^2(\mathcal{W}')} \le c_8. \tag{4.17}$$

By (4.17), since $\frac{\partial w_{m,l}}{\partial t} = k\Delta\theta_{m,l} - \mathbf{v}_{m,l}\cdot\nabla\theta_{m,l}$, we get that the sequences $\{w_{m,l}\}_{l\in\mathbb{N}}$ and $\{\theta_{m,l}\}_{l\in\mathbb{N}}$ are bounded in $H^1(0,T;\mathcal{W}^{'})$. By Theorem 3.1, there exists a subsequence, relabeled by l, such that for $l \to +\infty$ and $m \in \mathbb{N}$ fixed

$$\theta_{m,l} \to \theta_m \text{ strongly in } L^2(Q) \quad \text{and} \quad \mathbf{v}_{m,l} \to \mathbf{v}_m \text{ strongly in } L^2(0,T;H). \tag{4.18}$$

By passing to the limit for $l \to +\infty$ in (4.15) and (4.16), we prove that θ_m and $\mathbf{v}_m$ satisfy

$$\begin{aligned}\int_Q \alpha_m' \frac{\partial \theta_m}{\partial t}\varphi_s + \nabla\theta_m\cdot\nabla\varphi_s + \{\mathbf{v}_m\cdot\nabla\theta_m\}\varphi_s dx\,dt \\ = -\int_0^T \int_{\Gamma_2} g_m(\theta_m)\varphi_s\,d\sigma\,dt;\end{aligned} \tag{4.19}$$

$$\begin{aligned}\int_Q \frac{\partial \mathbf{v}_m}{\partial t}\boldsymbol{\psi}_s + \nu\nabla\mathbf{v}_m : \nabla\boldsymbol{\psi}_s + \{(\mathbf{v}_m\cdot\nabla)\mathbf{v}_m\}\boldsymbol{\psi}_s + \beta(\theta_m)\mathbf{v}_m\boldsymbol{\psi}_s\,dx\,dt \\ = \int_Q \mathbf{f}(\theta_m)\boldsymbol{\psi}_s\,dx\,dt.\end{aligned} \tag{4.20}$$

We now pass to the limit in the constitutive relation $w_{m,l} = \alpha_m(\theta_{m,l})$. It follows from (4.1) that

$$\|w_{m,l} - \alpha_m(\theta_m)\|_{L^2(Q)} \leq m^2 \|\theta_{m,l} - \theta_m\|_{L^2(Q)}.$$

Consequently, by (4.18), for $l \to +\infty$,

$$w_{m,l} = \alpha_m(\theta_{m,l}) \to \alpha_m(\theta_m) \quad \text{strongly in } L^2(Q).$$

By the uniqueness of the weak limit of $w_{m,l}$ in $L^2(Q)$, it can be seen that $w_m = \alpha_m(\theta_m)$. We now claim that

$$\int_0^T \int_{\Gamma_2} g_m(\theta_{m,l})\, \varphi_s \to \int_0^T \int_{\Gamma_2} g_m(\theta_m)\, \varphi_s.$$

It is a well-known result that for every $\varepsilon > 0$ there exists a constant $C(\varepsilon)$ such that

$$\int_0^T \int_{\Gamma_2} |\theta_{m,l} - \theta_m|^2 \, d\sigma \, dt \leq \varepsilon \|\nabla\theta_{m,l} - \nabla\theta_m\|^2_{L^2(Q)} + C(\varepsilon)\|\theta_{m,l} - \theta_m\|^2_{L^2(Q)}.$$

Since the norm $\|\nabla\theta_{m,l} - \nabla\theta_m\|_{L^2(Q)}$ is equibounded by a constant (depending on m but not on l) and the sequence $\theta_{m,l}$ strongly converges to θ_m in $L^2(Q)$, the above inequality implies that for $l \to \infty$

$$\theta_{m,l} \to \theta_m \quad \text{strongly in } L^2(\Sigma). \tag{4.21}$$

Moreover, there exists a constant $c(m)$, depending on m, such that

$$\|g_m(x,t,\theta_{m,l}) - g_m(x,t,\theta_m)\|_{L^1(\Sigma)} \leq c(m)\|\theta_{m,l}(x,t) - \theta_m(x,t)\|_{L^1(\Sigma)}.$$

Hence, by (4.21), for $l \to \infty$

$$\int_0^T \int_{\Gamma_2} g_m(\theta_{m,l})\, \varphi_s d\sigma \, dx \longrightarrow \int_0^T \int_{\Gamma_2} g_m(\theta_m)\, \varphi_s d\sigma \, dx.$$

4.5. Passage to the limit for $m \to \infty$. To conclude the proof we have to pass to the limit for $m \to \infty$.

We take as test functions in (4.19) and (4.20), respectively, w_m and $\mathbf{v}_m$. By standard calculations, we find that the sequences $\{w_m\}_{m\in\mathbb{N}}$ and $\{\mathbf{v}_m\}_{m\in\mathbb{N}}$ are uniformly bounded in $L^\infty(Q) \cap L^2(0,T;\mathcal{W}) \cap H^1(0,T;\mathcal{W}')$ and in $L^2(0,T;V) \cap L^\infty(0,T;H) \cap H^1(0,T;V')$ by a real positive constant not depending on m. By virtue of Theorem 3.1, the sequences $\{\theta_m\}_{m\in\mathbb{N}}$ and $\{\mathbf{v}_m\}_{m\in\mathbb{N}}$ strongly converge in $L^2(Q)$ and in $L^2(0,T;H)$, respectively, to θ and $\mathbf{v}$.

We now show that $\theta = \alpha^{-1}(w)$. By the monotonicity of the sequence δ_m, for every $s \in L^\infty(Q)$

$$\begin{aligned} &\int_0^T \int_\Omega [\delta_m(w_m(x,t)) - \delta_m(s(x,t))](w_m(x,t) - s(x,t))\, dx\, dt \\ &\qquad = \int_0^T \int_\Omega [\theta_m(x,t) - \delta_m(s)](w_m - s)\, dx\, dt \geq 0. \end{aligned} \tag{4.22}$$

We remark that

$$\int_0^T \int_\Omega \delta_m(w_m) w_m = \int_0^T \int_\Omega \theta_m w_m \, dx \, dt \longrightarrow \int_0^T \int_\Omega \theta w \, dx \, dt.$$

Since $\delta_m \to \alpha^{-1}$ in $C^0(\mathbb{R})$, by passing to the limit for $m \to \infty$ in (4.22), we get

$$\int_0^T \int_\Omega [\theta(x,t) - \alpha^{-1}(s(x,t))](w(x,t) - s(x,t)) \, dx \, dt \geq 0 \quad \forall s \in L^\infty(Q). \tag{4.23}$$

Since the real function α^{-1} is monotone, (4.23) implies that

$$\int_0^T \int_\Omega [\theta - \alpha^{-1}(w)](w - s) \, dx \, dt \geq 0 \quad \forall s \in L^\infty(Q),$$

from which we get

$$\theta = \alpha^{-1}(w) \quad \text{a.e. in } Q.$$

We now claim that

$$\int_0^T \int_{\Gamma_2} g_m(\theta_m) \gamma_0 \varphi \, dx \longrightarrow \int_0^T \int_{\Gamma_2} g(\gamma_0 \theta) \gamma_0 \varphi \, dx \quad \forall \varphi \in C^\infty(\overline{Q}).$$

Since the sequence $g_m(\theta_m)$ is bounded in $L^2(\Sigma)$, it follows that $g_m(\theta_m)$ weakly converges to ξ in $L^2(\Sigma)$. We prove that $\xi = g(\gamma_0 \theta)$. By virtue of the monotonicity of the functions g_m,

$$\int_\Sigma [g_m(\theta_m) - g_m(\gamma_0 s)] \gamma_0(\theta_m - s) \, dx \, dt \geq 0 \quad \forall s \in L^2(0,T; H^1(\Omega)) : \gamma_0 s \in L^\infty(\Sigma).$$

For $m \to \infty$,

$$\int_0^T \int_{\Gamma_2} g_m(\theta_m) \theta_m \, dx \, dt \to \int_0^T \int_{\Gamma_2} \xi(\gamma_0 \theta) dx \, dt.$$

On the other hand, by Lebesgue's dominate convergence theorem

$$g_m(x,t,\gamma_0 s(x,t)) \to g(x,t,\gamma_0 s(x,t)) \quad \text{strongly in } L^2(\Sigma).$$

It follows that

$$\int_0^T \int_{\Gamma_2} [\xi - g(\gamma_0 s)] \gamma_0(\theta - s) \, dx \geq 0 \quad \forall s \in L^2(0,T; H^1(\Omega)) : \gamma_0 s \in L^\infty(\Sigma).$$

Since g is a monotone function, we conclude that

$$\xi = g(\gamma_0 \theta) \quad \text{a.e. on } \Sigma. \qquad \square$$

REFERENCES

[1] C. Beckermann and R. Viskanta, *Effect of solid subcooling on natural convection melting of a pure metal*, ASME J. Heat Transfer, 11 (1989), pp. 416–424.

[2] A.D. Brent, V.R. Voller, and K.J. Reid, *Enthalpy-porosity technique for modeling convection-diffusion phase change: Application to the melting of a pure metal*, Numer. Heat Transfer, 13 (1988), pp. 297–318.

[3] H. Brézis, *Opérateurs Maximaux Monotones et Semi-Groups de Contractions dans les Espaces de Hilbert*, North-Holland, Amsterdam, 1973.

[4] J.R. Cannon and E. DiBenedetto, *On the existence of weak-solutions to an n-dimensional Stefan problem with nonlinear boundary conditions*, SIAM J. Math. Anal., 11 (1980), pp. 632–645.

[5] J.R. Cannon and E. DiBenedetto, *The steady state Stefan problem with convection, with mixed temperature and nonlinear heat flux boundary conditions*, in Free Boundary Problems, Pavia, 1979, Istituto Nazionale Alta Matematica, Francesco Severi, Rome, 1980, pp. 231–265.

[6] J.R. Cannon, E. DiBenedetto, and G.H. Knightly, *The bidimensional Stefan problem with convection: The time dependent case*, Comm. Partial Differential Equations, 8 (1983), pp. 1549–1604.

[7] E. Casella and M. Giangi, *An analytical and numerical study of the Stefan problem with convection by means of an enthalpy method*, Math. Methods Appl. Sci., 24 (2001), pp. 623–639.

[8] P. Constantin and C. Foias, *Navier-Stokes Equations*, University of Chicago Press, Chicago, IL, 1988.

[9] A. Damlamian and N. Kenmochi, *Le problème de Stefan avec conditions latérales variables*, Hiroshima Math. J., 10 (1980), pp. 271–293.

[10] E. DiBenedetto and M. O'Leary, *Three-dimensional conduction-convection problems with change of phase*, Arch. Ration. Mech. Anal., 123 (1993), pp. 99–116.

[11] A. Fasano and M. Primicerio, *Il problema di Stefan con condizioni al contorno nonlineari*, Ann. Scuola Norm. Sup. Pisa (3), 26 (1972), pp. 711–737.

[12] M. Giangi and F. Stella, *Melting of a pure metal on a vertical wall: Numerical simulation*, Numer. Heat Transfer, Part A, 38 (2000), pp. 193–208.

[13] M. Giangi, F. Stella, and T.A. Kowalewski, *Phase change problems with free convection: Fixed grid numerical simulation*, Comput. Visual. Sci., 2 (1999), pp. 123–130.

[14] J.L. Lions, *Quelque méthodes de résolution des problémes aux limitex non-lineaires*, Dunod, Gauthier-Villars, Paris, 1969.

[15] A. Rady and A.K. Mohanty, *Natural convection during melting and solidification of pure metals in a cavity*, Numer. Heat Transfer, Part A, 29 (1996), pp. 49–63.

[16] A. Visintin, *Sur le problème de Stefan avec flux non lineaire*, Boll. Un. Mat. Ital. C (5), 18 (1981), pp. 63–86.

[17] V.R. Voller, M. Cross, and N.C. Markatos, *An enthalpy method for convection diffusion phase change*, Internat. J. Numer. Methods Engrg., 24 (1987), pp. 271–284.

[18] X.F. Wang, *The Stefan problem with non linear convection*, J. Partial Differential Equations, 5 (1992), pp. 66–86.

SIAM J. MATH. ANAL.
Vol. 35, No. 2, pp. 389–422

GLOBAL EXISTENCE OF SMOOTH SOLUTIONS OF THE *N*-DIMENSIONAL EULER–POISSON MODEL*

G. ALÌ†

Abstract. The global existence of smooth solutions of the Cauchy problem for the N-dimensional Euler–Poisson model for semiconductors is established, under the assumption that the initial data is a perturbation of a stationary solution of the drift-diffusion equations with zero electron velocity, which is proved to be unique. The resulting evolutionary solutions converge asymptotically in time to the unperturbed state. The singular relaxation limit is also discussed.

Key words. Euler–Poisson, semiconductors, asymptotic behavior, smooth solutions

AMS subject classifications. 35L65, 76X05, 35M10

DOI. 10.1137/S0036141001393225

1. Introduction.

Hydrodynamic models for semiconductors were introduced about thirty years ago [8, 9] to describe the electron flow in semiconductor devices when the transport of energy plays a crucial role, as in submicron devices or in the occurrence of high field phenomena [39]. These models all consist of a set of balance laws for the moments of the electron distribution density, derived from the infinite hierarchy of moment equations of the semiclassical Boltzmann equation for semiconductors, coupled with the electric potential through a Poisson equation. Closure relations for the moment fluxes and the collision terms can be determined by using the maximum entropy principle [33, 18], physically set in the framework of extended thermodynamics [40, 30] (see [6], and also [5, 7] for a review).

In this paper, we consider the Euler–Poisson model, obtained from the first three moments: electron density, momentum, and energy. In rescaled variables, let n, $\boldsymbol{u} = (u^1, \dots, u^N)$, p, e, T, and ϕ denote the electron number density, the electron velocity, the electron pressure, the electron internal energy, the electron temperature, and the electric potential, respectively. All the dependent variables are functions of $(\boldsymbol{x}, t) \equiv (x^1, \dots, x^N, t) \in \mathbb{R}^N \times \mathbb{R}$. The (nondimensional) Euler–Poisson model consists of a hydrodynamic part,

$$\frac{\partial n}{\partial t} + \sum_{r=1}^{N} \partial_r (n u^r) = 0, \tag{1.1}$$

$$\frac{\partial}{\partial t}(n u^i) + \sum_{r=1}^{N} \partial_r \left(n u^i u^r + p \delta^{ir} \right) = n \partial_i \phi - \frac{n u^i}{\tau}, \tag{1.2}$$

$$\begin{aligned} \frac{\partial}{\partial t} \left(\frac{n |\boldsymbol{u}|^2}{2} + e \right) + \sum_{r=1}^{N} \partial_r \left[\left(\frac{n |\boldsymbol{u}|^2}{2} + e + p \right) u^r \right] \\ = \sum_{r=1}^{N} n u^r \partial_r \phi - \frac{1}{\sigma} \left[\frac{n |\boldsymbol{u}|^2}{2} + \left(\frac{\partial e}{\partial T} \right)_n (T - T^*) \right], \end{aligned} \tag{1.3}$$

*Received by the editors July 30, 2001; accepted for publication (in revised form) January 15, 2003; published electronically July 18, 2003.
http://www.siam.org/journals/sima/35-2/39322.html
†Istituto per le Applicazioni del Calcolo, CNR, via P. Castellino, 111 I-80131, Napoli, Italy (ali@iam.na.cnr.it).

supplemented by the Poisson equation

$$\Delta\phi = n - b. \tag{1.4}$$

Here, $\partial_r = \partial/\partial x^r$, $\Delta = \sum_{r=1}^N \partial_r^2$, $|\boldsymbol{u}|^2 = \sum_{r=1}^N (u^r)^2$, and δ^{ir} is the Kronecker symbol, equal to 1 if $i = r$, and equal to 0 otherwise. The positive constants τ and σ are the (rescaled) momentum relaxation time and the (rescaled) energy relaxation time, respectively, T^* is the (constant) equilibrium temperature, and the function $b(\boldsymbol{x})$ is the doping profile, satisfying the conditions

$$b \in L^\infty(\mathbb{R}^N), \qquad \sup_{\boldsymbol{x}\in\mathbb{R}^N} b(\boldsymbol{x}) = b^+ \geq \inf_{\boldsymbol{x}\in\mathbb{R}^N} b(\boldsymbol{x}) = b^- > 0, \tag{1.5}$$

$$Db \equiv (\partial_1 b, \partial_2 b, \dots, \partial_N b) \in H^s(\mathbb{R}^N), \qquad s > \frac{N}{2} + 1. \tag{1.6}$$

Consistent with extended thermodynamics, the system (1.1)–(1.4) is closed once we specify a convex state function $e(n, S)$, which satisfies the differential relation

$$d\left(\frac{e}{n}\right) = T d\left(\frac{S}{n}\right) - p d\left(\frac{1}{n}\right). \tag{1.7}$$

Then the temperature T and the pressure p are given as functions of the number density n and of the entropy S by the relations

$$T = \frac{\partial e}{\partial S}, \qquad p = \frac{\partial e}{\partial n} n + \frac{\partial e}{\partial S} S - e. \tag{1.8}$$

In particular, the partial derivative of e with respect to T keeping n constant, which appears in (1.3), can be expressed as

$$\left(\frac{\partial e}{\partial T}\right)_n = T \left(\frac{\partial S}{\partial T}\right)_n = \frac{\partial e/\partial S}{\partial T/\partial S}.$$

We expect (1.1)–(1.4) to determine a unique solution once we specify initial data for the hydrodynamical quantities (that is, for the electron velocity and for two thermodynamic variables) and appropriate boundary conditions. Heuristically speaking, when we extend the Euler–Poisson equations to the whole space, we do not have the problem of modeling contacts or insulated parts of the boundary. All we need to do is look for an equilibrium state and require that the state described by the solution be equal to that equilibrium state at infinity. In fact, we will prove the existence of a unique state $(n^*, \boldsymbol{u}^*, S^*, D\phi^*)$ in total thermodynamic equilibrium [1], which allows us to assign conditions at infinity for the hydrodynamic variables and the electric field. This result is consistent with the one-dimensional analysis [36, 37], which shows that it is natural to assign initially the electric field $E^i = \partial_i \phi$ compatibly with the Poisson equation instead of specifying the electric potential at infinity.

With this motivation, the system (1.1)–(1.4) is supplemented with the initial data

$$\begin{aligned} &n(\boldsymbol{x}, 0) = n_0(\boldsymbol{x}), \quad u^i(\boldsymbol{x}, 0) = u_0^i(\boldsymbol{x}), \\ &S(\boldsymbol{x}, 0) = S_0(\boldsymbol{x}), \quad \partial_i\phi(\boldsymbol{x}, 0) = E_0^i(\boldsymbol{x}), \end{aligned} \tag{1.9}$$

with

$$\sum_{i=1}^N \partial_i E_0^i = n_0 - b, \tag{1.10}$$

and with the boundary conditions

$$(1.11)\qquad \begin{aligned} &n(\boldsymbol{x},t)-n^*(\boldsymbol{x}),\quad u^i(\boldsymbol{x},t)-u^{*i}(\boldsymbol{x}),\\ &S(\boldsymbol{x},t)-S^*(\boldsymbol{x}),\quad E^i(\boldsymbol{x},t)-E^{*i}(\boldsymbol{x})\in H^s(\mathbb{R}^N),\qquad s>\frac{N}{2}+1. \end{aligned}$$

Here, we have introduced the electric field $E^{*i}=\partial_i\phi^*$. Of course, the condition (1.11) tells us much more than the behavior of the solution at infinity, but it also sets the natural functional space needed to study the global existence of smooth solutions of a hyperbolic system. In addition, we require that the electric flux be zero at infinity, that is,

$$(1.12)\qquad \frac{\partial \boldsymbol{E}}{\partial t}\in H^s(\mathbb{R}^N),\qquad s>\frac{N}{2}+1.$$

The only requirement for the unperturbed state is that

$$(1.13)\qquad \tau\sigma\,\|Dn^*\|_{L^2}^2<C^*$$

for a certain constant C^* which depends only on the equation of state and on $b^{\pm}$.

Now, we assume that the initial data is sufficiently close to the state in total thermodynamic equilibrium; that is, the differences

$$n_0(\boldsymbol{x})-n^*(\boldsymbol{x}),\quad u_0^i(\boldsymbol{x})-0,\quad S_0(\boldsymbol{x})-S^*(\boldsymbol{x}),\quad E_0^i(\boldsymbol{x})-E^{*i}(\boldsymbol{x})$$

belong to $H^s(\mathbb{R}^N)$, $s>\frac{N}{2}+1$, and their H^s-norms are small enough. Under these assumptions, we will show that the solution of the initial value problem (1.1)–(1.4) exists uniquely and globally in time and that it is a classical solution for $t>0$. Moreover, it decays exponentially in the H^s-norm to the stationary solution, according to the estimate

$$(1.14)\qquad \begin{aligned} &\|(n-n^*,\boldsymbol{u},S-S^*,\boldsymbol{E}-\boldsymbol{E}^*)(\cdot,t)\|_{H^s}^2\\ &\le Ke^{-c\tau t}\,\|(n-n^*,\boldsymbol{u},S-S^*,\boldsymbol{E}-\boldsymbol{E}^*)(\cdot,0)\|_{H^s}^2, \end{aligned}$$

with K and c positive constants which depend only on the equation of state, $e=e(n,S)$, on the equilibrium density, n^*, and on the product of the relaxation times, $\tau\sigma$. These results extend similar results obtained in [3] for the one-dimensional Euler–Poisson model.

Some remarks are in order. First, the condition (1.6) for the doping profile is just a technical one. In fact, the H^s-norm of b can be arbitrarily large, and we can assume the doping to be approximated by an appropriate function satisfying (1.6). Nevertheless, the condition (1.13) for $\|Dn^*\|_{L^2}$ can be read as a severe restriction for $\|Db\|_{L^2}$, which rules out the possibility of dopings with a sharp short well discontinuity, at least when $\tau\sigma$ is not small. This is a consequence of Theorem 3.1, where we prove some a priori estimates which ensure the existence of an equilibrium state. The restriction on the doping profile can be removed in the one-dimensional case, at least for a polytropic equation of state, as shown in [3]. We expect a similar result to hold also for higher space dimensions, although the proof may require different techniques from the ones used in this work.

Second, the smallness of the initial velocity and of $\|Db\|_{L^2}$ restricts implicitly our results to the subsonic case. Under transonic and supersonic flow conditions, the issue of global existence for the full hydrodynamic model remains essentially an open problem, since smooth solutions cease to exist in finite time [12].

A third remark concerns the singular relaxation limit, that is, the behavior of the solutions of the Euler–Poisson equations as the momentum relaxation time τ tends to zero and the product $\tau\sigma$ tends to a positive constant. We will show in section 4 that the estimate (1.14) is sufficient to establish the relaxation limit under the scaling $\tau t \to t$, $\boldsymbol{u}/\tau \to \boldsymbol{u}$. This limit provides a connection between the class of (hyperbolic) hydrodynamic models and the class of (parabolic) energy-transport models [2, 4]. For a review of the energy-transport models we refer to [13].

The Cauchy problem and the initial-boundary value problem of the one-dimensional hydrodynamic model for semiconductors and its relaxation limit have been studied extensively by many authors for isentropic flows, both in the stationary case [16, 20] and in the evolutionary case [11, 28, 29, 34, 36, 37, 41, 45, 46]. For the full Euler–Poisson model with additional heat conductivity, a preliminary numerical and theoretical study can be found in [21]. More recently, for the same model, the global existence of smooth solutions in a bounded domain for small initial data has been proved in [10, 47], under the assumption that the doping profile is close enough to a constant function. In [10], the singular relaxation hyperbolic–parabolic limit is also studied, assuming a uniform bound for an appropriate combination of the relaxation times. For the Euler–Poisson, in [3] the asymptotic behavior of solutions to the Cauchy problem and the convergence to the steady state solution are proven. Gasser and Natalini [22] have studied the zero relaxation convergence of the weak solutions to the corresponding drift-diffusion equations.

Very little is known so far in the multidimensional case. Besides the local classical solutions obtained in [31, 35], only steady state solutions in the subsonic case [17] and dynamic solutions with geometrical structure (symmetry) [11, 19, 25, 26] or without vorticity [23] have been studied. The existence of global smooth solutions for the general multidimensional model have been established in [24] for the isentropic case and in [27] for the full hydrodynamic system with heat conduction.

The approach followed in this paper relies essentially on the extended thermodynamic formulation of the system (1.1)–(1.4), that is, on the Legendre duality of the pressure and the internal energy expressed by (1.8). In particular, using new high order estimates, we are able to obtain arbitrarily smooth solutions if the initial data is smooth enough (in the sense of $H^s(\mathbb{R}^N)$ for any $s > N/2 + 1$). The proof of the estimate (1.14) extends energy methods already used for the one-dimensional model in [3] and for the parabolic energy-transport model in [14, 15] (see also [1]). To our knowledge, similar energy estimates appeared for the first time in the classical paper [38]. We mention also the recent paper [44], which studies the existence of smooth solutions for a general class of hyperbolic systems with relaxation.

The plan of the paper is the following. In section 2, we introduce the notation for the derivatives and for the norms that will be used throughout the paper. Also, we state some technical lemmas, whose proofs are given in the appendix. In section 3, we study the equilibrium states and their existence and their dependence on the doping profile. The subsequent section is the main section of the paper, where we state and prove the theorems concerning the global existence of a classical solution to the Euler–Poisson system and its singular relaxation limit. All these results depend on a key a priori estimate, which is established in the last two sections. Namely, in section 5 we study the positive definiteness of some Liapunov functionals which are used in the proof of the a priori estimate. The proof itself is given in section 6.

2. Notation and basic lemmas. We start this section by introducing two different notations for the components of the kth derivative operator [43]. The derivative

operator is defined by

$$D = (\partial_1, \partial_2, \dots, \partial_N).$$

Then, for any $k \geq 0$, the kth derivative operator is defined by

$$D^0 \stackrel{\text{def}}{=} I, \qquad D^k \stackrel{\text{def}}{=} DD^{k-1}, \qquad k \geq 1.$$

Let J be a k-tuple of integers between 1 and N, $J = (j_1, \dots, j_k)$. We set

$$\partial_J = \partial_{j_1}\partial_{j_2}\cdots\partial_{j_k},$$

and $|J| = k$, the total order of differentiation. Since partial derivatives with different indices commute, it is convenient to introduce an alternative notation. Let α be an N-tuple of nonnegative integers, $\alpha = (\alpha_1, \dots, \alpha_N)$. We set

$$\partial^\alpha = \partial_1^{\alpha_1}\partial_2^{\alpha_2}\cdots\partial_N^{\alpha_N}.$$

The total order of derivation is $|\alpha| = \sum_{i=1}^N \alpha_i$. If $|J| = |\alpha| = k$ and $f \in C^k(\mathbb{R}^N)$, then

$$\partial_J f = \partial^\alpha f,$$

with $\alpha_i = \#\{l : j_l = i\}$. In this case, we write $\alpha = \alpha(J)$.

For any $f, g \in C^k(\mathbb{R}^N)$, we consider $D^k f$, $D^k g$ as vectors in $(C^0)^{N^k}$ and define the Euclidean norm and scalar product

$$\left|D^k f\right| = \left(\sum_{|J|=k} (\partial_J f)^2\right)^{1/2}, \qquad D^k f \cdot D^k g = \sum_{|J|=k} (\partial_J f)(\partial_J g).$$

Using the α-multi-index notation, we have

$$\left|D^k f\right| = \left(\sum_{|\alpha|=k} \nu(\alpha)(\partial^\alpha f)^2\right)^{1/2}, \qquad D^k f \cdot D^k g = \sum_{|\alpha|=k} \nu(\alpha)(\partial^\alpha f)(\partial^\alpha g),$$

with

$$\nu(\alpha) = \#\{J : \alpha = \alpha(J)\} = \frac{k!}{\alpha!}, \qquad \alpha! = \alpha_1! \cdots \alpha_N!.$$

We can extend these definitions to a vector-valued function $\boldsymbol{v} = (v^1, \dots, v^N) \in (C^k(\mathbb{R}^N))^N$ by

$$\left|D^k \boldsymbol{v}\right| = \sum_{r=1}^N \sum_{|J|=k} (\partial_J v^r)^2, \qquad D^k \boldsymbol{v} \cdot D^k \boldsymbol{w} = \sum_{r=1}^N \sum_{|J|=k} (\partial_J v^r)(\partial_J w^r).$$

The symbol $\otimes$ will denote the tensorial product, which combines two symmetric tensors with k and l indices to give a symmetric tensor with $k+l$ indices. In particular, we have the following formula for the kth derivative of the product of two functions:

$$D^k(fg) = \sum_{r=0}^k \binom{k}{r} (D^r f) \otimes (D^{k-r} g). \tag{2.1}$$

All the functional spaces will be considered on $\mathbb{R}^N$, so we will omit the argument. We will use the following norms:

$$\|f\| = \|f\|_{L^2} = \left(\int |f(\boldsymbol{x})|^2 \, d\boldsymbol{x} \right)^{\frac{1}{2}},$$
$$\|f\|_{H^k} = \sum_{i=0}^{k} \left\| D^i f \right\|,$$
$$\|f\|_{L^\infty} = \sup_{\boldsymbol{x} \in \mathbb{R}^N} |f(\boldsymbol{x})|,$$
$$\|f\|_{C^k} = \sum_{i=0}^{k} \left\| D^i f \right\|_{L^\infty}.$$

In the following, the symbol C will denote a generic positive constant. Sometimes we will write $C(a_1, \dots, a_m)$ to signify that the generic constant C depends on the arguments $a_1, \dots, a_m$. We will use the following classical lemma [35, 32].

LEMMA 2.1 (Moser-type calculus). *If f, $g \in H^k \cap L^\infty$, then we have*

$$\left\| D^k(fg) \right\| \leq C(k) \left(\|f\|_{L^\infty} \left\| D^k g \right\| + \|g\|_{L^\infty} \left\| D^k f \right\| \right). \tag{2.2}$$

If $f \in H^k$, $Df \in L^\infty$, $g \in H^{k-1} \cap L^\infty$, then we have

$$\left\| D^k(fg) - f D^k(g) \right\| \leq C(k) \left(\|Df\|_{L^\infty} \left\| D^{k-1} g \right\| + \|g\|_{L^\infty} \left\| D^k f \right\| \right). \tag{2.3}$$

If $F(w)$ is a smooth vector-valued function and $f(\boldsymbol{x})$ is a continuous function which takes values into a compact subset of the domain Ω of F, with $f \in H^k \cap L^\infty$, then we have

$$\left\| D^k F(f) \right\| \leq C(k, F, \|f\|_{L^\infty}) \left\| D^k f \right\|, \tag{2.4}$$

where

$$C(k, F, \|f\|_{L^\infty}) = C(k) \left| \frac{\partial F}{\partial w} \right|_{C^{k-1}} \left(\sum_{\mu=1}^{k} \|f\|_{L^\infty}^{\mu-1} \right),$$

with

$$\left| \frac{\partial F}{\partial w} \right|_{C^k} = \sup_{w \in \Omega} \sum_{i=0}^{k} \left| \left(\frac{\partial}{\partial w} \right)^i F(w) \right|.$$

We also use the following results, which will be proved in the appendix.

LEMMA 2.2. *If $1 \leq j \leq k$ and $D^j f$, $g \in H^{k-j}$, then we have*

$$\left\| D^k(fg) - \sum_{r=0}^{j-1} \binom{k}{r} (D^r f) \otimes (D^{k-r} g) \right\| \leq C(k, j) \left\| D^j f \right\|_{H^{k-j}} \|g\|_{H^{k-j}}. \tag{2.5}$$

LEMMA 2.3. *If $F(w)$ is a smooth vector-valued function and $f(\boldsymbol{x})$, $g(\boldsymbol{x})$ are continuous functions which take values into a compact subset of the domain Ω of F, with $g \in L^\infty$, $Dg \in H^{k-1}$, and $f - g \in H^k \cap L^\infty$, then we have*

$$\left\| D^k(F(f) - F(g)) \right\| \leq C(F, f, g) \left(\left\| D^k(f - g) \right\| + \|Dg\|_{H^{k-1}} \|f - g\|_{H^{k-1}} \right), \tag{2.6}$$

$$\left\| D^k(F(f) - F(g)) - \frac{\partial F}{\partial w}(g) D^k(f - g) \right\| \leq C(F, f, g) \left(\|f - g\|_{L^\infty} \left\| D^k(f - g) \right\| + \|Dg\|_{H^{k-1}} \|f - g\|_{H^{k-1}} \right), \tag{2.7}$$

where the constant C depends on the C^k-norm of the derivative of F, on $\|g\|_{L^\infty}$, and on $\|f-g\|_{L^\infty}$.

COROLLARY 2.4. *Under the same hypothesis as that of Lemma 2.3, we have*

$$\begin{aligned}(2.8)\qquad &\left\|D^{k-1}(F(f)Df-F(g)Dg)\right\|\\ &\quad\le C(F,f,g)(\left\|D^k(f-g)\right\|+\|Dg\|_{H^{k-1}}\|f-g\|_{H^{k-1}}),\\ (2.9)\qquad &\left\|D^{k-1}(F(f)Df-F(g)Dg)-F(g)D^k(f-g)\right\|\\ &\quad\le C(F,f,g)(\|f-g\|_{L^\infty}\left\|D^k(f-g)\right\|+\|Dg\|_{H^{k-1}}\|f-g\|_{H^{k-1}}).\end{aligned}$$

Moreover, assuming $Dg\in H^k\cap L^\infty$ and $D(f-g)\in L^\infty$, we have

$$\begin{aligned}(2.10)\qquad &\Big\|D^k\left[F(f)\partial f-F(g)\partial g\right]-F(f)D^k\partial(f-g)\\ &\quad-kD\left[F(g)\right]\otimes D^{k-1}\left[\partial(f-g)\right]-\frac{\partial F}{\partial w}(g)D^k(f-g)\partial g\Big\|\\ &\qquad\le C(F,f,g)(\|f-g\|_{C^1}\left\|D^k(f-g)\right\|+\|Dg\|_{H^k}\|f-g\|_{H^{k-1}}),\end{aligned}$$

where the constant C depends on the C^k-norm of the derivative of F, on $\|g\|_{C^1}$, and on $\|f-g\|_{C^1}$.

We close this section with a technical lemma.

LEMMA 2.5. *If $f\in H^k$, $k>1$, then for any positive constants K and η we can choose positive $\eta_1,\dots,\eta_k$ such that*

$$(2.11)\qquad \sum_{j=1}^{k}\eta_j\|f\|_{H^{j-1}}\left\|D^jf\right\|\le K\left(\eta\|f\|^2+\sum_{j=1}^{k}\eta_j\left\|D^jf\right\|^2\right).$$

In particular, the thesis of the lemma is satisfied by

$$\eta_j=2\eta\left(\frac{K^2}{1+K^2}\right)^j.$$

All the results of this section hold also when f and g are vector-valued functions.

3. Total thermodynamic equilibrium. In this section we characterize a state in total thermodynamic equilibrium [1] and show that it satisfies the drift-diffusion equation with zero electron flux. The main result of the section is the proof of the a priori estimates (3.11)–(3.12), which guarantee the existence of such a state. Moreover, we prove that there exists a unique state in total thermodynamic equilibrium.

A state $(n^*,\boldsymbol{u}^*,S^*,\phi^*)$ in total thermodynamic equilibrium is characterized by the conditions

$$(3.1)\qquad \boldsymbol{u}^*=0,$$

$$(3.2)\qquad \left(\frac{\partial e}{\partial n}\right)^*=\phi^*+\text{constant},\qquad \left(\frac{\partial e}{\partial S}\right)^*=T^*=\text{constant}.$$

Moreover, the electron density is related to the electric potential by the Poisson equation

$$(3.3)\qquad \Delta\phi^*=n^*-b.$$

Deriving the second relation in (1.8) and using (3.2), it is possible to see that (n^*, S^*, ϕ^*) satisfies the stationary drift-diffusion model

$$\begin{aligned} Dp^* &= n^* D\phi^*, \\ \Delta\phi^* &= n^* - b, \end{aligned} \tag{3.4}$$

with the constraint

$$\left(\frac{\partial e}{\partial S}\right)^* = T^*. \tag{3.5}$$

In other words, $(n, \boldsymbol{u}, S, \phi) = (n^*, 0, S^*, \phi^*)$ is a stationary solution of the Euler–Poisson system.

Two other relevant identities can be recovered by deriving (3.2) and solving for Dn^* and DS^*. We obtain

$$DS^* = -\left(\frac{\partial T/\partial n}{\partial T/\partial S}\right)^* Dn^*, \tag{3.6}$$

$$\left(\frac{H(e)}{\partial T/\partial S}\right)^* Dn^* = D\phi^*, \tag{3.7}$$

where the superscript * denotes evaluation at equilibrium, and the Hessian determinant of e, $H(e)$, is defined by

$$H(e) = \frac{\partial^2 e}{\partial n^2}\frac{\partial^2 e}{\partial S^2} - \left(\frac{\partial^2 e}{\partial n \partial S}\right)^2. \tag{3.8}$$

We remark that both $H(e)$ and $\partial T/\partial S = \partial^2 e/\partial S^2$ are positive, due to the convexity of $e(n, S)$.

For a polytropic equation of state,

$$e = \Gamma \exp\left(\frac{S}{\Gamma n}\right) n^\gamma, \qquad \Gamma = \frac{1}{\gamma - 1},$$

the identities (3.6) and (3.7) reduce to

$$DS^* = \left(\frac{S^*}{n^*} - 1\right) Dn^*, \qquad \frac{T^*}{n^*} Dn^* = D\phi^*.$$

In this case, the entropy at equilibrium is given by

$$S^* = -n^* \log\left(\frac{n^*}{T^{*\Gamma}}\right).$$

In the remainder of this section, we consider the system

$$\begin{aligned} \Theta(n^*, S^*) Dn^* &= D\phi^*, \\ \Delta\phi^* &= n^* - b, \end{aligned} \tag{3.9}$$

where

$$\Theta(n, S) = \frac{H(e)}{\partial^2 e/\partial S^2} \geq 0, \tag{3.10}$$

and S^* is given as a function of n^* through (3.5). In the following, we will regard Θ as a function of density only and write $\Theta(n, S(n)) = \Theta(n)$.

THEOREM 3.1 (a priori estimates). *Let b satisfy* (1.5) *and* (1.6). *If $(n^*, \phi^*) \in C^2$ is a solution of* (3.9), *with $n^* - b \in H^s$, $s > \frac{N}{2} + 1$, then the following estimates hold:*

$$\inf_{\boldsymbol{x}\in\mathbb{R}^N} b(\boldsymbol{x}) \le n^*(\boldsymbol{x}) \le \sup_{\boldsymbol{x}\in\mathbb{R}^N} b(\boldsymbol{x}), \tag{3.11}$$

$$\|n^* - b\|^2_{H^{s+1}} \le C_s(\Theta, b)\, \|Db\|^2_{H^s}\,, \tag{3.12}$$

$$\|Dn^*\|^2_{H^s} \le c_s(\Theta, b)\, \|Db\|^2_{H^s}\,, \tag{3.13}$$

where the constants C_s and c_s depend only on the function $\Theta(n)$, $n \in [b^-, b^+]$, and on the function b.

Proof. Since $n^* - b \in H^s$, we have

$$\lim_{|\boldsymbol{x}|\to\infty} [n^*(\boldsymbol{x}) - b(\boldsymbol{x})] = 0.$$

Then, for any arbitrary small constant $\delta \ge 0$ there exists $A \ge 0$ such that

$$b(\boldsymbol{x}) - \delta \le n^*(\boldsymbol{x}) \le b(\boldsymbol{x}) + \delta \qquad \text{for all } |\boldsymbol{x}| \ge A. \tag{3.14}$$

Suppose that inside the open ball of radius A there exists a point $\boldsymbol{x}_0$ at which the density reaches its minimum, that is,

$$n^*(\boldsymbol{x}_0) = \min_{|\boldsymbol{x}|<A} n^*(\boldsymbol{x}).$$

Then we have $Dn^*(\boldsymbol{x}_0) = 0$, and $D^2 n^*(\boldsymbol{x}_0)$ is a positive definite quadratic form. This, in particular, implies $\Delta n^*(\boldsymbol{x}_0) \ge 0$. Using (3.9), we can write a second order partial differential equation for n^*,

$$D \cdot (\Theta(n^*) Dn^*) = n^* - b. \tag{3.15}$$

It follows that $n^*(\boldsymbol{x}_0) - b(\boldsymbol{x}_0) = \Theta(n^*(\boldsymbol{x}_0))\Delta n^*(\boldsymbol{x}_0) \ge 0$, and thus

$$n^*(\boldsymbol{x}) \ge n^*(\boldsymbol{x}_0) \ge b^- \qquad \text{for all } |\boldsymbol{x}| < A. \tag{3.16}$$

Combining (3.14) and (3.16), we can conclude

$$n^*(\boldsymbol{x}) \ge b^- - \delta \qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^N.$$

The arbitrariness of δ yields the first inequality in (3.11). The proof of the second inequality is perfectly analogous.

Next, we prove the inequalities (3.12) and (3.13). Since n^* satisfies (3.11), we can find two positive constants Θ^-, Θ^+ such that

$$\Theta^- \le \Theta(n^*(\boldsymbol{x})) \le \Theta^+ \qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^N. \tag{3.17}$$

Moreover, using Moser's calculus and (3.11), for any $k > 0$ there exists a constant θ_{k-1}, depending on $|d\Theta/dn|_{C^{k-1}}$ and b^+, such that

$$\left\|D^k(\Theta^* Dn^*) - \Theta^* D^{k+1} n^*\right\| \le \theta_{k-1}\, \|Dn^*\|^2_{H^{k-1}}\,, \tag{3.18}$$

with $\Theta^* = \Theta(n^*)$.

We multiply (3.15) by n^*-b and integrate over the whole space. After integration by parts, we obtain

$$\begin{aligned}\|n^*-b\|^2 &= -\int \Theta^* D(n^*-b)\cdot Dn^* \\ &\le -\frac{1}{2}\int \Theta^* |D(n^*-b)|^2 + \frac{1}{2}\int \Theta^* |Db|^2 ,\end{aligned}$$

which implies

$$\|n^*-b\|^2 + \frac{1}{2}\Theta^- \|D(n^*-b)\|^2 \le \frac{1}{2}\Theta^+ \|Db\|^2 . \tag{3.19}$$

Using (3.19), we find

$$\|Dn^*\|^2 \le 2(\|D(n^*-b)\|^2 + \|Db\|^2) \le c_0 \|Db\|^2 , \tag{3.20}$$

with

$$c_0 = 2\left(1+\frac{c_0'}{\Theta^-}\right), \qquad c_0' = \Theta^+.$$

The proof of (3.12), (3.13) follows from the estimates

$$\left\|D^k(n^*-b)\right\|^2 + \frac{1}{2}\Theta^- \left\|D^{k+1}(n^*-b)\right\|^2 \le \frac{1}{2}c_k' \|Db\|_{H^k}^2 , \tag{3.21}$$

$$\|Dn^*\|_{H^k}^2 \le c_k \|Db\|_{H^k}^2 , \tag{3.22}$$

with

$$c_k' = \frac{1}{\Theta^-}\left(\Theta^{+2} + \theta_{k-1}^2 c_{k-1}^2 \|Db\|_{H^{k-1}}^2\right), \tag{3.23}$$

$$c_k = 2\left(1+\frac{1}{\Theta^-}\sum_{j=0}^{k} c_j'\right). \tag{3.24}$$

To prove these estimates, we proceed by induction. Assuming that (3.21), (3.22) hold for $k-1$, we apply the operator of derivation D^k to (3.15) and multiply scalarly times $D^k(n^*-b)$. Integrating by parts, we obtain

$$\begin{aligned}\left\|D^k(n^*-b)\right\|^2 &= -\int \Theta^* \left|D^{k+1}(n^*-b)\right|^2 \\ &\quad -\int D^{k+1}(n^*-b)\cdot\left[D^k(\Theta^* Dn^*) - \Theta^* D^{k+1}(n^*-b)\right].\end{aligned}$$

Using the Cauchy–Schwarz inequality, recalling (3.18), and applying the induction hypothesis, we have

$$\begin{aligned}&\left\|D^k(n^*-b)\right\|^2 \le -\Theta^- \left\|D^{k+1}(n^*-b)\right\|^2 \\ &\qquad + \left\|D^{k+1}(n^*-b)\right\|\left(\theta_{k-1}c_{k-1}\|Db\|_{H^{k-1}}^2 + \Theta^+\left\|D^{k+1}b\right\|\right) \\ &\le -\Theta^-\left\|D^{k+1}(n^*-b)\right\|^2 \\ &\qquad + \left\|D^{k+1}(n^*-b)\right\|\left(\Theta^{+2}+\theta_{k-1}^2c_{k-1}^2\|Db\|_{H^{k-1}}^2\right)^{\frac{1}{2}}\|Db\|_{H^k} \\ &\le -\frac{1}{2}\Theta^-\left\|D^{k+1}(n^*-b)\right\|^2 + \frac{1}{2\Theta^-}\left(\Theta^{+2}+\theta_{k-1}^2c_{k-1}^2\|Db\|_{H^{k-1}}^2\right)\|Db\|_{H^k}^2 ,\end{aligned}$$

which is equivalent to (3.21).

Finally, using (3.21), we have

$$\begin{aligned}
\|Dn^*\|^2_{H^k} &\le \|Dn^*\|^2_{H^{k-1}} + 2\left(\left\|D^{k+1}(n^*-b)\right\|^2 + \left\|D^{k+1}b\right\|^2\right)\\
&\le c_{k-1}\|Db\|^2_{H^{k-1}} + 2\left(\frac{c'_k}{\Theta^-}\|Db\|^2_{H^k} + \left\|D^{k+1}b\right\|^2\right)\\
&\le 2\left(1+\frac{1}{\Theta^-}\sum_{j=0}^{k-1}c'_j\right)\|Db\|^2_{H^{k-1}} + 2\left(\frac{c'_k}{\Theta^-}\|Db\|^2_{H^k} + \left\|D^{k+1}b\right\|^2\right)\\
&\le 2\left(1+\frac{1}{\Theta^-}\sum_{j=0}^{k}c'_j\right)\|Db\|^2_{H^k},
\end{aligned}$$

which concludes the proof of the lemma. ∎

THEOREM 3.2 (existence and uniqueness of the total equilibrium state). *Let b satisfy* (1.5) *and* (1.6). *Then there exists a unique solution $(n^*,\phi^*)\in C^2$ of* (3.9), *with $n^*-b\in H^s$, $s>\frac{N}{2}+1$.*

Proof. Using a fixed point argument, it is possible to prove the existence of a solution which satisfies the hypothesis of Theorem 3.1. Since the proof is standard, we omit the details. For the uniqueness, let us assume that both (n^*,ϕ^*) and $(n^\star,\phi^\star)$ satisfy (3.9) and n^*-b, $n^\star-b\in H^s$. Then, using Moser's calculus, $\pi(n^*)-\pi(n^\star)\in H^s$ for any smooth function $\pi(n)$. In particular, we consider a function π such that $d\pi/dn=\Theta$. It follows that for any $\delta\ge 0$ there exists $A\ge 0$ such that

$$-\delta \le \pi(n^*(\boldsymbol{x})) - \pi(n^\star(\boldsymbol{x})) \le \delta \qquad \text{for all } |\boldsymbol{x}|\ge A. \tag{3.25}$$

We prove by contradiction that the inequalities in (3.25) hold also for all $|\boldsymbol{x}|\le A$. In fact, let us assume that

$$\pi(n^*(\bar{\boldsymbol{x}})) - \pi(n^\star(\bar{\boldsymbol{x}})) \equiv \max_{|\boldsymbol{x}|\le A}\left[\pi(n^*(\boldsymbol{x})) - \pi(n^\star(\boldsymbol{x}))\right] > \delta. \tag{3.26}$$

Then we have

$$D\left[\pi(n^*(\bar{\boldsymbol{x}})) - \pi(n^\star(\bar{\boldsymbol{x}}))\right] = 0, \qquad \Delta\left[\pi(n^*(\bar{\boldsymbol{x}})) - \pi(n^\star(\bar{\boldsymbol{x}}))\right] \le 0.$$

Subtracting (3.9) for the two solutions, we find

$$D(\pi(n^*)-\pi(n^\star)) = D(\phi^*-\phi^\star), \qquad \Delta(\phi^*-\phi^\star) = n^*-n^\star,$$

which implies

$$n^*(\bar{\boldsymbol{x}}) - n^\star(\bar{\boldsymbol{x}}) = \Delta\left[\pi(n^*(\bar{\boldsymbol{x}})) - \pi(n^\star(\bar{\boldsymbol{x}}))\right] \le 0.$$

Since $d\pi/dn=\Theta\ge 0$, we end up with

$$\pi(n^*(\bar{\boldsymbol{x}})) - \pi(n^\star(\bar{\boldsymbol{x}})) \le 0, \tag{3.27}$$

which contradicts (3.26). Thus, the second inequality in (3.25) holds for all $\boldsymbol{x}$. In a similar way, we can prove that the first inequality in (3.25) is globally valid. In conclusion, we have proved that

$$|\pi(n^*)-\pi(n^\star)| \le \delta \qquad \text{for all } \delta\ge 0. \tag{3.28}$$

It follows that $\pi(n^*)=\pi(n^\star)$, and therefore $n^*=n^\star$. ∎

4. Global existence in time and asymptotic decay of the perturbation. In this section we state and prove the main theorem of the paper, Theorem 4.3, asserting the global existence in time of classical solutions of (1.1)–(1.4), which are perturbations of an equilibrium state, and the decay of these solutions to the unperturbed state. Also, we study the relaxation of the hydrodynamic solutions, that is, their asymptotic behavior as the relaxation time τ tends to zero, assuming that $\sigma\tau$ tends to $\beta > 0$.

First, we show that the hyperbolic–elliptic system (1.1)–(1.4), with the initial data (1.9), (1.10), is equivalent to a symmetric, hyperbolic system, with a nonlocal source term, for the hydrodynamic variables and the electric field $D\phi$.

From (1.1)–(1.4) and (1.7), we find the entropy balance equation,

$$\frac{\partial S}{\partial t} + \sum_{r=1}^{N} \partial_r (S u^r) = Q, \tag{4.1}$$

with

$$Q = \frac{1}{T}\left\{ \frac{n\,|\boldsymbol{u}|^2}{\tau} - \frac{1}{\sigma}\left[\frac{n\,|\boldsymbol{u}|^2}{2} + \left(\frac{\partial e}{\partial T}\right)_n (T - T^*) \right] \right\}.$$

Recalling (1.1), the condition (1.10) on the initial data and (1.12), it is possible to show that the constraint (1.4) can be replaced by the nonlocal evolutionary equation

$$\frac{\partial \boldsymbol{E}}{\partial t} + D\Delta^{-1} D \cdot (n\boldsymbol{u}) = 0. \tag{4.2}$$

As noted in [24], the symbol $\nabla\Delta^{-1}\nabla\cdot$ can be written as a sum of products of Riesz's transforms. Then, by the L^2 boundedness of the Riesz transform [42], for any function $\boldsymbol{w}$ in $(H^s)^N$, $s \geq 0$, we have

$$\left\| \nabla\Delta^{-1}\nabla \cdot \boldsymbol{w} \right\|_{H^s} \leq C_R \left\| \boldsymbol{w} \right\|_{H^s} \tag{4.3}$$

for some positive constant C_R. Equations (1.1), (1.2), (4.1), (4.2) constitute a quasilinear system of partial differential equations with a nonlocal source term. It can be written in the form

$$\frac{\partial U}{\partial t} + \sum_{j=1}^{N} A_j(U)\partial_j U = B(U, \boldsymbol{x}), \tag{4.4}$$

with $U = (n, \boldsymbol{u}, S, \boldsymbol{E})$ and

$$A_j(U) = \begin{pmatrix} u^j & n\boldsymbol{e}_j & 0 & 0 \\ \frac{1}{n}\frac{\partial p}{\partial n}\boldsymbol{e}^j & u^j I & \frac{1}{n}\frac{\partial p}{\partial S}\boldsymbol{e}^j & 0 \\ 0 & S\boldsymbol{e}_j & u^j & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$B(U, \boldsymbol{x}) = \begin{pmatrix} 0 \\ \boldsymbol{E} - \frac{\boldsymbol{u}}{\tau} \\ Q \\ -D\Delta^{-1} D \cdot (n\boldsymbol{u}) \end{pmatrix}.$$

Here,

$$I = (\delta^{ij})_{1\leq i,j\leq N}, \qquad \boldsymbol{e}^j = (\delta^{ij})_{1\leq i\leq N}, \qquad \boldsymbol{e}_j = (\boldsymbol{e}^j)^T.$$

The differential part of system (4.4) is hyperbolic and symmetric in Friedrichs's sense: for any $U \in G \equiv \{U : n > 0\}$ there is a positive definite symmetric matrix $\tilde{A}_0(U)$ smoothly varying with U, and a positive constant c, so that, for all $U \in G$,

1. $cV \cdot V \le (\tilde{A}_0(U)V) \cdot V \le c^{-1}V \cdot V$ for all $V \in G$;
2. $\tilde{A}_j(U) = \tilde{A}_0(U)A_j(U)$ is symmetric.

Specifically, recalling (1.8), the symmetry condition is satisfied with

$$\tilde{A}_0(U) = \begin{pmatrix} \frac{\partial^2 e}{\partial n^2} & 0 & \frac{\partial^2 e}{\partial n \partial S} & 0 \\ 0 & nI & 0 & 0 \\ \frac{\partial^2 e}{\partial n \partial S} & 0 & \frac{\partial^2 e}{\partial S^2} & 0 \\ 0 & 0 & 0 & I \end{pmatrix}.$$

It is well known that a system which is hyperbolic and symmetric in Friedrichs's sense admits locally a unique classical solution in H^s, $s > N/2 + 1$, if the initial data belong to H^s (cf. [35]). Using (4.3), we can easily extend this result to prove the local existence of a solution to (4.4) which is a perturbation of an equilibrium state. We introduce the notation $\delta F(U) = F(U) - F(U^*)$ for any function $F(U)$, where $U^* = (n^*, 0, S^*, \boldsymbol{E}^*)$, with $\boldsymbol{E}^* = D\phi^*$, is the equilibrium state given by (3.9). We can consider the perturbations

$$\delta n = n - n^*, \qquad \delta \boldsymbol{u} = \boldsymbol{u}, \qquad \delta S = S - S^*, \qquad \delta \boldsymbol{E} = \boldsymbol{E} - \boldsymbol{E}^*.$$

We also use the obvious notation

$$U_0(\boldsymbol{x}) = U(\boldsymbol{x}, 0), \qquad \delta U_0(\boldsymbol{x}) = U_0(\boldsymbol{x}) - U^*(\boldsymbol{x}).$$

The perturbation δU satisfies the system

$$\frac{\partial \delta U}{\partial t} + \sum_{j=1}^{N} A_j(U)\partial_j \delta U = \delta B(U, \boldsymbol{x}) - \sum_{j=1}^{N} \delta A_j(U)\partial_j U^*. \tag{4.5}$$

In particular, from (1.4) and $(3.9)_2$, we derive immediately

$$D \cdot \delta \boldsymbol{E} = \delta n. \tag{4.6}$$

Since the differential structure of (4.5) and (4.4) is the same, we can state immediately the following theorem.

THEOREM 4.1 (local existence and uniqueness). *Let $\tau\sigma \neq 0$ and $\delta U_0 \in H^s$, with $s > \frac{N}{2} + 1$. Then there is a time $T > 0$ such that the equations* (4.5) *have a unique classical solution $\delta U(x,t) \in C^1(\mathbb{R}^N \times [0,T))$, with*

$$\delta U \in C^0([0,T], H^s) \cap C^1([0,T], H^{s-1}).$$

This local classical solution can be prolonged locally if the initial data is close enough to the equilibrium solution. The prolongation of a local solution given by Theorem 4.1 resides entirely on the following lemma, whose proof will be given later in subsequent sections.

LEMMA 4.2. *Assuming $0 < \tau < \sigma$ and $Dn^* \in H^s(\mathbb{R}^N)$, let $\delta U = (\delta n, \delta \boldsymbol{u}, \delta S, \delta \boldsymbol{E})$ be the solution of* (4.5) *given by Theorem* 4.1*, and let $\delta \hat{U} = (\delta n, \delta S, \delta \boldsymbol{E})$. There exist positive constants C^*, ϵ, c, and K such that if*

$$\tau\sigma \left\| Dn^* \right\|^2 \le C^*$$

and

$$\frac{1}{\tau^2}\left\|\boldsymbol{u}(\cdot,t)\right\|_{H^s}^2+\left\|\delta\hat{U}(\cdot,t)\right\|_{H^s}^2\leq\epsilon^2\quad\textit{for all } t\in[0,T],$$

then the following a priori estimates hold:

$$\left\|\boldsymbol{u}(\cdot,t)\right\|_{H^s}^2\leq Ke^{-\frac{ct}{\tau}}\left\|\delta U(\cdot,0)\right\|_{H^s}^2\quad\textit{for all } t\in[0,T], \tag{4.7}$$

$$\left\|\delta\hat{U}(\cdot,t)\right\|_{H^s}^2\leq Ke^{-c\tau t}\left\|\delta U(\cdot,0)\right\|_{H^s}^2\quad\textit{for all } t\in[0,T]. \tag{4.8}$$

The constants C^*, ϵ, c, and K depend only on the equation of state $e=e(n,S)$, on the equilibrium density n^*, and on the product of the relaxation times, $\sigma\tau$.

In Lemma 4.2, we keep explicit track of the relaxation time τ, since we are going to study the limit $\tau\to 0$ at the end of this section. If τ is a given constant, the statement of the lemma simplifies. In particular, if we assume $0<\tau\leq 1$, we have

$$e^{-\frac{ct}{\tau}}\leq e^{-c\tau t}, \tag{4.9}$$

and the two estimates (4.7), (4.8) can be comprised into a single estimate.

Now, we are ready to state the global existence theorem announced at the beginning of this section.

THEOREM 4.3 (global existence and asymptotic decay). *Under the same hypothesis as that of Lemma* 4.2, *if $\tau\leq 1$ and $\delta U_0\in H^s$, there exists positive constants C^* and ϵ such that if*

$$\tau\sigma\left\|Dn^*\right\|^2\leq C^*$$

and

$$\frac{1}{\tau^2}\left\|\boldsymbol{u}_0\right\|_{H^s}^2+\left\|\delta\hat{U}_0\right\|_{H^s}^2<\epsilon'^2, \tag{4.10}$$

then the equations (4.5) *have the unique classical solution $\delta U(x,t)\in C^1(\mathbb{R}\times[0,\infty))$. Furthermore,*

$$\delta U\in C^0([0,\infty),H^s)\cap C^1([0,\infty),H^{s-1})$$

and

$$\left\|\delta U(\cdot,t)\right\|_{H^s}^2\leq Ke^{-c\tau t}\left\|\delta U(\cdot,0)\right\|_{H^s}^2, \tag{4.11}$$

where c and K are positive constants given by Lemma 4.2.

Proof. Theorem 4.1 yields immediately the local H^s existence of the unique classical solution to the initial value problem for (4.4). We introduce the constant $\epsilon_1=\epsilon/\sqrt{2K}<\epsilon$, where ϵ is given by Lemma 4.2, and assume that the inequality (4.10) is satisfied with $\epsilon'=\epsilon_1$. By continuity, we can determine a time $T_1>0$ such that

$$\frac{1}{\tau^2}\left\|\boldsymbol{u}(\cdot,t)\right\|_{H^s}^2+\left\|\delta\hat{U}(\cdot,t)\right\|_{H^s}^2\leq\epsilon_1^2\quad\text{for all } t\in[0,T_1]. \tag{4.12}$$

The inequality (4.12) is still satisfied if we choose $\epsilon'=\min\{\epsilon_1,\epsilon_1 cT_1/2\}$ in (4.10). We will prove that this choice of ϵ' is sufficient to ensure

$$\frac{1}{\tau^2}\left\|\boldsymbol{u}(\cdot,t)\right\|_{H^s}^2+\left\|\delta\hat{U}(\cdot,t)\right\|_{H^s}^2<\epsilon^2\quad\text{for all } t\geq 0. \tag{4.13}$$

We proceed by contradiction and assume that there exists $\bar{T} > T_1$ such that (4.13) is satisfied for all $0 \le t < \bar{T}$, and

$$\frac{1}{\tau^2}\left\|\boldsymbol{u}(\cdot,\bar{T})\right\|_{H^s}^2 + \left\|\delta\hat{U}(\cdot,\bar{T})\right\|_{H^s}^2 = \epsilon^2. \tag{4.14}$$

Using Lemma 4.2, we get

$$\frac{1}{\tau^2}\left\|\boldsymbol{u}(\cdot,\bar{T})\right\|_{H^s}^2 \le K\epsilon'^2 \sup_{0<\tau\le 1}\left\{\tau^{-2}e^{-\frac{cT_1}{\tau}}\right\} = \begin{cases} K\epsilon'^2 e^{-cT_1} & \text{if} \quad \dfrac{cT_1}{2} \ge 1, \\ K\epsilon'^2\left(\dfrac{2}{ecT_1}\right)^2 & \text{if} \quad \dfrac{cT_1}{2} < 1. \end{cases}$$

If $cT_1/2 \ge 1$, we have $\epsilon' = \epsilon_1$ and conclude that

$$\frac{1}{\tau^2}\left\|\boldsymbol{u}(\cdot,\bar{T})\right\|_{H^s}^2 \le K\epsilon_1^2 e^{-cT_1} \le \frac{\epsilon^2}{2e^2} < \frac{\epsilon^2}{2}.$$

If $cT_1/2 < 1$, we have $\epsilon' = \epsilon_1 cT_1/2$, and we find

$$\frac{1}{\tau^2}\left\|\boldsymbol{u}(\cdot,\bar{T})\right\|_{H^s}^2 \le K\left(\frac{\epsilon_1 cT_1}{2}\right)^2\left(\frac{2}{ecT_1}\right)^2 = \frac{\epsilon^2}{2e^2} < \frac{\epsilon^2}{2}.$$

Using Lemma 4.2, it is simple to see that

$$\left\|\delta\hat{U}(\cdot,\bar{T})\right\|_{H^s}^2 \le K\epsilon'^2 \sup_{0<\tau\le 1}\left\{e^{-\tau cT_1}\right\} = K\epsilon'^2 < \frac{\epsilon^2}{2}.$$

It follows that

$$\frac{1}{\tau^2}\left\|\boldsymbol{u}(\cdot,\bar{T})\right\|_{H^s}^2 + \left\|\delta\hat{U}(\cdot,\bar{T})\right\|_{H^s}^2 < \epsilon^2, \tag{4.15}$$

which contradicts (4.14). Therefore, (4.13) holds true. Using Lemma 4.2, the local solution can be prolonged for all times and the estimates (4.7) and (4.8) hold globally. The final estimate (4.11) is obtained by adding (4.7) and (4.8) and recalling (4.9). ∎

We remark that the constants in the estimate (4.11) are independent on τ. We can use this fact to study the behavior of the solution as τ tends to zero, assuming that $\tau\sigma$ tends to a certain positive constant β. We define the rescaled variables

$$U'(\boldsymbol{x},t') \equiv (n',\boldsymbol{u}',S',E')(\boldsymbol{x},t') = \left(n,\frac{1}{\tau}\boldsymbol{u},S,E\right)\left(\boldsymbol{x},\frac{1}{\tau}t'\right)$$

and the initial data

$$U'(\boldsymbol{x},0) = U_0'(\boldsymbol{x}) \equiv (n_0,\boldsymbol{u}_0',S_0,E_0)(\boldsymbol{x}).$$

In other words, we are assuming that $\boldsymbol{u}(\boldsymbol{x},0) = \tau\boldsymbol{u}_0'(\boldsymbol{x})$. The rescaled variables satisfy the system

$$A^\tau\frac{\partial U'}{\partial t'} + \sum_{j=1}^N A_j'(U')\partial_j U' = B'(U',\boldsymbol{x}), \tag{4.16}$$

with $A^\tau = \mathrm{diag}(1, \tau^2 I, 1, I)$ and

$$A'_j(U') = \begin{pmatrix} u'^j & n'\boldsymbol{e}_j & 0 & 0 \\ \frac{1}{n'}\frac{\partial p'}{\partial n'}\boldsymbol{e}^j & \tau^2 u'^j I & \frac{1}{n'}\frac{\partial p'}{\partial S'}\boldsymbol{e}^j & 0 \\ 0 & S'\boldsymbol{e}_j & u'^j & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$B'(U', \boldsymbol{x}) = \begin{pmatrix} 0 \\ \boldsymbol{E}' - \boldsymbol{u}' \\ \frac{n'}{T'}\left\{\left(1 - \frac{\tau}{2\sigma}\right)|\boldsymbol{u}'|^2 - \frac{1}{2\tau\sigma}\left(\frac{\partial e'}{\partial T'}\right)_{n'}(T' - T^*)\right\} \\ -D\Delta^{-1}D \cdot (n'\boldsymbol{u}') \end{pmatrix}.$$

In terms of the rescaled variables, the estimate (4.11) becomes

$$\tau^2 \left\|\boldsymbol{u}'(\cdot, t')\right\|^2_{H^s} + \left\|\delta\hat{U}'(\cdot, t')\right\|^2_{H^s} \leq Ke^{-ct'}\left(\tau^2 \left\|\boldsymbol{u}'(\cdot, 0)\right\|^2_{H^s} + \left\|\delta\hat{U}'(\cdot, 0)\right\|^2_{H^s}\right). \tag{4.17}$$

Since $\sigma\tau$ is strictly positive and bounded, this estimate is uniformly valid as τ tends to zero, and the limit function is a solution of (4.16) with $\tau = 0$ and $\sigma\tau = \beta$. Then it is immediate to derive the following singular limit result.

THEOREM 4.4 (relaxation). *Let as assume that $\tau\sigma$ tends to $\beta > 0$ as τ tends to 0. For any fixed $\tau > 0$, let $U = (n^\tau, \boldsymbol{u}^\tau, S^\tau, \boldsymbol{E}^\tau)(\boldsymbol{x}, t)$ be a global solution of (4.4), satisfying (4.7), and let $\hat{U}^\tau = (n^\tau, S^\tau, \boldsymbol{E}^\tau)(\boldsymbol{x}, t)$. Then there exists some function $\hat{U}^0 = (n^0, S^0, \boldsymbol{E}^0)$ which is a smooth solution of (4.16), with $\tau = 0$, and such that, as τ tends to zero,*

$$(\hat{U}^\tau - \hat{U}^*)\left(\boldsymbol{x}, \frac{1}{\tau}t'\right) \to (\hat{U}^0 - \hat{U}^*)(\boldsymbol{x}, t') \quad \text{in } C([0, \infty); H^s).$$

Furthermore,

$$\left\|(\hat{U}^0 - \hat{U}^*)(\cdot, t')\right\|^2_{H^s} \leq K'e^{-c't'}\left\|(\hat{U}^0 - \hat{U}^*)(\cdot, 0)\right\|^2_{H^s}, \tag{4.18}$$

where c' and K' are positive constants.

5. Positive definiteness of some functionals of Liapunov type. This section deals with the positive definiteness of some functionals of Liapunov type which will be used in the subsequent section. The main results are summarized in Lemma 5.3.

We fix an integer $s > N/2 + 1$ and introduce the energy densities

$$\mathcal{H}_0 = \frac{n}{2}|\boldsymbol{u}|^2 - \frac{\lambda_0 n}{\sigma n^*}\boldsymbol{u} \cdot \delta\boldsymbol{E} + \frac{1}{2}|\delta\boldsymbol{E}|^2 + e - e^* - \left(\frac{\partial e}{\partial n}\right)^* \delta n - \left(\frac{\partial e}{\partial S}\right)^* \delta S, \tag{5.1}$$

$$\mathcal{H}_i = \frac{n}{2}\left|D^i\boldsymbol{u}\right|^2 - \frac{\lambda_i n}{\sigma n^*}D^i\boldsymbol{u} \cdot D^i\delta\boldsymbol{E} + \frac{1}{2}\left|D^i\delta\boldsymbol{E}\right|^2 + \frac{1}{2}\frac{\partial^2 e}{\partial n^2}\left|D^i\delta n\right|^2 + \frac{\partial^2 e}{\partial n \partial S}D^i\delta n \cdot D^i\delta S + \frac{1}{2}\frac{\partial^2 e}{\partial S^2}\left|D^i\delta S\right|^2, \quad 1 \leq i \leq s. \tag{5.2}$$

Here, λ_0 and λ_i are positive constants. With an appropriate choice of these constants, $\mathcal{H}_0$ and $\mathcal{H}_i$ are positive semidefinite, as asserted by the following lemma.

LEMMA 5.1. *If $0 < \tau < \sigma$, $\sigma\tau \geq \beta > 0$, $\lambda_i < \min\{1/2, \beta \inf n^*\}$, and $|\delta n| < \inf |n^*|$, then there exist positive constants $k_{\mathcal{H}}$, $K_{\mathcal{H}}$ (depending only on β, λ_i, $e(n,S)$ and n^*) such that $\mathcal{H}_i$ satisfies*

$$k_{\mathcal{H}} \left|D^i \delta U\right|^2 \leq \mathcal{H}_i \leq K_{\mathcal{H}} \left|D^i \delta U\right|^2.$$

Proof. We can decompose the energy densities as

$$\mathcal{H}_i = g_i(D^i\delta n, D^i \delta S) + h_i(D^i \boldsymbol{u}, D^i \delta \boldsymbol{E}), \quad i \geq 0, \tag{5.3}$$

where, for any $X^i, Y^i \in \mathbb{R}^{N^i} \times \mathbb{R}^{N^i}$ and $\boldsymbol{X}^i, \boldsymbol{Y}^i \in (\mathbb{R}^{N^i})^N \times (\mathbb{R}^{N^i})^N$, the quadratic forms g_i and h_i are defined by

$$g_i(X^i, Y^i) = \begin{cases} e(n^* + X^i, S^* + Y^i) - e^* - \left(\dfrac{\partial e}{\partial n}\right)^* X^i - \left(\dfrac{\partial e}{\partial S}\right)^* Y^i, & i = 0, \\ \dfrac{1}{2}\dfrac{\partial^2 e}{\partial n^2}\left|X^i\right|^2 + \dfrac{\partial^2 e}{\partial n \partial S} X^i \cdot Y^i + \dfrac{1}{2}\dfrac{\partial^2 e}{\partial S^2}\left|Y^i\right|^2, & i > 0, \end{cases}$$

$$h_i(\boldsymbol{X}^i, \boldsymbol{Y}^i) = \frac{n}{2}\left|\boldsymbol{X}^i\right|^2 - \frac{\lambda_i n}{\sigma n^*} \boldsymbol{X}^i \cdot \boldsymbol{Y}^i + \frac{1}{2}\left|\boldsymbol{Y}^i\right|^2, \quad i \geq 0.$$

The positive definiteness of g_i follows from the convexity of the internal energy with respect to n and S. We consider the following inequality:

$$\tau \left|\boldsymbol{X}^i \cdot \boldsymbol{Y}^i\right| \leq n^* \tau^2 \left|\boldsymbol{X}^i\right|^2 + \frac{1}{4n^*}\left|\boldsymbol{Y}^i\right|^2.$$

Then it is possible to derive

$$\begin{aligned} h_i(\boldsymbol{X}^i, \boldsymbol{Y}^i) &\geq \frac{n}{2}\left(1 - \frac{2\lambda_i \tau}{\sigma}\right)\left|\boldsymbol{X}^i\right|^2 + \frac{1}{2}\left(1 - \frac{\lambda_i}{\sigma\tau n^*} + \left(1 - \frac{\delta n}{n^*}\right)\frac{\lambda_i}{2\sigma\tau n^*}\right)\left|\boldsymbol{Y}^i\right|^2 \\ &\geq \frac{n}{2}\left(1 - 2\lambda_i\right)\left|\boldsymbol{X}^i\right|^2 + \frac{1}{2}\left(1 - \frac{\lambda_i}{\sigma\tau n^*}\right)\left|\boldsymbol{Y}^i\right|^2. \end{aligned}$$

Here, we have used $\tau < \sigma$, $|\delta n| < \inf n^*$. Recalling that $\sigma\tau \geq \beta$ and $\lambda_i < \min\{1/2, \beta \inf n^*\}$, the thesis of the lemma follows immediately. ∎

Now, we introduce the functionals

$$\begin{aligned} \mathcal{D}_0 = \frac{n^*}{2\tau}(1 - 2\bar{C}_R\lambda_0)\left|\boldsymbol{u}\right|^2 + \frac{1}{\sigma}\Bigg\{ &\frac{\lambda_0}{2}|\delta \boldsymbol{E}|^2 + \frac{\lambda_0}{(n^*)^2}\left(\frac{\partial p}{\partial S}\right)^* [\delta n D S^* \\ &- \delta S D n^*] \cdot \delta \boldsymbol{E} + \left[\left(\frac{\partial S}{\partial T}\right)_n \left(\frac{\partial T}{\partial n}\right)^2 + \frac{\lambda_0}{n}\frac{\partial p}{\partial n}\right]^* \delta n^2 \\ &+ \left(2\frac{\partial T}{\partial n} + \frac{\lambda_0}{n}\frac{\partial p}{\partial S}\right)^* \delta n \delta S + \left(\frac{\partial T}{\partial S}\right)^* \delta S^2 \Bigg\}, \end{aligned} \tag{5.4}$$

$$\begin{aligned} \mathcal{D}_i = \frac{n^*}{2\tau}(1 - 2\bar{C}\lambda_i)\left|D^i \boldsymbol{u}\right|^2 + \frac{1}{\sigma}\Bigg\{ &\frac{\lambda_i}{2}\left|D^i \delta \boldsymbol{E}\right|^2 \\ &+ \left[\left(\frac{\partial S}{\partial T}\right)_n \left(\frac{\partial T}{\partial n}\right)^2 + \frac{\lambda_0}{n}\frac{\partial p}{\partial n}\right]^* \left|D^i \delta n\right|^2 \\ &+ \left(2\frac{\partial T}{\partial n} + \frac{\lambda_i}{n}\frac{\partial p}{\partial S}\right)^* D^i \delta n \cdot D^i \delta S + \left(\frac{\partial T}{\partial S}\right)^* \left|D^i \delta S\right|^2 \Bigg\}, \quad i \geq 1, \end{aligned} \tag{5.5}$$

with $\bar{C} = \sup n^* / \inf n^*$, $\bar{C}_R = \bar{C} C_R$. The constant C_R is defined by (4.3), and we have $\bar{C}_R \geq \bar{C} \geq 1$. We prove that the quadratic forms $\mathcal{D}_0$ and $\mathcal{D}_i$ are positive definite with an appropriate choice of λ_0 and λ_i.

LEMMA 5.2. *Let $0 < \tau < \sigma$, $\lambda_i < 1/2\bar{C}_R$, and*

$$0 < \lambda_0 < \inf \left\{ \frac{H(e)}{\left(\frac{1}{2n}\frac{\partial p}{\partial S}\right)^2 \left(1 + 2\Theta \frac{|Dn|^2}{n^2}\right)} \right\}^*, \qquad 0 < \lambda_i < \inf \left\{ \frac{H(e)}{\left(\frac{1}{2n}\frac{\partial p}{\partial S}\right)^2} \right\}^*,$$

where $H(e)$ and Θ are defined by (3.8) *and* (3.10)*, respectively. Then there exist positive constants $k_{\mathcal{D}}$, $K_{\mathcal{D}}$ (depending only on λ_i, $e(n,S)$, and n^*) such that $\mathcal{D}_i$ ($i \geq 0$) satisfies*

$$k_{\mathcal{D}} \left(\frac{1}{\tau} \left| D^i \boldsymbol{u} \right|^2 + \frac{1}{\sigma} \left| D^i \delta \hat{U} \right|^2 \right) \leq \mathcal{D}_i \leq K_{\mathcal{D}} \left(\frac{1}{\tau} \left| D^i \boldsymbol{u} \right|^2 + \frac{1}{\sigma} \left| D^i \delta \hat{U} \right|^2 \right). \tag{5.6}$$

Proof. We can decompose the quadratic form $\mathcal{D}_0$ as

$$\mathcal{D}_0(\boldsymbol{u}, \delta \boldsymbol{E}, \delta n, \delta S) = \frac{n^*}{2\tau}(1 - 2\bar{C}_R \lambda_0) + \frac{1}{\sigma} \tilde{\mathcal{D}}(\delta \hat{U}).$$

The coefficient of $|\boldsymbol{u}|^2$ is positive, since $\lambda_0 < 1/2\bar{C}_R$. The matrix associated to the quadratic form $\tilde{\mathcal{D}}$ with respect to $\delta \hat{U}$ is

$$A = \begin{pmatrix} \frac{\lambda_0}{2} I & \frac{\lambda_0}{2n^2}\frac{\partial p}{\partial S} DS & -\frac{\lambda_0}{2n^2}\frac{\partial p}{\partial S} Dn \\ \frac{\lambda_0}{2n^2}\frac{\partial p}{\partial S}(DS)^T & \left(\frac{\partial S}{\partial T}\right)_n \left(\frac{\partial T}{\partial n}\right)^2 + \frac{\lambda_0}{n}\frac{\partial p}{\partial n} & \frac{\partial T}{\partial n} + \frac{\lambda_0}{2n}\frac{\partial p}{\partial S} \\ -\frac{\lambda_0}{2n^2}\frac{\partial p}{\partial S}(Dn)^T & \frac{\partial T}{\partial n} + \frac{\lambda_0}{2n}\frac{\partial p}{\partial S} & \frac{\partial T}{\partial S} \end{pmatrix}^*.$$

The matrix A is positive definite if its determinant is positive together with the determinants of the following minors:

$$\left(\frac{\partial T}{\partial S} \right)^*, \qquad \begin{pmatrix} \left(\frac{\partial S}{\partial T}\right)_n \left(\frac{\partial T}{\partial n}\right)^2 + \frac{\lambda_0}{n}\frac{\partial p}{\partial n} & \frac{\partial T}{\partial n} + \frac{\lambda_0}{2n}\frac{\partial p}{\partial S} \\ \frac{\partial T}{\partial n} + \frac{\lambda_0}{2n}\frac{\partial p}{\partial S} & \frac{\partial T}{\partial S} \end{pmatrix}^*.$$

Explicitly, these conditions amount to

$$\left(\frac{\partial T}{\partial S}\right)^* > 0,$$

$$\lambda_0 \left(H(e) - \lambda_0 \left(\frac{1}{2n}\frac{\partial p}{\partial S} \right)^2 \right)^* > 0,$$

$$\frac{\lambda_0^2}{2} \left(H(e) - \lambda_0 \left(\frac{1}{2n}\frac{\partial p}{\partial S} \right)^2 \left(1 + \frac{2Dn \cdot Dp}{n^3} \right) \right)^* > 0.$$

Recalling (3.4) and (3.9), we see that

$$\left(1 + \frac{2Dn \cdot Dp}{n^3} \right)^* = \left(1 + \frac{2Dn \cdot D\phi}{n^2} \right)^* = \left(1 + 2\Theta \frac{|Dn|^2}{n^2} \right)^* \geq 1.$$

Then the previous inequalities are satisfied altogether if

$$0 < \lambda_0 < \inf \left\{ \frac{H(e)}{\left(\frac{1}{2n}\frac{\partial p}{\partial S}\right)^2 \left(1 + 2\Theta \frac{|Dn|^2}{n^2}\right)} \right\}^*,$$

which holds by hypothesis. We can conclude that $\mathcal{D}_0$ is positive definite. Then we can determine appropriate constants $k_{\mathcal{D}}$, $K_{\mathcal{D}}$, independent on τ and σ, such that (5.6) holds for $i = 0$. In the same way, after replacing λ_0 with λ_i and putting $Dn^* = 0$ in the matrix A, it is possible to prove the positive definiteness of $\mathcal{D}_i$, $i \geq 1$. ∎

The following lemma is a useful combination of Lemmas 5.1 and 5.2.

LEMMA 5.3. *If* $0 < \tau < \sigma$, $\sigma\tau \geq \beta > 0$, $|\delta n| < \inf |n^*|$, *and*

$$\lambda_i < \min\left\{ \frac{1}{2\bar{C}_R}, \beta \inf n^* \right\}, \quad \lambda_i < \inf \left\{ \frac{H(e)}{\left(\frac{1}{2n}\frac{\partial p}{\partial S}\right)^2 \left(1 + 2\Theta \frac{|Dn|^2}{n^2}\right)} \right\}^*, \quad i \geq 0, \tag{5.7}$$

then there exist some positive constants $k_{\mathcal{H}}$, $K_{\mathcal{H}}$, $k_{\mathcal{D}}$, $K_{\mathcal{D}}$ *(depending only on* λ_i, $e(n, S)$, n^*, *and* β*) such that the functions* $\mathcal{H}_i$ *and* $\mathcal{D}_i$ *(*$i \geq 0$*) satisfy*

$$k_{\mathcal{H}} \left\| D^i \delta U \right\|^2 \leq \int \mathcal{H}_i \, d\boldsymbol{x} \leq K_{\mathcal{H}} \left\| D^i \delta U \right\|^2, \tag{5.8}$$

$$\begin{aligned} k_{\mathcal{D}} \left(\frac{1}{\tau} \left\| D^i \boldsymbol{u} \right\|^2 + \frac{1}{\sigma} \left\| D^i \delta \hat{U} \right\|^2 \right) \\ \leq \int \mathcal{D}_i \, d\boldsymbol{x} \leq K_{\mathcal{D}} \left(\frac{1}{\tau} \left\| D^i \boldsymbol{u} \right\|^2 + \frac{1}{\sigma} \left\| D^i \delta \hat{U} \right\|^2 \right). \end{aligned} \tag{5.9}$$

6. A priori estimates. This section is entirely devoted to the proof of Lemma 4.2. For some fixed positive number T, we assume that a solution δU of (4.5) exists and $\delta U(\boldsymbol{x}, t) \in H^s(\mathbb{R}^N)$ for an integer $s > N/2 + 1$ for all $t \in (0, T)$. We introduce the vector

$$\delta U_\tau = \left(\delta n, \frac{1}{\tau} \boldsymbol{u}, \delta S, \delta \boldsymbol{E} \right) \tag{6.1}$$

and define

$$\mathcal{U}(T) = \sup_{0 \leq t \leq T} \left\| \delta U_\tau(\cdot, t) \right\|_{H^s}. \tag{6.2}$$

Using the standard Sobolev inequalities, there exists a positive constant $C_{\mathcal{U}}$ such that

$$\sup_{0 \leq t \leq T} \left\| \delta U_\tau(\cdot, t) \right\|_{C^r} \leq C_{\mathcal{U}} \mathcal{U}(T), \qquad r < s - \frac{N}{2}. \tag{6.3}$$

For any function $f \in H^s$, we introduce the norm

$$\|f\|_{\eta,s}^2 = \sum_{i=0}^{s} \eta_i \|f\|_{H^s}^2,$$

where $\eta_0 = 1$, and η_i, $i = 1, 2, \ldots, s$, are positive constants to be determined. Also, we introduce the energy

$$\mathcal{W} = \int \left(\sum_{i=0}^{s} \eta_i \mathcal{H}_i \right) d\boldsymbol{x}, \tag{6.4}$$

where the energy densities $\mathcal{H}_i$ are defined by (5.1), (5.2), and the constants λ_i, $i \geq 0$, are chosen according to the condition (5.7). Then, using Lemma 5.3, it is immediate to see that $\mathcal{W}$ is equivalent to the H^s-norm of δU. Lemma 4.2 follows from the subsequent energy estimate.

LEMMA 6.1. *If $0 < \tau < \sigma$ and $Dn^* \in H^s(\mathbb{R}^N)$, there exist positive constants C^* (depending only on the function $e(n,S)$ and on $b^{\pm}$), ϵ, η_i, $i = 1, 2, \dots, s$, such that if*

$$\sigma\tau \left\| Dn^* \right\|^2 \leq C^* \tag{6.5}$$

and the solution is so small that $\mathcal{U}(T) \leq \epsilon$, then the following a priori estimate holds for $t \in [0,T]$:

$$\left\| \boldsymbol{u}(\cdot,t) \right\|_{\eta,s} \leq \frac{\mathcal{W}(0)}{k_{\mathcal{H}}} e^{-\frac{ct}{\tau}}, \qquad \left\| \delta\hat{U}(\cdot,t) \right\|_{\eta,s} \leq \frac{\mathcal{W}(0)}{k_{\mathcal{H}}} e^{-c\tau t}, \tag{6.6}$$

where the constant c depends only on the equation of state, on the equilibrium state, and on the product $\tau\sigma$, and $k_{\mathcal{H}}$ is given by (5.8).

Proof. To begin with, we consider the function $\mathcal{H}_0$ defined by (5.1). Deriving with respect to t, using (1.1)–(1.3), (4.1), and (4.2), and integrating on the whole space, we obtain

$$\begin{aligned}
\frac{d}{dt}\int \mathcal{H}_0 \, d\boldsymbol{x} &= \int \left\{ \frac{\partial}{\partial t}\left(\frac{n}{2}|\boldsymbol{u}|^2 + e \right) - \left(\frac{\partial e}{\partial n} \right)^* \frac{\partial n}{\partial t} - \left(\frac{\partial e}{\partial S} \right)^* \frac{\partial S}{\partial t} \right. \\
&\quad \left. + \delta\boldsymbol{E} \cdot \frac{\partial \delta\boldsymbol{E}}{\partial t} - \frac{\lambda_0}{\sigma n^*} \left[n\boldsymbol{u} \cdot \frac{\partial \delta\boldsymbol{E}}{\partial t} + \frac{\partial}{\partial t}(n\boldsymbol{u}) \cdot \delta\boldsymbol{E} \right] \right\} d\boldsymbol{x} \\
&= \int \left\{ n\boldsymbol{u} \cdot \boldsymbol{E} + TQ - \frac{1}{\tau} n \left| \boldsymbol{u} \right|^2 + \phi^* D \cdot (n\boldsymbol{u}) \right. \\
&\quad - T^*Q - \delta\boldsymbol{E} \cdot D\Delta^{-1} D \cdot (n\boldsymbol{u}) + \frac{\lambda_0}{\sigma n^*} n\boldsymbol{u} \cdot D\Delta^{-1} D \cdot (n\boldsymbol{u}) \\
&\quad \left. + \frac{\lambda_0}{\sigma n^*} \left[D \cdot (n\boldsymbol{u} \otimes \boldsymbol{u}) + D\,(p - p^*) - n\delta\boldsymbol{E} - \delta n \boldsymbol{E}^* + \frac{n\boldsymbol{u}}{\tau} \right] \cdot \delta\boldsymbol{E} \right\} d\boldsymbol{x}.
\end{aligned} \tag{6.7}$$

Integrating by parts and observing that $\delta\boldsymbol{E}$ is a gradient, it is immediate to see that

$$\int \left\{ n\boldsymbol{u} \cdot \boldsymbol{E} + \phi^* D \cdot (n\boldsymbol{u}) - \delta\boldsymbol{E} \cdot D\Delta^{-1} D \cdot (n\boldsymbol{u}) \right\} d\boldsymbol{x} = 0.$$

Moreover, using the estimate (4.3), we get

$$\int \left\{ \frac{\lambda_0}{\sigma n^*} n\boldsymbol{u} \cdot D\Delta^{-1} D \cdot (n\boldsymbol{u}) \right\} d\boldsymbol{x} \leq \frac{\lambda_0 C_R}{\sigma \inf n^*} \left\| n\boldsymbol{u} \right\|^2.$$

Also, by Schwarz's inequality, and recalling the first condition in (5.7), we have

$$\begin{aligned}
&\int \frac{\lambda_0 n}{\sigma n^*} \left(-\delta\boldsymbol{E} + \frac{\boldsymbol{u}}{\tau} \right) \cdot \delta\boldsymbol{E} \, d\boldsymbol{x} \\
&\leq \int \left\{ \frac{n}{2\tau} \left| \boldsymbol{u} \right|^2 - \frac{\lambda_0 n}{2\sigma n^*} \left(2 - \frac{\lambda_0}{\sigma\tau n^*} \right) \left| \delta\boldsymbol{E} \right|^2 \right\} d\boldsymbol{x} \\
&\leq \int \left\{ \frac{n}{2\tau} \left| \boldsymbol{u} \right|^2 - \frac{\lambda_0 n}{2\sigma n^*} \left| \delta\boldsymbol{E} \right|^2 \right\} d\boldsymbol{x}.
\end{aligned} \tag{6.8}$$

Using these results in (6.7), we find

$$\frac{d}{dt}\int \mathcal{H}_0\,d\boldsymbol{x} \le -\int\left\{\frac{n}{2\tau}\left(1-\lambda_0\frac{2\tau C_R n}{\sigma\inf n^*}\right)|\boldsymbol{u}|^2 - Q\delta T + \frac{\lambda_0 n}{2\sigma n^*}|\delta\boldsymbol{E}|^2\right\}d\boldsymbol{x} + \int\left\{\frac{\lambda_0}{\sigma n^*}\delta\boldsymbol{E}\cdot\left[D\cdot(n\boldsymbol{u}\otimes\boldsymbol{u}) + D\,(p-p^*) - \delta n\boldsymbol{E}^*\right]\right\}d\boldsymbol{x} \equiv I_{01}+I_{02}, \tag{6.9}$$

where

$$Q\delta T = \left(1-\frac{\tau}{2\sigma}\right)\frac{n}{T}\delta T\frac{|\boldsymbol{u}|^2}{\tau} - \frac{1}{\sigma T}\left(\frac{\partial e}{\partial T}\right)_n(\delta T)^2.$$

Taylor expanding $T(n,S)$ and $e(n,S)$ around (n^*,S^*), and recalling that $\tau<\sigma$, we can estimate

$$I_{01} \le C\mathcal{U}\left(\frac{1}{\tau}\|\boldsymbol{u}\|^2 + \frac{1}{\sigma}\left\|\delta\hat{U}\right\|^2\right) - \int\left\{\frac{n^*}{2\tau}(1-2\bar{C}_R\lambda_0)|\boldsymbol{u}|^2 + \frac{1}{\sigma T^*}\left(\frac{\partial e}{\partial T}\right)_n^*\left[\left(\frac{\partial T}{\partial n}\right)^*\delta n + \left(\frac{\partial T}{\partial S}\right)^*\delta S\right]^2 + \frac{\lambda_0}{2\sigma}|\delta\boldsymbol{E}|^2\right\}d\boldsymbol{x}, \tag{6.10}$$

where $\bar{C}_R = C_R\sup n^*/\inf n^* \ge 1$.

To estimate I_{02}, we integrate by parts and use (4.6). We obtain

$$\begin{aligned} I_{02} &= -\int\frac{\lambda_0\tau n}{\sigma n^*}\left[(\boldsymbol{u}\cdot D)\delta\boldsymbol{E} - \frac{Dn^*}{n^*}(\boldsymbol{u}\cdot\delta\boldsymbol{E})\right]\cdot\frac{\boldsymbol{u}}{\tau}\,d\boldsymbol{x} \\ &\quad -\int\frac{\lambda_0}{\sigma}\left[(p-p^*)\left(\frac{\delta n}{n^*} - \frac{1}{(n^*)^2}\delta\boldsymbol{E}\cdot Dn^*\right) + \frac{\delta n}{(n^*)^2}\delta\boldsymbol{E}\cdot Dp^*\right]d\boldsymbol{x} \\ &\le C\mathcal{U}\left(\frac{1}{\tau}\|\boldsymbol{u}\|^2 + \frac{1}{\sigma}\left\|\delta\hat{U}\right\|^2\right) - \int\frac{\lambda_0\delta n}{\sigma n^*}\left[\left(\frac{\partial p}{\partial n}\right)^*\delta n + \left(\frac{\partial p}{\partial S}\right)^*\delta S\right]d\boldsymbol{x} \\ &\quad -\int\frac{\lambda_0}{\sigma}\left(\frac{1}{n^2}\frac{\partial p}{\partial S}\right)^*(\delta n DS^* - \delta S Dn^*)\cdot\delta\boldsymbol{E}\,d\boldsymbol{x}. \end{aligned} \tag{6.11}$$

Using (6.10) and (6.11) in (6.7), we find

$$\frac{d}{dt}\int\mathcal{H}_0\,d\boldsymbol{x} \le -\int\mathcal{D}_0\,d\boldsymbol{x} + \tau C_0'\mathcal{U}\left\|\delta U_\tau\right\|^2, \tag{6.12}$$

where C_0' is a positive constant and $\mathcal{D}_0$ is the quadratic form defined by (5.4). In conclusion, using (5.9), at order zero we can estimate the energy as

$$\frac{d}{dt}\left(\int\mathcal{H}_0\,d\boldsymbol{x}\right) \le -k_{\mathcal{D}}\left(\frac{1}{\tau}\|\boldsymbol{u}\|^2 + \frac{1}{\sigma}\left\|\delta\hat{U}\right\|^2\right) + \tau C_0'\mathcal{U}\left\|\delta U_\tau\right\|^2. \tag{6.13}$$

Next, we consider the function $\mathcal{H}_k$, defined by (5.2), with $1\le k\le s$. We can write

$$\begin{aligned} \frac{d}{dt}\int\mathcal{H}_k\,d\boldsymbol{x} &= \int\frac{\partial}{\partial t}\left\{\frac{1}{2}n\left|D^k\boldsymbol{u}\right|^2 + \frac{1}{2}\left|D^k\delta\boldsymbol{E}\right|^2\right\}d\boldsymbol{x} \\ &\quad +\int\frac{\partial}{\partial t}\left\{\frac{1}{2}\frac{\partial^2 e}{\partial n^2}\left|D^k\delta n\right|^2 + \frac{\partial^2 e}{\partial n\partial S}D^k\delta n\cdot D^k\delta S + \frac{1}{2}\frac{\partial^2 e}{\partial S^2}\left|D^k\delta S\right|^2\right\}d\boldsymbol{x} \\ &\quad -\int\frac{\partial}{\partial t}\left\{\frac{\lambda_k n}{\sigma n^*}D^k\boldsymbol{u}\cdot D^k\delta\boldsymbol{E}\right\}d\boldsymbol{x} = I_1+I_2+I_3. \end{aligned} \tag{6.14}$$

Using the perturbation equation (4.5), we find

$$
\begin{aligned}
I_1 &= -\int \left\{ D\cdot(n\boldsymbol{u})\frac{1}{2}\left|D^k\boldsymbol{u}\right|^2 + D^k\delta\boldsymbol{E}\cdot D^k(n\boldsymbol{u}) \right.\\
&\left. + nD^k\boldsymbol{u}\cdot D^k\left[(\boldsymbol{u}\cdot D)\boldsymbol{u} + \frac{1}{n}Dp - \frac{1}{n^*}Dp^* - \delta\boldsymbol{E} + \frac{\boldsymbol{u}}{\tau}\right]\right\} d\boldsymbol{x}\\
&= -\int nD^k\boldsymbol{u}\cdot\left[D^k((\boldsymbol{u}\cdot D)\boldsymbol{u}) - (\boldsymbol{u}\cdot D)(D^k\boldsymbol{u})\right] d\boldsymbol{x}\\
&- \int D^k\delta\boldsymbol{E}\cdot\left[D^k(n\boldsymbol{u}) - nD^k\boldsymbol{u}\right] d\boldsymbol{x} - \int nD^k\boldsymbol{u}\cdot D^k\left(\frac{1}{n}Dp - \frac{1}{n^*}Dp^*\right) d\boldsymbol{x}\\
&- \int \frac{n}{\tau}\left|D^k\boldsymbol{u}\right|^2 d\boldsymbol{x} \equiv I_{11} + I_{12} + I_{13} + I_{14}.
\end{aligned}
$$

Applying Lemma 2.1, the first integral in the previous expression can be estimated as

$$
\begin{aligned}
I_{11} &\le \|n\|_{L^\infty}\left\|D^k\boldsymbol{u}\right\|\sum_{r=1}^{N}\left\|D^k(u^r(\partial_r\boldsymbol{u})) - u^rD^k(\partial_r\boldsymbol{u})\right\|\\
&\le C\left\|D^k\boldsymbol{u}\right\|\sum_{r=1}^{N}\left(\|Du^r\|_{L^\infty}\left\|D^{k-1}\partial_r\boldsymbol{u}\right\| + \|\partial_r\boldsymbol{u}\|_{L^\infty}\left\|D^ku^r\right\|\right)\\
&\le \tau C\mathcal{U}\left\|D^k\boldsymbol{u}\right\|^2.
\end{aligned}
\tag{6.15}
$$

In a similar way, we find

$$
\begin{aligned}
I_{12} &\le \left\|D^k\delta\boldsymbol{E}\right\|\left(\left\|D^k(\delta n\boldsymbol{u})\right\| + \|\delta n\|_{L^\infty}\left\|D^k\boldsymbol{u}\right\| + \left\|D^k(n^*\boldsymbol{u}) - n^*D^k\boldsymbol{u}\right\|\right)\\
&\le C\left\|D^k\delta\boldsymbol{E}\right\|\left(\|\delta n\|_{L^\infty}\left\|D^k\boldsymbol{u}\right\| + \|\boldsymbol{u}\|_{L^\infty}\left\|D^k\delta n\right\| + \|Dn^*\|_{H^{k-1}}\|\boldsymbol{u}\|_{H^{k-1}}\right)\\
&\le \tau C\mathcal{U}\left\|D^k\delta\hat{U}\right\|\left(\frac{1}{\tau}\left\|D^k\boldsymbol{u}\right\| + \left\|D^k\delta\hat{U}\right\|\right)\\
&+ C\|Dn^*\|_{H^{k-1}}\|\boldsymbol{u}\|_{H^{k-1}}\left\|D^k\delta\boldsymbol{E}\right\|.
\end{aligned}
\tag{6.16}
$$

Using the Corollary 2.4, we can estimate

$$
\begin{aligned}
I_{13} &\le C\left(\left\|\delta\hat{U}\right\|_{C^1}\left\|D^k\delta\hat{U}\right\| + \|Dn^*\|_{H^k}\left\|\delta\hat{U}\right\|_{H^{k-1}}\right)\|D^k\boldsymbol{u}\|\\
&- \sum_{r=1}^{N}\int D^ku^r\cdot\left\{\frac{\partial p}{\partial n}D^k\partial_r\delta n + \frac{\partial p}{\partial S}D^k\partial_r\delta S\right.\\
&+ knD\left(\frac{1}{n}\frac{\partial p}{\partial n}\right)^*\otimes D^{k-1}(\partial_r\delta n) + knD\left(\frac{1}{n}\frac{\partial p}{\partial S}\right)^*\otimes D^{k-1}(\partial_r\delta S)\\
&+ n\left[\frac{\partial}{\partial n}\left(\frac{1}{n}\frac{\partial p}{\partial n}\right)^*D^k\delta n + \frac{\partial}{\partial S}\left(\frac{1}{n}\frac{\partial p}{\partial n}\right)^*D^k\delta S\right]\partial_rn^*\\
&\left.+ n\left[\frac{\partial}{\partial n}\left(\frac{1}{n}\frac{\partial p}{\partial S}\right)^*D^k\delta n + \frac{\partial}{\partial S}\left(\frac{1}{n}\frac{\partial p}{\partial S}\right)^*D^k\delta S\right]\partial_rS^*\right\} d\boldsymbol{x}\\
&\le C\left(\mathcal{U}\left\|D^k\delta\hat{U}\right\| + \|Dn^*\|_{H^k}\left\|\delta\hat{U}\right\|_{H^{k-1}}\right)\|D^k\boldsymbol{u}\|\\
&- \sum_{r=1}^{N}\int D^ku^r\cdot\partial_r\left\{\frac{\partial p}{\partial n}D^k\delta n + \frac{\partial p}{\partial S}D^k\delta S\right\} d\boldsymbol{x}
\end{aligned}
\tag{6.17}
$$

$$-\sum_{r=1}^{N}\int D^k u^r\cdot\left\{kn^*D\left(\frac{1}{n}\frac{\partial p}{\partial n}\right)^*\otimes D^{k-1}(\partial_r\delta n)\right.$$
$$\left.+kn^*D\left(\frac{1}{n}\frac{\partial p}{\partial S}\right)^*\otimes D^{k-1}(\partial_r\delta S)-\frac{1}{n^*}\partial_r p^* D^k\delta n\right\}d\boldsymbol{x}.$$

Next, we turn our attention to I_2. We can write

$$\begin{aligned}
I_2 &= -\int\left\{\left(\frac{\partial^2 e}{\partial n^2}D^k\delta n+\frac{\partial^2 e}{\partial n\partial S}D^k\delta S\right)D^kD\cdot(n\boldsymbol{u})\right.\\
&+\left(\frac{\partial^2 e}{\partial n\partial S}D^k\delta n+\frac{\partial^2 e}{\partial S^2}D^k\delta S\right)D^k[D\cdot(S\boldsymbol{u})-Q]\\
&\left.-\frac{1}{2}\frac{\partial}{\partial t}\frac{\partial^2 e}{\partial n^2}D^k\delta nD^k\delta n-\frac{\partial}{\partial t}\frac{\partial^2 e}{\partial n\partial S}D^k\delta nD^k\delta S-\frac{1}{2}\frac{\partial}{\partial t}\frac{\partial^2 e}{\partial S^2}D^k\delta SD^k\delta S\right\}d\boldsymbol{x}\\
&=\int\left\{\frac{1}{2}\left|D^k\delta n\right|^2\left[\frac{\partial}{\partial t}\frac{\partial^2 e}{\partial n^2}+D\cdot\left(\frac{\partial^2 e}{\partial n^2}\boldsymbol{u}\right)\right]+D^k\delta n\cdot D^k\delta S\left[\frac{\partial}{\partial t}\frac{\partial^2 e}{\partial n\partial S}\right.\right.\\
&\left.\left.+D\cdot\left(\frac{\partial^2 e}{\partial n\partial S}\boldsymbol{u}\right)\right]+\frac{1}{2}\left|D^k\delta S\right|^2\left[\frac{\partial}{\partial t}\frac{\partial^2 e}{\partial S^2}+D\cdot\left(\frac{\partial^2 e}{\partial S^2}\boldsymbol{u}\right)\right]\right\}d\boldsymbol{x}\\
&-\int\left\{\left(\frac{\partial^2 e}{\partial n^2}D^k\delta n+\frac{\partial^2 e}{\partial n\partial S}D^k\delta S\right)\cdot\left[D^kD\cdot(n\boldsymbol{u})-nD^kD\cdot\boldsymbol{u}-(\boldsymbol{u}\cdot D)D^k\delta n\right]\right.\\
&\left.+\left(\frac{\partial^2 e}{\partial n\partial S}D^k\delta n+\frac{\partial^2 e}{\partial S^2}D^k\delta S\right)\cdot\left[D^kD\cdot(S\boldsymbol{u})-SD^kD\cdot\boldsymbol{u}-(\boldsymbol{u}\cdot D)D^k\delta S\right]\right\}d\boldsymbol{x}\\
&-\int\left[n\left(\frac{\partial^2 e}{\partial n^2}D^k\delta n+\frac{\partial^2 e}{\partial n\partial S}D^k\delta S\right)+S\left(\frac{\partial^2 e}{\partial n\partial S}D^k\delta n+\frac{\partial^2 e}{\partial S^2}D^k\delta S\right)\right]D^kD\cdot\boldsymbol{u}\,d\boldsymbol{x}\\
&+\int\left(\frac{\partial^2 e}{\partial n\partial S}D^k\delta n+\frac{\partial^2 e}{\partial S^2}D^k\delta S\right)D^kQ\,d\boldsymbol{x}=I_{21}+I_{22}+I_{23}+I_{24}.
\end{aligned}$$

In order to estimate I_{21}, we observe that, for any function $a(n,S)$,

$$\begin{aligned}
\frac{\partial a}{\partial t}+D\cdot(a\boldsymbol{u})&=\left(a-\frac{\partial a}{\partial n}n-\frac{\partial a}{\partial S}S\right)D\cdot\boldsymbol{u}+\frac{\partial a}{\partial S}Q\\
&\le C\left(\|D\boldsymbol{u}\|_{L^\infty}+\frac{1}{\tau}\|\boldsymbol{u}\|_{L^\infty}^2+\frac{1}{\sigma}\left\|\delta\hat{U}\right\|_{L^\infty}^2\right)\le\tau C\mathcal{U}.
\end{aligned}$$

We can conclude that

$$I_{21}\le\tau C\mathcal{U}\left\|D^k\delta\hat{U}\right\|^2. \tag{6.18}$$

To estimate I_{22}, we recall that $n=n^*+\delta n$, $S=S^*+\delta S$. Then, posing $\mu=n,S$ and using Moser's calculus, we can estimate

$$\begin{aligned}
&\left\|D^kD\cdot(\delta\mu\boldsymbol{u})-\delta\mu D^kD\cdot\boldsymbol{u}-(\boldsymbol{u}\cdot D)D^k\delta\mu\right\|\\
&\le\left\|D^k(\delta\mu D\cdot\boldsymbol{u})-\delta\mu D^kD\cdot\boldsymbol{u}\right\|+\left\|D^k(\boldsymbol{u}\cdot D\delta\mu)-(\boldsymbol{u}\cdot D)D^k\delta\mu\right\|\\
&\le C(\|D\boldsymbol{u}\|_{L^\infty}\left\|D^k\delta\mu\right\|+\|D\delta\mu\|_{L^\infty}\left\|D^k\boldsymbol{u}\right\|)\\
&\le\tau C\mathcal{U}\left(\frac{1}{\tau}\left\|D^k\boldsymbol{u}\right\|+\left\|D^k\delta\hat{U}\right\|\right).
\end{aligned} \tag{6.19}$$

Also, recalling Lemma 2.2, we have

$$
\begin{aligned}
(6.20)\quad & \left\|D^k D\cdot(\mu^*\boldsymbol{u})-\mu^* D^k D\cdot\boldsymbol{u}-D^k\boldsymbol{u}\cdot D\mu^*-kD\mu^*\otimes(D^{k-1}(D\cdot\boldsymbol{u}))\right\| \\
&\le \left\|D^k(\mu^* D\cdot\boldsymbol{u})-\mu^* D^k D\cdot\boldsymbol{u}-kD\mu^*\otimes(D^{k-1}(D\cdot\boldsymbol{u}))\right\| \\
&+\sum_{r=1}^N\left\|D^k(u^r\partial_r\mu^*)-D^k u^r\partial_r\mu^*\right\| \\
&\le C\left\|D^2\mu^*\right\|_{H^{k-2}}\|D\cdot\boldsymbol{u}\|_{H^{k-2}}+C\sum_{r=1}^N\|D\partial_r\mu^*\|_{H^{k-1}}\|u^r\|_{H^{k-1}} \\
&\le C\left\|D^2\mu^*\right\|_{H^{k-1}}\|\boldsymbol{u}\|_{H^{k-1}}.
\end{aligned}
$$

Using (6.19) and (6.20) we obtain

$$
\begin{aligned}
&I_{22}\le\tau C\mathcal{U}\left(\frac{1}{\tau}\left\|D^k\boldsymbol{u}\right\|+\left\|D^k\delta\hat{U}\right\|\right)\left\|D^k\delta\hat{U}\right\|+C\left\|D^2n^*\right\|_{H^{k-1}}\|\boldsymbol{u}\|_{H^{k-1}}\left\|D^k\delta\hat{U}\right\| \\
&-\int\left\{\left(\frac{\partial^2 e}{\partial n^2}D^k\delta n+\frac{\partial^2 e}{\partial n\partial S}D^k\delta S\right)^*\cdot\left[D^k\boldsymbol{u}\cdot Dn^*+kDn^*\otimes(D^{k-1}(D\cdot\boldsymbol{u}))\right]\right. \\
&\left.+\left(\frac{\partial^2 e}{\partial n\partial S}D^k\delta n+\frac{\partial^2 e}{\partial S^2}D^k\delta S\right)^*\cdot\left[D^k\boldsymbol{u}\cdot DS^*+kDS^*\otimes(D^{k-1}(D\cdot\boldsymbol{u}))\right]\right\}d\boldsymbol{x} \\
&=\tau C\mathcal{U}\left(\frac{1}{\tau}\left\|D^k\boldsymbol{u}\right\|+\left\|D^k\delta\hat{U}\right\|\right)\left\|D^k\delta\hat{U}\right\|+C\left\|D^2n^*\right\|_{H^{k-1}}\|\boldsymbol{u}\|_{H^{k-1}}\left\|D^k\delta\hat{U}\right\| \\
&-\int\left\{D^k\delta n\cdot\left[D^k\boldsymbol{u}\cdot D\left(\frac{\partial e}{\partial n}\right)^*+kD\left(\frac{\partial e}{\partial n}\right)^*\otimes(D^{k-1}(D\cdot\boldsymbol{u}))\right]\right. \\
&\left.+D^k\delta S\cdot\left[D^k\boldsymbol{u}\cdot D\left(\frac{\partial e}{\partial S}\right)^*+kD\left(\frac{\partial e}{\partial S}\right)^*\otimes(D^{k-1}(D\cdot\boldsymbol{u}))\right]\right\}d\boldsymbol{x} \\
&=\tau C\mathcal{U}\left(\frac{1}{\tau}\left\|D^k\boldsymbol{u}\right\|+\left\|D^k\delta\hat{U}\right\|\right)\left\|D^k\delta\hat{U}\right\|+C\left\|D^2n^*\right\|_{H^{k-1}}\|\boldsymbol{u}\|_{H^{k-1}}\left\|D^k\delta\hat{U}\right\| \\
&-\sum_{r=1}^N\int D^k\delta n\cdot\left[D^k u^r\partial_r\phi^*+kD\phi^*\otimes(D^{k-1}(\partial_r u^r))\right]d\boldsymbol{x}.
\end{aligned}
$$

Here, we have used the equilibrium condition (3.2). Integrating two times by parts, we arrive at the estimate

$$
\begin{aligned}
(6.21)\quad & I_{22}\le\tau C\mathcal{U}\left(\frac{1}{\tau}\left\|D^k\boldsymbol{u}\right\|+\left\|D^k\delta\hat{U}\right\|\right)\left\|D^k\delta\hat{U}\right\| \\
&+C\left\|D^2n^*\right\|_{H^{k-1}}\|\boldsymbol{u}\|_{H^{k-1}}\left\|D^k\delta\hat{U}\right\| \\
&-\sum_{r=1}^N\int D^k u^r\cdot\left[D^k\delta n\partial_r\phi^*+kD\phi^*\otimes(D^{k-1}(\partial_r\delta n))\right]d\boldsymbol{x}.
\end{aligned}
$$

Next, using (1.8) and integrating by parts, we can write

$$
(6.22)\qquad I_{23}=\sum_{r=1}^N\int D^k u^r\cdot\partial_r\left(\frac{\partial p}{\partial n}D^k\delta n+\frac{\partial p}{\partial S}D^k\delta S\right)d\boldsymbol{x}.
$$

This term cancels out with the first integral on the right-hand side of (6.17). Next, we come to I_{24}. We observe that

$$\begin{aligned}
&\left\| D^k Q + \frac{1}{\sigma}\left(\frac{\partial S}{\partial T}\right)_n D^k \delta T \right\| \\
&= \tau \left\| \left(1 - \frac{\tau}{2\sigma}\right) D^k \left(\frac{n\,|\boldsymbol{u}|^2}{T\tau^2} \right) - \frac{1}{\sigma\tau}\left[D^k \left(\left(\frac{\partial S}{\partial T}\right)_n \delta T \right) - \left(\frac{\partial S}{\partial T}\right)_n D^k \delta T \right] \right\| \\
&\le \tau C \mathcal{U} \left(\frac{1}{\tau} \left\| D^k \boldsymbol{u} \right\| + \left\| D^k \delta \hat{U} \right\| \right) + \tau C \left\| D n^* \right\|_{H^{k-1}} \left\| \delta \hat{U} \right\|_{H^{k-1}}.
\end{aligned}$$

Then, proceeding as before, we find

$$\begin{aligned}
I_{24} &= \int \left(\frac{\partial T}{\partial n} D^k \delta n + \frac{\partial T}{\partial S} D^k \delta S \right) \left(D^k Q + \frac{1}{\sigma} \frac{\partial S}{\partial T} D^k \delta T \right) d\boldsymbol{x} \\
&- \int \frac{1}{\sigma} \frac{\partial S}{\partial T} \left(\frac{\partial T}{\partial n} D^k \delta n + \frac{\partial T}{\partial S} D^k \delta S \right) \left(D^k \delta T - \frac{\partial T}{\partial n} D^k \delta n - \frac{\partial T}{\partial S} D^k \delta S \right) d\boldsymbol{x} \\
&- \int \frac{1}{\sigma} \frac{\partial S}{\partial T} \left(\frac{\partial T}{\partial n} D^k \delta n + \frac{\partial T}{\partial S} D^k \delta S \right)^2 d\boldsymbol{x} \\
&\le \tau C \mathcal{U} \left(\frac{1}{\tau} \left\| D^k \boldsymbol{u} \right\| + \left\| D^k \delta \hat{U} \right\| \right) \left\| D^k \delta \hat{U} \right\| \\
&+ \tau C \left\| D n^* \right\|_{H^{k-1}} \left\| \delta \hat{U} \right\|_{H^{k-1}} \left\| D^k \delta \hat{U} \right\| \\
&- \int \frac{1}{\sigma} \left(\frac{\partial S}{\partial T} \right)^* \left[\left(\frac{\partial T}{\partial n} \right)^* D^k \delta n + \left(\frac{\partial T}{\partial S} \right)^* D^k \delta S \right]^2 d\boldsymbol{x}.
\end{aligned} \tag{6.23}$$

Summing up, using (6.15), (6.16), (6.17), (6.18), (6.21), (6.22), and (6.23), we obtain the estimate

$$\begin{aligned}
I_1 + I_2 &\le \tau C \mathcal{U} \left\| D^k \delta U_\tau \right\|^2 + \tau C \left\| D n^* \right\|_{H^k} \left\| \delta U_\tau \right\|_{H^{k-1}} \left\| D^k \delta U_\tau \right\| \\
&- \sum_{r=1}^{N} \int D^k u^r \cdot k \left\{ \left[D\phi^* + n^* D \left(\frac{1}{n} \frac{\partial p}{\partial n} \right)^* \right] \otimes D^{k-1} (\partial_r \delta n) \right. \\
&\left. + n^* D \left(\frac{1}{n} \frac{\partial p}{\partial S} \right)^* \otimes D^{k-1} (\partial_r \delta S) \right\} d\boldsymbol{x} - \int \frac{n}{\tau} \left| D^k \boldsymbol{u} \right|^2 d\boldsymbol{x} \\
&- \int \frac{1}{\sigma} \left(\frac{\partial S}{\partial T} \right)^* \left[\left(\frac{\partial T}{\partial n} \right)^* D^k \delta n + \left(\frac{\partial T}{\partial S} \right)^* D^k \delta S \right]^2 d\boldsymbol{x}, \\
&\le \tau C \mathcal{U} \left\| D^k \delta U_\tau \right\|^2 + \tau C \left\| D n^* \right\|_{H^k} \left\| \delta U_\tau \right\|_{H^{k-1}} \left\| D^k \delta U_\tau \right\| \\
&+ c^* \left\| D n^* \right\| \left\| D^k \boldsymbol{u} \right\| \left\| D^k \delta \hat{U} \right\| - \int \frac{n}{\tau} \left| D^k \boldsymbol{u} \right|^2 d\boldsymbol{x} \\
&- \int \frac{1}{\sigma} \left(\frac{\partial S}{\partial T} \right)^* \left[\left(\frac{\partial T}{\partial n} \right)^* D^k \delta n + \left(\frac{\partial T}{\partial S} \right)^* D^k \delta S \right]^2 d\boldsymbol{x},
\end{aligned} \tag{6.24}$$

where the constant c^* depends only on the function $e(n, S)$ and on $b^{\pm}$.

Next, we consider I_3.

$$
\begin{aligned}
I_3 &= \int \frac{\lambda_k}{\sigma n^*} \left\{ n D^k \boldsymbol{u} \cdot D^k \left(D\Delta^{-1} D \cdot (n\boldsymbol{u}) \right) + D \cdot (n\boldsymbol{u}) D^k \boldsymbol{u} \cdot D^k \delta \boldsymbol{E} \right. \\
&\left. + n D^k \left(\boldsymbol{u} \cdot D\boldsymbol{u} + \frac{1}{n} Dp - \frac{1}{n^*} Dp^* - \delta \boldsymbol{E} + \frac{1}{\tau} \boldsymbol{u} \right) \cdot D^k \delta \boldsymbol{E} \right\} d\boldsymbol{x} \\
&= \int \frac{\lambda_k n}{\sigma n^*} D^k \boldsymbol{u} \cdot D^k \left(D\Delta^{-1} D \cdot (n\boldsymbol{u}) \right) d\boldsymbol{x} \\
&+ \int \frac{\lambda_k}{\sigma n^*} \left[D \cdot (n\boldsymbol{u}) D^k \boldsymbol{u} + n D^k ((\boldsymbol{u} \cdot D)\boldsymbol{u}) \right] \cdot D^k \delta \boldsymbol{E} \, d\boldsymbol{x} \\
&+ \int \frac{\lambda_k n}{\sigma n^*} D^k \left[\frac{1}{n} Dp - \frac{1}{n^*} Dp^* \right] \cdot D^k \delta \boldsymbol{E} \, d\boldsymbol{x} \\
&\int \frac{\lambda_k n}{\sigma n^*} \left(D^k \delta \boldsymbol{E} - \frac{1}{\tau} D^k \boldsymbol{u} \right) \cdot D^k \delta \boldsymbol{E} \, d\boldsymbol{x} \equiv I_{31} + I_{32} + I_{33} + I_{34}.
\end{aligned}
$$

We can estimate I_{31} immediately as

$$
\begin{aligned}
I_{31} &\leq \frac{\lambda_k}{\sigma \inf n^*} \left\| n D^k \boldsymbol{u} \right\| \left\| D^k D\Delta^{-1} D \cdot (n\boldsymbol{u}) \right\| \\
&= \frac{\lambda_k}{\sigma \inf n^*} \left\| n D^k \boldsymbol{u} \right\| \left\| D^{k-1} D \cdot (n\boldsymbol{u}) \right\| \\
&\leq \frac{\lambda_k}{\sigma \inf n^*} \left\| n D^k \boldsymbol{u} \right\| \left(\left\| D^k (n\boldsymbol{u}) - n D^k \boldsymbol{u} \right\| + \left\| n D^k \boldsymbol{u} \right\| \right) \\
&\leq \frac{\lambda_k}{\sigma \inf n^*} \left\| n D^k \boldsymbol{u} \right\|^2 + \frac{C}{\sigma} \left\| D^k \boldsymbol{u} \right\| \left(\| Dn^* \|_{L^\infty} \left\| D^{k-1} \boldsymbol{u} \right\| \right. \\
&\left. + \left\| D^k n^* \right\| \| \boldsymbol{u} \|_{L^\infty} + \| \boldsymbol{u} \|_{L^\infty} \left\| D^k \delta n \right\| + \| \delta n \|_{L^\infty} \left\| D^k \boldsymbol{u} \right\| \right) \\
&\leq \frac{\lambda_k}{\sigma \inf n^*} \left\| n D^k \boldsymbol{u} \right\|^2 \\
&+ \tau C \left\| D^k \delta U_\tau \right\| \left(\| Dn^* \|_{H^{k-1}} \| \delta U_\tau \|_{H^{k-1}} + \mathcal{U} \left\| D^k \delta U_\tau \right\| \right).
\end{aligned} \tag{6.25}
$$

For I_{32}, integrating by parts, recalling that $\delta \boldsymbol{E} = D\delta\phi$, and using (4.6), we find

$$
\begin{aligned}
I_{32} &\leq C \frac{\lambda_k}{\sigma} \left(\| n \|_{L^\infty} \| D\boldsymbol{u} \| + \| \boldsymbol{u} \|_{L^\infty} \| Dn \| \right) \left\| D^k \boldsymbol{u} \right\| \left\| D^k \delta \boldsymbol{E} \right\| \\
&+ \sum_{j,r=1}^{N} \int \frac{\lambda_k}{\sigma} \partial_j D^{k-1} ((\boldsymbol{u} \cdot D) u^r) \cdot \left\{ \frac{n}{n^*} \partial_j D^{k-1} \delta E^r \right\} d\boldsymbol{x} \\
&\leq C \frac{\lambda_k}{\sigma \tau} \mathcal{U} \left\| D^k \boldsymbol{u} \right\| \left\| D^k \delta \boldsymbol{E} \right\| \\
&- \sum_{j,r=1}^{N} \int \frac{\lambda_k}{\sigma} D^{k-1} ((\boldsymbol{u} \cdot D) u^r) \cdot \partial_j \left\{ \frac{n}{n^*} \partial_r D^{k-1} \delta E^j \right\} d\boldsymbol{x} \\
&\leq \tau C \mathcal{U} \left\| D^k \delta U_\tau \right\|^2 + \frac{\lambda_k}{\sigma} \sum_{r=1}^{N} \left\| D^{k-1} ((\boldsymbol{u} \cdot D) u^r) \right\| \left\{ \left\| \frac{n}{n^*} \partial_r D^{k-1} \delta n \right\| \right. \\
&\left. + \left\| D \left(\frac{\delta n}{n^*} \right) \right\|_{L^\infty} \left\| \partial_r D^{k-1} \delta \boldsymbol{E} \right\| \right\} \\
&\leq \tau C \mathcal{U} \left\| D^k \delta U_\tau \right\|^2 + \frac{C}{\sigma} \| \boldsymbol{u} \|_{L^\infty} \left\| D^k \boldsymbol{u} \right\| \left(\left\| D^k \delta n \right\| + \| \delta n \|_{C^1} \left\| D^k \delta \boldsymbol{E} \right\| \right).
\end{aligned} \tag{6.26}
$$

In a similar way, we find

$$
\begin{aligned}
I_{33} &= -\int \frac{\lambda_k}{\sigma} D^{k-1}\left(\frac{1}{n}Dp - \frac{1}{n^*}Dp^*\right)\cdot \sum_{r=1}^{N}\partial_r\left\{\frac{n}{n^*}D^{k-1}\partial_r\delta\boldsymbol{E}\right\}\,d\boldsymbol{x}\\
&\le \frac{\lambda_k}{\sigma}\left\|D^{k-1}\left(\frac{Dp}{n}-\frac{Dp^*}{n^*}\right)\right\|\left\|\sum_{r=1}^{N}\partial_r\left(\frac{\delta n}{n^*}\right)D^{k-1}\partial_r\delta\boldsymbol{E}+\frac{\delta n}{n^*}D^k\delta n\right\|\\
&+\frac{\lambda_k}{\sigma}\left\|D^{k-1}\left(\frac{Dp}{n}-\frac{Dp^*}{n^*}\right)-\left(\frac{1}{n}\frac{\partial p}{\partial n}\right)^* D^k\delta n-\left(\frac{1}{n}\frac{\partial p}{\partial S}\right)^* D^k\delta S\right\|\left\|D^k\delta n\right\|\\
&-\int\frac{\lambda_k}{\sigma}\left\{\left(\frac{1}{n}\frac{\partial p}{\partial n}\right)^* D^k\delta n+\left(\frac{1}{n}\frac{\partial p}{\partial S}\right)^* D^k\delta S\right\}\cdot D^k\delta n\,d\boldsymbol{x}.
\end{aligned}
$$

Using Lemma 2.3 and Corollary 2.4, we can estimate

$$
\begin{aligned}
I_{33} &\le \tau C\mathcal{U}\left\|D^k\delta U_\tau\right\|^2+\tau C\left\|Dn^*\right\|_{H^k}\left\|\delta U_\tau\right\|_{H^{k-1}}\left\|D^k\delta U_\tau\right\|\\
&-\int\frac{\lambda_k}{\sigma}\left\{\left(\frac{1}{n}\frac{\partial p}{\partial n}\right)^* D^k\delta n+\left(\frac{1}{n}\frac{\partial p}{\partial S}\right)^* D^k\delta S\right\}\cdot D^k\delta n\,d\boldsymbol{x}.
\end{aligned}
\tag{6.27}
$$

Proceeding as in (6.8), we find

$$
I_{34}\le\int\left\{\frac{n}{2\tau}\left|D^k\boldsymbol{u}\right|^2-\frac{\lambda_0 n}{2\sigma n^*}\left|D^k\delta\boldsymbol{E}\right|^2\right\}d\boldsymbol{x}.
\tag{6.28}
$$

In conclusion, using (6.25), (6.26), (6.27), and (6.28), we can estimate I_3 as

$$
\begin{aligned}
I_3 &\le \tau C\mathcal{U}\left\|D^k\delta U_\tau\right\|^2+\tau C\left\|Dn^*\right\|_{H^k}\left\|\delta U_\tau\right\|_{H^{k-1}}\left\|D^k\delta U_\tau\right\|\\
&-\int\frac{\lambda_k}{\sigma}\left\{\left(\frac{1}{n}\frac{\partial p}{\partial n}\right)^* D^k\delta n+\left(\frac{1}{n}\frac{\partial p}{\partial S}\right)^* D^k\delta S\right\}\cdot D^k\delta n\,d\boldsymbol{x}\\
&-\int\left\{\frac{\lambda_k}{2\sigma}\left|D^k\delta\boldsymbol{E}\right|^2-\left(\frac{1}{2\tau}+\frac{\lambda_k\sup n^*}{\sigma\inf n^*}\right)n^*\left|D^k\boldsymbol{u}\right|^2\right\}d\boldsymbol{x}.
\end{aligned}
\tag{6.29}
$$

Using the estimates (6.24) and (6.29) in (6.14), we find

$$
\begin{aligned}
\frac{d}{dt}\int\mathcal{H}_k\,d\boldsymbol{x} &\le -\int\mathcal{D}_k\,d\boldsymbol{x}+c^*\left\|Dn^*\right\|\left\|D^k\boldsymbol{u}\right\|\left\|D^k\delta\hat{U}\right\|\\
&+\tau C_k'\mathcal{U}\left\|D^k\delta U_\tau\right\|^2+\tau C_k''\left\|Dn^*\right\|_{H^k}\left\|\delta U_\tau\right\|_{H^{k-1}}\left\|D^k\delta U_\tau\right\|,
\end{aligned}
\tag{6.30}
$$

where C_k' and C_k'' are positive constants which depend only on the equilibrium state and on $\sigma\tau$. Next, we observe that the assumption

$$
\sigma\tau\left\|Dn^*\right\|^2\le C^*\equiv\frac{k_{\mathcal{D}}^2}{4c^{*2}}
\tag{6.31}
$$

implies

$$
\begin{aligned}
&c^*\left\|Dn^*\right\|\left\|D^k\boldsymbol{u}\right\|\left\|D^k\delta\hat{U}\right\|\\
&\le\frac{1}{2}c^*\left\|Dn^*\right\|\left(\frac{k_{\mathcal{D}}}{2\tau c^*\left\|Dn^*\right\|}\left\|D^k\boldsymbol{u}\right\|^2+\frac{2\tau c^*\left\|Dn^*\right\|}{k_{\mathcal{D}}}\left\|D^k\delta\hat{U}\right\|^2\right)\\
&\le\frac{k_{\mathcal{D}}}{4}\left(\frac{1}{\tau}\left\|D^k\boldsymbol{u}\right\|^2+\frac{1}{\sigma}\left\|D^k\delta\hat{U}\right\|^2\right).
\end{aligned}
\tag{6.32}
$$

Thus, using (5.9) and the assumption (6.31), we obtain

$$\frac{d}{dt}\int \mathcal{H}_k\,d\boldsymbol{x} \le -\frac{3}{4}k_{\mathcal{D}}\left(\frac{1}{\tau}\left\|D^k\boldsymbol{u}\right\|^2+\frac{1}{\sigma}\left\|D^k\delta\hat{U}\right\|^2\right) + \tau C_k'\mathcal{U}\left\|D^k\delta U_\tau\right\|^2+\tau C_k''\left\|Dn^*\right\|_{H^k}\left\|\delta U_\tau\right\|_{H^{k-1}}\left\|D^k\delta U_\tau\right\|. \tag{6.33}$$

Next, we use (6.13) and (6.33) in order to estimate the energy $\mathcal{W}$, given by (6.4). We find

$$\frac{d\mathcal{W}}{dt}=\sum_{k=0}^{s}\eta_k\frac{d}{dt}\int\mathcal{H}_k\,d\boldsymbol{x}\le-\tau k_{\mathcal{D}}^*\left(\left\|\delta U_\tau\right\|^2+\frac{3}{4}\sum_{k=1}^{s}\eta_k\left\|D^k\delta U_\tau\right\|^2\right) +\tau C'\mathcal{U}\sum_{k=0}^{s}\eta_k\left\|D^k\delta U_\tau\right\|^2+\tau C''\left\|Dn^*\right\|_{H^s}\sum_{k=1}^{s}\eta_k\left\|\delta U_\tau\right\|_{H^{k-1}}\left\|D^k\delta U_\tau\right\|, \tag{6.34}$$

with

$$k_{\mathcal{D}}^*=\frac{k_{\mathcal{D}}}{\max\{1,\tau\sigma\}},\qquad C'=\max\{C_0',C_1',\dots,C_s'\},\qquad C''=\max\{C_1'',\dots,C_s''\}.$$

Recalling Lemma 2.5, we pose

$$\eta=2,\qquad K=\frac{k_{\mathcal{D}}^*}{4C''\left\|Dn^*\right\|_{H^s}}$$

and choose

$$\eta_k=2\eta\left(\frac{K^2}{1+K^2}\right)^k,\qquad k=1,\dots,s.$$

Then it follows that

$$C''\left\|Dn^*\right\|_{H^s}\sum_{k=1}^{s}\eta_k\left\|\delta U_\tau\right\|_{H^{k-1}}\left\|D^k\delta U_\tau\right\| \le\frac{k_{\mathcal{D}}^*}{4}\left(2\left\|\delta U_\tau\right\|^2+\sum_{k=1}^{s}\eta_k\left\|D^k\delta U_\tau\right\|^2\right). \tag{6.35}$$

Using (6.35) in (6.34), we get

$$\frac{d\mathcal{W}}{dt}\le-\tau\left(\frac{k_{\mathcal{D}}^*}{2}-C'\mathcal{U}\right)\sum_{k=0}^{s}\eta_k\left\|D^k\delta U_\tau\right\|^2.$$

If the solution satisfies

$$\mathcal{U}(T)\le\epsilon\equiv\frac{k_{\mathcal{D}}^*}{4C'}, \tag{6.36}$$

we have

$$\frac{d\mathcal{W}}{dt}\le-\frac{\tau k_{\mathcal{D}}^*}{4}\sum_{k=0}^{s}\eta_k\left\|D^k\delta U_\tau\right\|^2. \tag{6.37}$$

After integration with respect to time and recalling Lemma 5.3, we find

$$\left\|D^k\delta U(\cdot,t)\right\|^2_{\eta,s} \le K - \tau c\int_0^t \left\|D^k\delta U_\tau(\cdot,t')\right\|^2_{\eta,s}\,dt', \tag{6.38}$$

with

$$K = \frac{\mathcal{W}(0)}{k_{\mathcal{H}}} \qquad c = \frac{k^*_{\mathcal{D}}}{4k_{\mathcal{H}}}. \tag{6.39}$$

In particular, from (6.38) we can derive

$$\left\|D^k\boldsymbol{u}(\cdot,t)\right\|^2_{\eta,s} \le K - \frac{c}{\tau}\int_0^t \left\|D^k\boldsymbol{u}(\cdot,t')\right\|^2_{\eta,s}\,dt', \tag{6.40}$$

$$\left\|D^k\delta\hat{U}(\cdot,t)\right\|^2_{\eta,s} \le K - \tau c\int_0^t \left\|D^k\delta\hat{U}(\cdot,t')\right\|^2_{\eta,s}\,dt'. \tag{6.41}$$

The sought a priori estimate follows immediately from (6.40) and (6.41) after applying Gronwall's lemma. ∎

Appendix. In this appendix, we prove Lemmas 2.2, 2.3, 2.4, and 2.5.

Proof of Lemma 2.2. From the identity (2.1), we find

$$\left\|D^k(fg) - \sum_{r=0}^{j-1}\binom{k}{r}(D^r f)\otimes(D^{k-r}g)\right\| \le \sum_{r=0}^{k-j}\binom{k}{j+r}\left\|D^r(D^j f)\right\|\left\|D^{k-j-r}g\right\|.$$

Using Schwarz's inequality, we obtain the thesis. ∎

Proof of Lemma 2.3. If $k=1$, we find immediately that

$$\begin{aligned}
&\|D(F(f)-F(g))\| \le \left|\frac{\partial F}{\partial w}\right|_{C^0}\|D(f-g)\| + \left\|\frac{\partial F}{\partial w}(f)-\frac{\partial F}{\partial w}(g)\right\|\|Dg\|,\\
&\left\|D(F(f)-F(g)) - \frac{\partial F}{\partial w}(f)D(f-g)\right\|\\
&\le \left\|\left(\frac{\partial F}{\partial w}(f)-\frac{\partial F}{\partial w}(g)\right)D(f-g)\right\| + \left\|\frac{\partial F}{\partial w}(f)-\frac{\partial F}{\partial w}(g)\right\|\|Dg\|.
\end{aligned}$$

The estimates (2.6) and (2.7) follow from the identity

$$\frac{\partial F}{\partial w}(f)-\frac{\partial F}{\partial w}(g) = \left\{\int_0^1 \frac{\partial}{\partial w}\frac{\partial F}{\partial w}(g+\lambda(f-g))d\lambda\right\}(f-g). \tag{A.1}$$

Next, let α be an N-tuple of nonnegative integers, with $|\alpha| = k \ge 2$. Then we have

$$\begin{aligned}
&D^\alpha(F(f)-F(g)) \qquad\qquad (A.2)\\
&\quad= \sum_{\mu_1+\cdots+\mu_r=\alpha} C_\mu\left[(D^{\mu_1}f)\cdots(D^{\mu_r}f)F^{(r)}(f) - (D^{\mu_1}g)\cdots(D^{\mu_r}g)F^{(r)}(g)\right]\\
&\quad= \sum_{\mu_1+\cdots+\mu_r=\alpha} C_\mu\Big\{D^{\mu_1}(f-g)\cdots D^{\mu_r}(f-g)F^{(r)}(f) + \Big[(D^{\mu_1}f)\cdots(D^{\mu_r}f)\\
&\qquad - (D^{\mu_1}g)\cdots(D^{\mu_r}g) - D^{\mu_1}(f-g)\cdots D^{\mu_r}(f-g)\Big]F^{(r)}(f)\\
&\qquad + (D^{\mu_1}g)\cdots(D^{\mu_r}g)\left[F^{(r)}(f)-F^{(r)}(g)\right]\Big\},
\end{aligned}$$

where $F^{(r)}$ denotes the rth derivative of F with respect to its argument. We write $f = g+(f-g)$ in (A.2), evaluate the L^2-norm, and use the Cauchy–Schwarz inequality. Then the Gagliardo–Nirenberg inequality yields

$$\|D^{\mu_1}(f-g)\cdots D^{\mu_r}(f-g)\| \le C\,\|f-g\|_{L^\infty}^{r-1}\left\|D^k(f-g)\right\|. \tag{A.3}$$

Furthermore, we have

$$\begin{aligned}
&\left\|(D^{\mu_1}g)\cdots(D^{\mu_r}g)\left[F^{(r)}(f)-F^{(r)}(g)\right]\right\| \\
&\le \|(D^{\mu_1}g)\cdots(D^{\mu_r}g)\|\left\|F^{(r)}(f)-F^{(r)}(g)\right\| \\
&\le C\,\|g\|_{L^\infty}^{r-1}\left\|D^k g\right\|\left\|\left\{\int_0^1\left(\frac{\partial F^{(r)}}{\partial w}\right)(g+\lambda(f-g))d\lambda\right\}(f-g)\right\| \\
&\le C\left|F^{(r+1)}\right|_{C^0}\|g\|_{L^\infty}^{r-1}\left\|D^k g\right\|\|f-g\|.
\end{aligned} \tag{A.4}$$

Also, for $r \ge 2$, we find

$$\begin{aligned}
&\|(D^{\mu_1}f)\cdots(D^{\mu_r}f)-(D^{\mu_1}g)\cdots(D^{\mu_r}g)-D^{\mu_1}(f-g)\cdots D^{\mu_r}(f-g)\| \\
&\le C\sum_{i=1}^{r-1}\|D^{\mu_1}(f-g)\cdots D^{\mu_i}(f-g)\|\,\|(D^{\mu_{i+1}}g)\cdots(D^{\mu_r}g)\| \\
&\le C\sum_{i=1}^{r-1}\|f-g\|_{L^\infty}^{i-1}\left\|D^{|\mu_1+\cdots+\mu_i|}(f-g)\right\|\|g\|_{L^\infty}^{r-i-1}\left\|D^{k-|\mu_1+\cdots+\mu_i|}g\right\|.
\end{aligned} \tag{A.5}$$

In (A.5), we note that the number i of derivatives of $f-g$ cannot exceed $|\mu_1+\cdots+\mu_i|$, and the number $r-i$ of derivatives of g cannot exceed $k-|\mu_1+\cdots+\mu_i|$. Using the estimates (A.3), (A.4), (A.5), we find

$$\begin{aligned}
&\|D^\alpha(F(f)-F(g))\| \le C\sum_{i=1}^{k}\left|F^{(i)}\right|_{C^0}\|f-g\|_{L^\infty}^{i-1}\left\|D^k(f-g)\right\| \\
&+C\sum_{r=1}^{k-1}\sum_{i=1}^{r}\sum_{j=1}^{k-r}\left|F^{(i+j)}\right|_{C^0}\|f-g\|_{L^\infty}^{i-1}\|g\|_{L^\infty}^{j-1}\|D^r(f-g)\|\,\left\|D^{k-r}g\right\| \\
&+C\sum_{j=1}^{k}\left|F^{(j+1)}\right|_{C^0}\|g\|_{L^\infty}^{j-1}\|f-g\|\,\left\|D^k g\right\| \\
&\le C\left|\frac{\partial F}{\partial w}\right|_{C^{k-1}}\left(\sum_{i=1}^{k}\|f-g\|_{L^\infty}^{i-1}\right)\left\|D^k(f-g)\right\| \\
&+C\left|\frac{\partial F}{\partial w}\right|_{C^k}\left(\sum_{i=1}^{k}\sum_{j=1}^{k}\|f-g\|_{L^\infty}^{i-1}\|g\|_{L^\infty}^{j-1}\right)\sum_{r=0}^{k-1}\|D^r(f-g)\|\,\left\|D^{k-r}g\right\|.
\end{aligned} \tag{A.6}$$

The estimate (2.6) follows immediately, with

$$C(F,f,g) = C\left|\frac{\partial F}{\partial w}\right|_{C^k}\left(\sum_{i=1}^{k}\sum_{j=1}^{k}\|f-g\|_{L^\infty}^{i-1}\|g\|_{L^\infty}^{j-1}\right).$$

To prove (2.7), we proceed in the same way to obtain

$$
\begin{aligned}
(\mathrm{A}.7)\quad & \left\| D^\alpha (F(f)-F(g)) - \frac{\partial F}{\partial w}(f) D^\alpha (f-g) \right\| \\
& \le \left\| \left(\frac{\partial F}{\partial w}(f) - \frac{\partial F}{\partial w}(g) \right) D^\alpha (f-g) \right\| \\
& + C \left| \frac{\partial F}{\partial w} \right|_{C^{k-1}} \left(\sum_{i=2}^{k} \|f-g\|_{L^\infty}^{i-2} \right) \|f-g\|_{L^\infty} \left\| D^k (f-g) \right\| \\
& + C \left| \frac{\partial F}{\partial w} \right|_{C^{k}} \left(\sum_{i=1}^{k} \sum_{j=1}^{k} \|f-g\|_{L^\infty}^{i-1} \|g\|_{L^\infty}^{j-1} \right) \sum_{r=0}^{k-1} \|D^r (f-g)\| \, \left\| D^{k-r} g \right\| .
\end{aligned}
$$

The estimate (2.7) follows from the identity (A.1). ∎

Proof of Corollary 2.4. The proof of (2.8) and (2.9) is perfectly analogous to the proof of (2.6) and (2.7). To prove (2.10), we observe that

$$
\begin{aligned}
& \Big\| D^k \left[F(f)\partial f - F(g)\partial g \right] - F(f) D^k \partial (f-g) \\
& - kD\left[F(g)\right] \otimes D^{k-1}\left[\partial (f-g)\right] - \frac{\partial F}{\partial w}(g) D^k (f-g) \partial g \Big\| \\
& \le \left\| D^k \left[(F(f)-F(g))\partial (f-g) \right] - (F(f)-F(g)) D^k \partial (f-g) \right\| \\
& + \left\| D^k \left[F(g)\partial (f-g) \right] - F(g) D^k \partial (f-g) - kD\left[F(g)\right] \otimes D^{k-1}\left[\partial (f-g)\right] \right\| \\
& + \left\| D^k \left[(F(f)-F(g))\partial g \right] - D^k (F(f)-F(g)) \partial g \right\| \\
& + \left\| D^k \left[F(f)-F(g) \right] - \frac{\partial F}{\partial w}(g) D^k (f-g) \right\| \|\partial g\|_{L^\infty} .
\end{aligned}
$$

The thesis follows from Lemmas 2.1, 2.2, and 2.3. ∎

Proof of Lemma 2.5. Using the Cauchy–Schwarz inequality, we have

$$
\begin{aligned}
& \sum_{j=1}^{k} \eta_j \|f\|_{H^{j-1}} \left\| D^j f \right\| \le \frac{1}{2} \sum_{j=1}^{k} \eta_j \left(\frac{1}{\alpha_j} \|f\|_{H^{j-1}}^2 + \alpha_j \left\| D^j f \right\|^2 \right) \\
& = \frac{1}{2} \sum_{j=1}^{k} \eta_j \alpha_j \left\| D^j f \right\|^2 + \frac{1}{2} \sum_{j=1}^{k} \sum_{i=0}^{j-1} \frac{\eta_j}{\alpha_j} \left\| D^i f \right\|^2 \\
& = \frac{1}{2} \sum_{j=1}^{k} \eta_j \alpha_j \left\| D^j f \right\|^2 + \frac{1}{2} \sum_{j=0}^{k-1} \sum_{i=j+1}^{k} \frac{\eta_i}{\alpha_i} \left\| D^j f \right\|^2 \\
& = \frac{1}{2} \eta_k \alpha_k \left\| D^k f \right\|^2 + \frac{1}{2} \sum_{j=1}^{k-1} \left\{ \eta_j \alpha_j + \sum_{i=j+1}^{k} \frac{\eta_i}{\alpha_i} \right\} \left\| D^j f \right\|^2 + \frac{1}{2} \sum_{i=1}^{k} \frac{\eta_i}{\alpha_i} \|f\|^2 ,
\end{aligned}
$$

where α_j, $j = 1, \dots, k$, are constants to be chosen. The thesis amounts to proving that we can choose η_j and α_j, $j = 1, \dots, k$, such that

$$
\begin{aligned}
& \eta_k \alpha_k = 2K \eta_k , \\
& \eta_j \alpha_j + \sum_{i=j+1}^{k} \frac{\eta_i}{\alpha_i} = 2K \eta_j , \quad j = 1, \dots, k-1 ,
\end{aligned}
$$

$$\sum_{i=1}^{k} \frac{\eta_i}{\alpha_i} = 2K\eta.$$

Recursively, we obtain

$$\begin{aligned}
&2K - \alpha_k = 0,\\
&2K - \alpha_j = \sum_{i=j+1}^{k} \frac{\eta_i}{\alpha_i \eta_j} = \sum_{i=j+2}^{k} \frac{\eta_i}{\alpha_i \eta_j} + \frac{\eta_{j+1}}{\alpha_{j+1}\eta_j}\\
&\qquad = \left(2K - \alpha_{j+1} + \frac{1}{\alpha_{j+1}}\right) \frac{\eta_{j+1}}{\eta_j} > 0, \quad j = 1, \dots, k-1,\\
&\eta_1 \left(2K - \alpha_1 + \frac{1}{\alpha_1}\right) = 2K\eta.
\end{aligned}$$

Then we can choose

$$\begin{aligned}
&\alpha_j = 2(1-\beta_j)K, \quad j = 1, \dots, k-1, \qquad \alpha_k = 2K,\\
&\eta_1 = \frac{2K\eta}{2K - \alpha_1 + \frac{1}{\alpha_1}}, \qquad \eta_j = \frac{2K\beta_{j-1}\eta_{j-1}}{2K - \alpha_j + \frac{1}{\alpha_j}}, \quad j = 2, \dots, k,
\end{aligned}$$

for any constants $\beta_j < 1$, $j = 1, \dots, k-1$. In particular, choosing $\beta_j = 1/2$, we get the simple expression

$$\eta_j = 2\eta \left(\frac{K^2}{1+K^2}\right)^j, \quad j = 1, \dots, k. \qquad \square$$

REFERENCES

[1] G. Albinus, *A Thermodynamically Motivated Formulation of the Energy Model of Semiconductor Devices*, Preprint 210, Weierstrass-Institut, Berlin, 1995.

[2] G. Alì, *Asymptotic Fluid Dynamic Models for Semiconductors*, Quaderno IAC 25/1995, Istituto per le Applicazioni del Calcolo, CNR, Roma, 1995.

[3] G. Alì, D. Bini, and S. Rionero, *Global existence and relaxation limit for smooth solutions to the Euler–Poisson model for semiconductors*, SIAM J. Math. Anal., 32 (2000), pp. 572–587.

[4] G. Alì, P. Marcati, and R. Natalini, *Hydrodynamic models for semiconductors*, Z. Angew. Math. Mech., 76 (1996), pp. 301–304.

[5] A. M. Anile, *An extended thermodynamic framework for the hydrodynamical modeling of semiconductors*, in Mathematical Problems in Semiconductors Physics (Rome, 1993), P. A. Marcati, P. A. Markowich, and R. Natalini, eds., Pitman Res. Notes Math. Ser. 3, Longman, Harlow, UK, 1995, pp. 3–41.

[6] A. M. Anile and S. Pennisi, *Thermodynamic derivation of the hydrodynamical model for charge transport in semiconductors*, Phys. Rev. B, 46 (1992), pp. 13186–13193.

[7] A. M. Anile and V. Romano, *Hydrodynamical modeling of charge carrier transport in semiconductors*, Meccanica, 35 (2000), pp. 249–296.

[8] K. Bløtekjær, *Transport equations for electrons in two-valley semiconductors*, IEEE Trans. Electron. Devices, ED-17 (1970), pp. 38–47.

[9] G. Baccarani and M. R. Wordeman, *An investigation of steady-state velocity overshoot effects in Si and GaAs devices*, Solid State Electr., 28 (1985), pp. 407–416.

[10] G.-Q. Chen, J. W. Jerome, and B. Zhang, *Existence and the singular relaxation limit for the inviscid hydrodynamic energy model*, in Modelling and Computation for Applications in Mathematics, Science, and Engineering (Evanston, IL, 1996), Numer. Math. Sci. Comput., Oxford University Press, New York, 1998, pp. 189–215.

[11] G.-Q. Chen and D. Wang, *Convergence of shock schemes for the compressible Euler-Poisson equations*, Comm. Math. Phys., 179 (1996), pp. 333–364.

[12] G.-Q. Chen and D. Wang, *Formation of singularities in compressible Euler-Poisson fluids with heat diffusion and damping relaxation*, J. Differential Equations, 144 (1998), pp. 44–65.

[13] P. Degond, *Mathematical modelling of microelectronics semiconductor devices*, in Some Current Topics on Nonlinear Conservation Laws, AMS/IP Stud. Adv. Math. 15, AMS, Providence, RI, 2000, pp. 77–110.

[14] P. Degond, S. Genieys, and A. Jüngel, *A steady-state system in nonequilibrium thermodynamics including thermal and electrical effects*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 867–872.

[15] P. Degond, S. Genieys, and A. Jüngel, *A system of parabolic equations in nonequilibrium thermodynamics including thermal and electrical effects*, J. Math. Pure Appl., 76 (1997), pp. 991–1015.

[16] P. Degond and P. A. Markowich, *On a one-dimensional steady-state hydrodynamic model for semiconductors*, Appl. Math. Lett., 3 (1990), pp. 25–29.

[17] P. Degond and P. A. Markowich, *A steady-state potential flow model for semiconductors*, Ann. Mat. Pura Appl. (4), 165 (1993), pp. 87–98.

[18] W. Dreyer, *Maximization of the entropy in non-equilibrium*, J. Phys. A, 20 (1987), pp. 6505–6517.

[19] S. Engelberg, H. L. Liu, and E. Tadmor, *Critical thresholds in the Euler–Poisson equations*, Indiana Univ. Math. J., 50 (2001), pp. 109–157.

[20] I. M. Gamba, *Stationary transonic solutions of a one-dimensional hydrodynamic model for semiconductors*, Comm. Partial Differential Equations, 17 (1992), pp. 553–577.

[21] C. L. Gardner, J. W. Jerome, and D. J. Rose, *Numerical methods for the hydrodynamic device model: Subsonic flow*, IEEE Trans. Electron. Devices, 8 (1989), pp. 501–507.

[22] I. Gasser and R. Natalini, *The energy transport and the drift diffusion equations as relaxation limits of the hydrodynamic model for semiconductors*, Quart. Appl. Math., 57 (1999), pp. 269–282.

[23] Y. Guo, *Smooth irrotational flows in the large to the Euler–Poisson system in* $\mathbb{R}^{3+1}$, Comm. Math. Phys., 195 (1998), pp. 249–265.

[24] L. Hsiao, P. Markowich, and S. Wang, *The asymptotic behavior of globally smooth solutions of the multidimensional isentropic hydrodynamic model for semiconductors*, J. Differential Equations, to appear.

[25] L. Hsiao and S. Wang, *The asymptotic behavior of global solutions to the hydrodynamic model with spherical symmetry*, Nonlinear Anal., to appear.

[26] L. Hsiao and S. Wang, *The asymptotic behavior of global solutions to the hydrodynamic model in the exterior domain*, Acta Math. Scientia, to appear.

[27] L. Hsiao and S. Wang, *The Cauchy Problem to the Multidimensional Full Hydrodynamic Model for Semiconductors*, preprint.

[28] L. Hsiao and K. J. Zhang, *The relaxation of the hydrodynamic model for semiconductors to the drift-diffusion equations*, J. Differential Equations, 165 (2000), pp. 315–354.

[29] F. Jochmann, *Global weak solutions of the one-dimensional hydrodynamic model for semiconductors*, Math. Models Methods Appl. Sci., 3 (1993), pp. 759–788.

[30] D. Jou, J. Casa-Vasquez, and G. Lebon, *Extended Irreversible Thermodynamics*, Springer-Verlag, Berlin, 1993.

[31] T. Kato, *The Cauchy problem for quasi linear symmetric hyperbolic systems*, Arch. Ration. Mech. Anal., 38 (1975), pp. 181–205.

[32] S. Klainerman and A. Majda, *Singular perturbation of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids*, Comm. Pure Appl. Math., 34 (1981), pp. 481–524.

[33] C. D. Levermore, *Moment closure hierarchies for kinetic theories*, J. Statist. Phys., 83 (1996), pp. 331–407.

[34] T. Luo, R. Natalini, and Z. Xin, *Large time behavior of the solutions to a hydrodynamic model for semiconductors*, SIAM J. Appl. Math., 59 (1998), pp. 810–830.

[35] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Appl. Math. Sci. 53, Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1984.

[36] P. Marcati and R. Natalini, *Weak solutions to a hydrodynamic model for semiconductors: The Cauchy problem*, Proc. Roy. Soc. Edinburgh Sect. A, 28 (1995), pp. 115–131.

[37] P. Marcati and R. Natalini, *Weak solutions to a hydrodynamic model for semiconductors and relaxation to the drift-diffusion equation*, Arch. Ration. Mech. Anal., 129 (1995), pp. 129–145.

[38] A. Matsumura and T. Nishida, *The initial value problem for the equations of motion of viscous and heat-conductive gases*, J. Math. Kyoto Univ., 20 (1980), pp. 67–104.

[39] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer-Verlag, Wien, New York, 1990.
[40] I. Müller and T. Ruggeri, *Extended Thermodynamics*, Springer-Verlag, Berlin, 1993.
[41] F. Poupaud, M. Rascle, and J.-P. Vila, *Global solutions to the isothermal Euler–Poisson system with arbitrarily large data*, J. Differential Equations, 123 (1995), pp. 93–121.
[42] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
[43] M. E. Taylor, *Partial Differential Equations* I. *Basic Theory*, Springer-Verlag, New York, 1996.
[44] Y. Zeng, *Gas dynamics in thermal nonequilibrium and general hyperbolic systems with relaxation*, Arch. Ration. Mech. Anal., 150 (1999), pp. 225–279.
[45] B. Zhang, *Convergence of the Godunov scheme for a simplified one-dimensional hydrodynamic model for semiconductor*, Comm. Math. Phys., 157 (1993), pp. 1–22.
[46] B. Zhang, *Global weak solutions of the Cauchy problem to a hydrodynamic model for semiconductors*, J. Partial Differential Equations, 12 (1999), pp. 369–383.
[47] C. Zhu and H. Hattori, *Asymptotic behavior of the solution to a nonisentropic hydrodynamic model of semiconductors*, J. Differential Equations, 144 (1998), pp. 353–389.

SIAM J. MATH. ANAL.
Vol. 35, No. 2, pp. 423–452

DUBUC–DESLAURIERS SUBDIVISION FOR FINITE SEQUENCES AND INTERPOLATION WAVELETS ON AN INTERVAL*

J. M. DE VILLIERS†, K. M. GOOSEN†, AND B. M. HERBST‡

Abstract. In this paper we consider a method of adapting Dubuc–Deslauriers subdivision, which is defined for bi-infinite sequences, to accommodate sequences of finite length. After deriving certain useful properties of the Dubuc–Deslauriers refinable function on $\mathbb{R}$, we define a multiscale finite sequence of functions on a bounded interval, which are then proved to be refinable. Using this fact, the resulting adapted interpolatory subdivision scheme for finite sequences is then shown to be convergent. Corresponding interpolation wavelets on an interval are defined, and explicit formulations of the resulting decomposition and reconstruction algorithms are calculated. Finally, we give two numerical examples on signature smoothing and two-dimensional feature extraction of the subdivision and wavelet algorithms.

Key words. subdivision, interpolation, refinable functions, wavelets, bounded interval

AMS subject classifications. 65D10, 65T60, 42C40, 41A05

DOI. 10.1137/S0036141001386830

1. Introduction.

Consider the following simple iterative procedure: Let M be the linear space of bi-infinite vector-valued sequences $c = \{c_j\} = \{c_j : j \in \mathbb{Z}\}$, and begin with an ordered sequence $c^{(0)} \in M$, called the *initial control points*. From these points, generate a new sequence $c^{(1)}$ of control points in M where the even-indexed new control points interpolate the old ones. In contrast to standard interpolation procedures, subdivision schemes generate the new points by taking a linear combination of the old control points. For example, if one generates the new points using

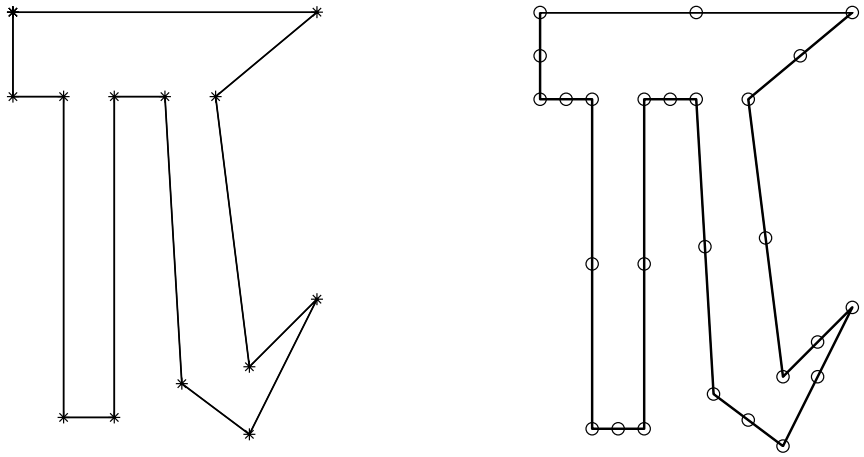
$$c^{(1)}_{2j} = c^{(0)}_j \quad \text{and} \quad c^{(1)}_{2j+1} = \tfrac{1}{2}(c^{(0)}_j + c^{(0)}_{j+1}), \quad j \in \mathbb{Z}, \tag{1.1}$$

then the odd-indexed new points are generated halfway between the old ones. This step can of course be repeated indefinitely, roughly "doubling" the number of points at each step. In this case the points fill in or converge to the straight lines connecting the initial control points (see Figure 1). Thus we obtain a continuous piecewise linear curve. In general, the existence and smoothness of this limit curve depend on the choice of the coefficients of the linear combination. The initial task is to find suitable choices for these coefficients which (i) ensure convergence to a limit curve, and (ii) yield appropriate smoothness of the limit curve.

In general, given a finitely supported real-valued sequence a, we define the corresponding subdivision operator $S : M \to M$ by

$$(Sc)_j = \sum_k a_{j-2k} c_k, \quad j \in \mathbb{Z}, \quad c \in M. \tag{1.2}$$

*Received by the editors March 23, 2001; accepted for publication (in revised form) September 6, 2002; published electronically July 18, 2003.
http://www.siam.org/journals/sima/35-2/38683.html
†Department of Mathematics, University of Stellenbosch, Private Bag X1, Matieland, 7602, Stellenbosch, South Africa (jmdv@sun.ac.za, karin@goose.sun.ac.za).
‡Department of Applied Mathematics, University of Stellenbosch, Private Bag X1, Matieland, 7602, Stellenbosch, South Africa (herbst@ibis.sun.ac.za).



(a) *The original control points* ($\star$). (b) *The new control points* ($\circ$) *after one iteration.*

FIG. 1. *Illustration of the iterative procedure* (1.1).

Unless stated otherwise we sum over all the integers. The resulting subdivision scheme is then defined, for a given initial sequence $c \in M$, by

$$c^{(0)} = c, \quad c^{(r+1)} = Sc^{(r)}, \quad r = 0, 1, \ldots, \tag{1.3}$$

or, equivalently,

$$c^{(0)} = c, \quad c^{(r)} = S^r c, \quad r = 1, \ldots. \tag{1.4}$$

The sequence $a = \{a_j\}$ is called the *mask* of the subdivision scheme.

There is no unique or best way of obtaining a mask. One possibility is to demand that if the original control points fall on a polynomial of degree $2N + 1$, then the newly generated control points must also lie on the same polynomial. This is, in fact, the idea behind Dubuc–Deslauriers subdivision [15, 12], which is a symmetric interpolatory scheme. In section 2.1, we develop this idea to eventually calculate an explicit formulation of the resulting mask sequence $\{a_j\}$.

To investigate the convergence of Dubuc–Deslauriers subdivision, we first refer to a result by Micchelli [24], according to which there exists a compactly supported fundamental interpolant $\phi \in C(\mathbb{R})$ which is refinable with respect to the sequence $\{a_j\}$. We then deduce, in Theorem 2.1, further properties of ϕ, including polynomial reproduction and information on its zeros, which prove to be necessary for our arguments in the subsequent two sections. It can then be shown as in Theorem 2.2 below that Dubuc–Deslauriers subdivision converges in the sense that, at each step of the iteration, the subdivision sequence lies entirely on the limit curve.

The Dubuc–Deslauriers subdivision operator has the form (1.2) and is therefore formulated for bi-infinite sequences. In section 3 we consider the case of finite sequences and propose a method to adapt Dubuc–Deslauriers subdivision to this situation. We base our construction on a multiscale finite sequence of fundamental interpolants $\{\phi_j^r\}$ defined on a bounded interval. Away from the boundaries the integer shifts of the original functions ϕ suffice, while the adjustments in the proximity of the boundaries preserve the polynomial reproduction property of ϕ. The resulting

adapted interpolatory subdivision scheme, which corresponds to Dubuc–Deslauriers subdivision away from the boundaries, then also preserves the above-mentioned polynomial filling property. In fact, most results derived in this paper depend on these polynomial filling or polynomial reproduction properties.

Our main result, as proved in Theorem 3.2, is that, on each fixed level r, the sequence $\{\phi_j^r\}$ is refinable in the sense that the function ϕ_j^r can be expressed as a finite linear combination of the (finite) sequence $\{\phi_j^{r+1}(2\,\bullet)\}$. Using this result, it is then a simple matter to prove, in Theorem 3.3, that our adapted Dubuc–Deslauriers subdivision for finite sequences converges. We end section 3 by calculating an explicit formulation for this adapted scheme.

The refinability of the sequence $\{\phi_j^r\}$ allows a multiresolutional construction of interpolation wavelets on an interval, and we pursue this idea in section 4. The main result of this section is the direct sum (nonorthogonal) space decomposition of Theorem 4.1, by virtue of which we obtain compactly supported symmetric interpolation wavelets. By giving up orthogonality we describe here the construction of symmetric interpolation wavelets resulting in finite decomposition and reconstruction formulas with rational coefficients given by the values of certain Lagrange polynomials at half integers. Also, a connection between interpolation wavelet decomposition on an interval and the adapted subdivision of section 3 is explained.

Alternative approaches to the construction of wavelets on an interval include work by Daubechies [11, section 10.7] and Cohen, Daubechies, and Vial [9], in which periodization and related methods are used for the construction of orthonormal wavelet bases on an interval. In the spline setting, explicit constructions of symmetric biorthogonal spline wavelets on an interval, as well as the corresponding decomposition and reconstruction algorithms, appear in Chui and Quak [8], Quak and Weyrich [25], Chui and de Villiers [6], and Chui [5, section 7.3.2].

The connection between the compactly supported orthonormal wavelets of Daubechies [10, 11] and Dubuc–Deslauriers subdivision has been noted by several authors (see, e.g., [24, section 3]) and exploited for the construction of biorthogonal interpolatory wavelets by Beylkin and Saito [4] and Bertoluzza and Naldi [2, 3]. A further study of the relationship between interpolation processes and wavelets construction appears in the paper [21] by Lee, Sharma, and Tan.

The idea, as used here in section 4, to construct interpolation wavelets by means of a nonorthogonal linear space decomposition and an interpolation operator has been studied for wavelets on $\mathbb{R}$ by Chui and Li [7] and used in the context of the Sweldens lifting scheme in [27].

Our interpolation wavelet decomposition and reconstruction algorithms for finite data sets, as given by (4.32), (4.33), and (3.34)–(3.38) below, are identical to those derived by Aràndiga, Donat, and Harten [1, equations (55) and (56)]. Making use of ideas developed by Harten [17, 18], these authors derive their equations from a general framework—our approach has the advantage that it allows an explicit construction of the underlying refinable sequence $\{\phi_j^r\}$ as demonstrated by our equation (3.12). This is perhaps closer to the unpublished work of Donoho [13] in the sense that both are based on the principle of polynomial extrapolation. An essential difference, however, lies in the way in which the associated nested sequence of linear spaces $\{V_r\}$ is defined. While Donoho's construction is consistently based on a polynomial extrapolation operator, we define the linear space V_r as the span of the sequence $\{\phi_j^r\}$. The nesting property of the $\{V_r\}$ then follows from the refinability of $\{\phi_j^r\}$, as proved in Theorem 3.2 below.

Note that the wavelets constructed in this manner do not have vanishing moments —the mean of our wavelet is not zero. In this regard our wavelets are similar to those mentioned by Mallat [22, p. 301]. We show, however, that the corresponding interpolation wavelet space can be characterized in terms of a projection operator which is exact on polynomials, thereby causing the interpolation wavelet coefficients of a function to be relatively small in those regions where the function exhibits local polynomial-like behavior. For the construction of (nonorthogonal) interpolation wavelets on the interval with vanishing moments, we refer to Donoho [14].

In [26], Schröder and Sweldens develop algorithms implementing interpolation wavelets on an interval without providing detailed proofs.

The main algorithms are illustrated in the final section using examples from signature verification and two-dimensional image processing.

2. Dubuc–Deslauriers subdivision for bi-infinite sequences. In this section we introduce Dubuc–Deslauriers subdivision as an optimally local curve filling iterative procedure which reproduces polynomials of a given odd degree. We next prove the existence of an associated refinable function, which is then shown to provide a limit curve for the Dubuc–Deslauriers subdivision scheme.

2.1. Construction of the mask. For a given nonnegative integer N, consider the problem of finding a minimally supported mask a such that the $(2N+1)$th degree polynomial filling property

$$\sum_k a_{j-2k} p(k) = p\left(\frac{j}{2}\right), \quad j \in \mathbb{Z}, \quad p \in \pi_{2N+1}, \tag{2.1}$$

holds. Here π_{2N+1} denotes the linear space of polynomials of degree $\leq 2N+1$.

For this purpose we introduce the Lagrange fundamental polynomials $L_k \in \pi_{2N+1}$, $k = -N, \ldots, N+1$, as defined by

$$L_k(x) = \prod_{k \neq j = -N}^{N+1} \frac{x-j}{k-j}, \quad x \in \mathbb{R}, \quad k = -N, \ldots, N+1, \tag{2.2}$$

for which

$$L_k(j) = \delta_{k,j}, \quad k, j = -N, \ldots, N+1, \tag{2.3}$$

and

$$\sum_{k=-N}^{N+1} p(k) L_k(x) = p(x), \quad x \in \mathbb{R}, \quad p \in \pi_{2N+1}. \tag{2.4}$$

Setting $j = 0$ and $j = 1$ in (2.1), and using (2.3) and (2.4), the equation (2.1) implies

$$\left.\begin{aligned} a_{2j} + \sum_{k \notin \{-N,\ldots,N+1\}} a_{2k} L_{-j}(k) &= \delta_{j,0}, \\ a_{2j+1} + \sum_{k \notin \{-N,\ldots,N+1\}} a_{1-2k} L_{-j}(k) &= L_{-j}\left(\tfrac{1}{2}\right), \end{aligned}\right\} \; j = -N-1, \ldots, N. \tag{2.5}$$

A necessary condition for a minimally supported mask $a = \{a_j\}$ to satisfy (2.1) is thus

$$a_{2j} = \delta_{j,0}, \qquad j \in \mathbb{Z}, \tag{2.6a}$$

$$a_{2j+1} = L_{-j}\left(\tfrac{1}{2}\right), \quad j = -N-1, \dots, N, \tag{2.6b}$$

$$a_{2j+1} = 0, \qquad j \geq N+1 \text{ or } j \leq -N-2. \tag{2.6c}$$

The choice (2.6) is also sufficient to fulfill (2.1). In fact, if $j = 2m,\ \ m \in \mathbb{Z}$, then, for $p \in \pi_{2N+1}$, (2.6a) implies

$$\sum_k a_{j-2k} p(k) = \sum_k a_{2m-2k} p(k) = \sum_k a_{2k} p(m-k) = p(m) = p\left(\frac{j}{2}\right),$$

whereas if $j = 2m+1,\ m \in \mathbb{Z}$, then (2.6b), (2.6c), and (2.4) give

$$\sum_k a_{j-2k} p(k) = \sum_k a_{2k+1} p(m-k) = \sum_{k=-N-1}^{N} L_{-k}\left(\tfrac{1}{2}\right) p(m-k) = p\left(m+\tfrac{1}{2}\right) = p\left(\frac{j}{2}\right).$$

Hence the subdivision scheme corresponding to the mask (2.6), which was introduced by Deslauriers and Dubuc in [12], is indeed a minimally supported mask sequence $a = \{a_j\}$ for which (2.1) holds.

Observe that (1.2) implies, with reference to (1.3), the interpolatory property

$$c_{2j}^{(r+1)} = c_j^{(r)}, \quad j \in \mathbb{Z}, \ \ r = 0, 1, \dots, \tag{2.7}$$

by virtue of which Dubuc–Deslauriers subdivision is called an *interpolatory* scheme.

We now derive an explicit expression for the mask. Since, for $k \in \{-N, \dots, N+1\}$, we have

$$\prod_{k \neq j = -N}^{N+1} \left(\tfrac{1}{2} - j\right) = \frac{1}{2^{2N+1}} \, \frac{1}{1-2k} \prod_{j=-N-1}^{N} (2j+1) = \frac{(-1)^N}{2^{4N+1}} \frac{1}{2k-1} \left[\frac{(2N+1)!}{N!}\right]^2$$

and

$$\prod_{k \neq j = -N}^{N+1} (k-j) = (-1)^{N+1+k} (N+k)!(N+1-k)!,$$

we deduce from (2.6) and (2.2) that the Dubuc–Deslauriers mask sequence $a = \{a_j\}$ has the explicit formulation

$$\left.\begin{aligned} a_{2j+1} &= \frac{N+1}{2^{4N+1}} \binom{2N+1}{N} \frac{(-1)^j}{2j+1} \binom{2N+1}{N+j+1}, && j = -N-1, \dots, N, \\ a_{2j} &= \delta_{j,0}, && j = -N, \dots, N, \\ a_j &= 0, && |j| \geq 2N+2. \end{aligned}\right\} \tag{2.8}$$

For example, if $N = 1$, (2.8) gives

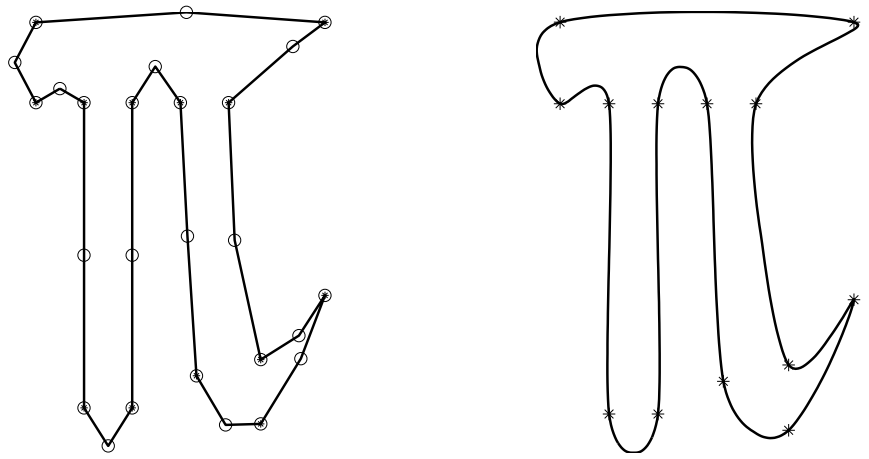
$$a_{2j+1} = \begin{cases} -\frac{1}{16}, & j = -2, \\ \frac{9}{16}, & j = -1, \\ \frac{9}{16}, & j = 0, \\ -\frac{1}{16}, & j = 1, \\ 0 & \text{otherwise.} \end{cases}$$

From (2.8) observe that the mask coefficients are symmetric, i.e.,

$$a_j = a_{-j}, \quad j \in \mathbb{Z}. \tag{2.9}$$

Therefore we call Dubuc–Deslauriers subdivision a *symmetric* scheme.

For $N = 0$, note that (2.8), (1.2), and (1.3) yield, for $r = 0$, the iteration procedure in (1.1). This subdivision scheme converges to a continuous piecewise linear function which interpolates the original control points (see Figure 1). In contrast, subdivision with the mask obtained by setting $N = 1$ in (2.8) converges to a smoother function, while still interpolating the original control points (see Figure 2). Incidentally, this example is not completely arbitrary. In fact Knuth based his construction of TEX fonts on ideas remarkably similar to subdivision schemes [19, Chapter 2], more than 10 years before the Dubuc–Deslauriers scheme was introduced in [12].



(a) *Original control points* ($\star$) *and one step of the subdivision algorithm* ($\circ$).

(b) *Original control points with limit curve.*

FIG. 2. *Illustration of Dubuc–Deslauriers subdivision for* $N = 1$.

We now show that, for any given initial sequence $c \in M$, the Dubuc–Deslauriers subdivision sequence $\{c^{(r)} : r = 0, 1, \dots\} \subset M$ converges to a limit curve which will be specified in terms of an associated refinable function.

2.2. The corresponding refinable function. Before investigating the convergence of Dubuc–Deslauriers subdivision, we establish the existence and properties of

the corresponding refinable function, following an approach employed by Micchelli in [24] (see also [16]).

THEOREM 2.1. *For any given nonnegative integer N, let the sequence $a = \{a_j\}$ be chosen as in* (2.6). *Then there exists a compactly supported function $\phi \in C(\mathbb{R})$ such that*

$$\phi(x) = \sum_j a_j \phi(2x - j), \quad x \in \mathbb{R}; \tag{2.10}$$

$$\phi(j) = \delta_{j,0}, \quad j \in \mathbb{Z}; \tag{2.11}$$

$$\phi(x) = 0, \quad x \notin (-2N - 1, 2N + 1); \tag{2.12}$$

$$\sum_j p(j)\phi(x - j) = p(x), \quad x \in \mathbb{R}, \quad p \in \pi_{2N+1}; \tag{2.13}$$

$$\phi(x) = \phi(-x), \quad x \in \mathbb{R}; \tag{2.14}$$

$$\phi\left(\frac{j}{2}\right) = a_j, \quad j \in \mathbb{Z}; \tag{2.15}$$

$$\phi(2N + 1 - 2^{-j}(N - \tfrac{1}{2} - k)) = 0, \quad k = 0, 1, \dots, \quad j = 0, 1, \dots. \tag{2.16}$$

Proof. The existence of a compactly supported function of $\phi \in C(\mathbb{R})$ satisfying the properties (2.10), (2.11), and (2.12) was proven in [24, Lemma 3.1, Theorem 4.1, and Corollary 4.1].

To prove (2.13), suppose $\ell \in \{0, 1, \dots, 2N + 1\}$, $k \in \mathbb{Z}$, and $r \in \{0, 1, \dots\}$. We shall prove that

$$\sum_j j^\ell \phi\left(\frac{k}{2^r} - j\right) = \left(\frac{k}{2^r}\right)^\ell, \tag{2.17}$$

which then implies (2.13), since the set $\{\frac{k}{2^r} : k \in \mathbb{Z}, \quad r = 0, 1, \dots\}$ is dense in $\mathbb{R}$, and since ϕ is a compactly supported continuous function on $\mathbb{R}$.

Noting that (2.17) is an immediate consequence of (2.11) if $r = 0$, we assume next that $r \geq 1$. Then, using consecutively (2.10), (2.1), and (2.11), we get

$$\begin{aligned}
\sum_j j^\ell \phi\left(\frac{k}{2^r} - j\right) &= \sum_j j^\ell \sum_m a_m \phi\left(\frac{k}{2^{r-1}} - 2j - m\right) \\
&= \sum_m \left[\sum_j a_{m-2j} j^\ell\right] \phi\left(\frac{k}{2^{r-1}} - m\right) \\
&= \frac{1}{2^\ell} \sum_m m^\ell \phi\left(\frac{k}{2^{r-1}} - m\right) \\
&\;\;\vdots \\
&= \frac{1}{2^{\ell r}} \sum_m m^\ell \phi(k - m) = \left(\frac{k}{2^r}\right)^\ell.
\end{aligned}$$

Similarly, the property (2.14) will be proved if we can show that, for $k \in \mathbb{Z}$ and $r \in \{0, 1, \dots\}$,

$$\phi\left(\frac{k}{2^r}\right) = \phi\left(-\frac{k}{2^r}\right). \tag{2.18}$$

For $r = 0$, (2.18) follows from (2.11), whereas for $r = 1$, it follows from (2.10), (2.9), and (2.11) that $\phi(-\frac{k}{2}) = \sum_j a_j \phi(-k - j) = \sum_j a_j \phi(-k + j) = \sum_j a_j \phi(k - j) = \phi(\frac{k}{2})$. If $r \geq 2$, we also use (1.2) to deduce that

$$\begin{aligned}
\phi\left(-\frac{k}{2^r}\right) &= \sum_j a_j \phi\left(-\frac{k}{2^{r-1}} - j\right) \\
&= \sum_j a_j \phi\left(-\frac{k}{2^{r-1}} + j\right) \\
&= \sum_j a_j \sum_\ell a_{-\ell} \phi\left(-\frac{k}{2^{r-2}} + 2j - \ell\right) \\
&= \sum_\ell \left[\sum_j a_{\ell-2j} a_j\right] \phi\left(-\frac{k}{2^{r-2}} + \ell\right) \\
&= \sum_\ell (Sa)_\ell \phi\left(-\frac{k}{2^{r-2}} + \ell\right) \\
&\;\;\vdots \\
&= \sum_\ell \left(S^{r-1}a\right)_\ell \phi(-k + \ell) = (S^{r-1}a)_k. \qquad (2.19)
\end{aligned}$$

A similar strategy shows that

$$\phi\left(\frac{k}{2^r}\right) = (S^{r-1}a)_k,$$

which, together with (2.19), proves (2.18) for $r \geq 2$.

Property (2.15) is an immediate consequence of (2.10) and (2.11).

Finally, we prove (2.16) by induction. For $j = 0$, property (2.16) follows from (2.15) and (2.6c). To advance the inductive hypothesis from j to $j + 1$, we use (2.10) and (2.6c) to deduce that, for $k \in \{0, 1, \dots\}$,

$$\begin{aligned}
\phi(2N + 1 - 2^{-(j+1)}(N - \tfrac{1}{2} - k)) &= \sum_\ell a_\ell \phi(4N + 2 - 2^{-j}(N - \tfrac{1}{2} - k) - \ell) \\
&= \sum_{\ell=0}^{4N+2} a_{2N+1-\ell} \phi(2N + 1 - 2^{-j}(N - \tfrac{1}{2} - [k + 2^j \ell])).
\end{aligned}$$

Using (2.12) we complete the proof. □

Remarks.

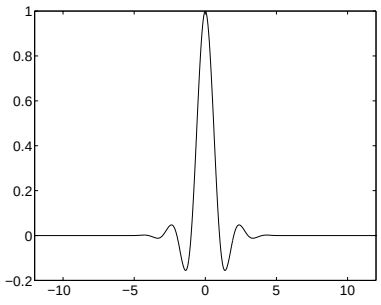
1. Equations (2.10) and (2.15) imply

$$\phi(x) = \sum_j \phi\left(\frac{j}{2}\right) \phi(2x - j), \quad x \in \mathbb{R}. \tag{2.20}$$

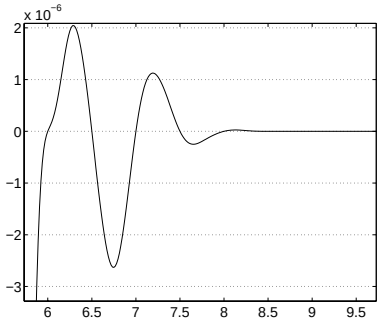
2. Equations (2.14) and (2.16) imply

$$\phi(-2N-1+2^{-j}(N-\tfrac{1}{2}-k))=0, \quad k=0,1,\ldots, \quad j=0,1,\ldots. \tag{2.21}$$

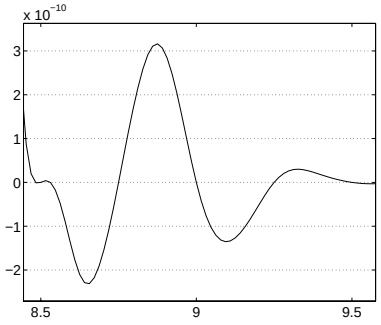
3. For $k \geq N$, (2.16) and (2.21) hold because the argument falls outside the support of ϕ, while for $k = 0, 1, \ldots, N-1$ the argument falls within the support. This in turn implies that ϕ has an infinite number of zeros within its support and that these zeros are clustered more densely toward the edges of the support, as illustrated in Figure 3.
4. A detailed study of the regularity of ϕ as a function of N is given in [12].



(a) ϕ *for* $N = 5$.



(b) *Magnification of* ϕ.



(c) *Further magnification of* ϕ.

FIG. 3. *Illustration of the refinable function with clustered zeros.*

2.3. Convergence of the subdivision scheme. We now show that the Dubuc–Deslauriers subdivision scheme converges in the sense that, for each $r \in \{0, 1, \ldots, \}$, the sequence $c^{(r)}$ lies entirely on the limit curve.

THEOREM 2.2. *For each initial sequence* $c = \{c_j\} \in M$, *the Dubuc–Deslauriers subdivision scheme* (1.3), (1.2), *and* (2.6) *converges to the function*

$$f(x) = \sum_j c_j \phi(x-j), \quad x \in \mathbb{R}, \tag{2.22}$$

with ϕ *defined as in Theorem* 2.1, *in the sense that*

$$c_k^{(r)} = f\left(\frac{k}{2^r}\right), \quad k \in \mathbb{Z}, \quad r = 0, 1, \ldots. \tag{2.23}$$

Proof. If $r = 0$, then (2.23) follows from (2.11). For $r \geq 1$, we use (2.10), (1.2), (1.3), and (2.11) to deduce, for $k \in \mathbb{Z}$,

$$\begin{aligned} f\left(\frac{k}{2^r}\right) &= \sum_j c_j \phi\left(\frac{k}{2^r} - j\right) \\ &= \sum_j c_j \sum_\ell a_{\ell-2j} \phi\left(\frac{k}{2^{r-1}} - \ell\right) \\ &= \sum_\ell (Sc)_k \phi\left(\frac{k}{2^{r-1}} - \ell\right) \\ &\;\;\vdots \\ &= \sum_\ell (S^r c)_\ell \phi(k-\ell) = (S^r c)_k = c_k^{(r)}, \end{aligned}$$

the last equality by virtue of (1.4). ∎

3. A modified subdivision scheme for finite sequences. The algorithms for bi-infinite sequences, as described in the previous sections, are applied mainly in the case of periodic sequences. For finitely supported sequences these algorithms must be modified to accommodate the boundaries. We consider here a method of adapting the Dubuc–Deslauriers subdivision scheme of section 2 to the situation where the initial sequence c is finite.

3.1. Construction of a modified scheme. We first show that refinable functions defined on an interval allow one to construct a subdivision scheme for finite sequences. We derive the specific properties of these refinable functions from those of the refinable (on $\mathbb{R}$) function ϕ of section 2, with appropriate modifications near the boundaries.

With N as in section 2, assume n is a positive integer with $n \geq 4N + 2$, and let r be a nonnegative integer. On the basis of Theorem 2.1, and the subsequent equation (2.20), we seek to construct a sequence $\{\phi_j^r\} = \{\phi_j^r : j = 0, 1, \ldots, 2^r n, \quad r = 0, 1, \ldots\}$ such that, for each fixed r,

$$\phi_j^r \in C[0, 2^r n], \quad j = 0, 1, \ldots, 2^r n; \tag{3.1}$$

$$\phi_j^r(x) = 0, \; x \notin \begin{cases} [0, 2N+1+j), & j = 0, 1, \ldots, 2N+1, \\ (-2N-1+j, 2N+1+j), & j = 2N+2, \ldots, 2^r n - 2N - 2, \\ (-2N-1+j, 2^r n], & j = 2^r n - 2N - 1, \ldots, 2^r n; \end{cases} \tag{3.2}$$

$$\phi_j^r(k) = \delta_{j,k}, \quad j, k = 0, 1, \ldots, 2^r n; \tag{3.3}$$

$$\sum_{j=0}^{2^r n} p(j) \phi_j^r(x) = p(x), \quad x \in [0, 2^r n], \quad p \in \pi_{2N+1}; \tag{3.4}$$

$$\phi_j^r(x) = \sum_{k=0}^{2^{r+1} n} \phi_j^r\left(\frac{k}{2}\right) \phi_k^{r+1}(2x), \quad x \in [0, 2^r n], \quad j = 0, 1, \ldots, 2^r n. \tag{3.5}$$

Denoting the linear space of vector-valued sequences

$$c = \{c_j\} = \{c_j : j = 0, 1, \dots, 2^r n\}$$

by M_r, the subdivision operator sequence $\{S_r : r = 0, 1, \dots, \dots\}$ for $S_r : M_r \to M_{r+1}$ is then defined by

$$(S_r c)_j = \sum_{k=0}^{2^r n} \phi_k^r \left(\frac{j}{2}\right) c_k, \quad j = 0, 1, \dots, 2^{r+1} n. \tag{3.6}$$

The corresponding subdivision scheme is defined by

$$c^{(0)} = c \in M_0, \quad c^{(r+1)} = S_r c^{(r)}, \quad r = 0, 1, \dots, \tag{3.7}$$

or, equivalently,

$$c^{(0)} = c, \quad c^{(r+1)} = S_r(\dots(S_1(S_0 c))\dots), \quad r = 0, 1, \dots. \tag{3.8}$$

Observe that (3.3), (3.6), and (3.7) imply

$$c_{2j}^{(r+1)} = c_j^{(r)}, \quad j = 0, 1, \dots, 2^r n, \quad r = 0, 1, \dots, \tag{3.9}$$

i.e., the subdivision scheme (3.6)–(3.7) is interpolatory, whereas (3.4) implies

$$\sum_{k=0}^{2^r n} \phi_k^r \left(\frac{j}{2}\right) p(k) = p\left(\frac{j}{2}\right), \quad j = 0, 1, \dots, 2^{r+1} n, \quad p \in \pi_{2N+1}, \tag{3.10}$$

according to which the subdivision scheme (3.6)–(3.7) has the $(2N+1)$th degree polynomial filling property. In (3.6)–(3.10) we have obtained the analogues of, respectively, (1.2), (1.3), (1.4), (2.7), and (2.1).

To find a sequence $\{\phi_j^r\}$ satisfying (3.1)–(3.5), we observe from (2.12) and (2.13) that

$$\sum_{j=0}^{2^r n} p(j)\phi(x-j) = \sum_j p(j)\phi(x-j) = p(x), \quad x \in [2N, 2^r n - 2N], \quad p \in \pi_{2N+1}. \tag{3.11}$$

Thus, the sequence $\{\phi(\cdot - j)\}$ provides suitable refinable functions away from the boundaries. The boundary modifications can again be based on property (2.1). Using arguments similar to the ones which led to the construction of the mask (2.6), we define, for each fixed $r \in \{0, 1, \dots\}$, the sequence $\{\phi_j^r\}$ on the interval $[0, 2^r n]$ by

$$\phi_j^r(x) = \phi(x-j) + \sum_{k=-2N}^{-1} L_{j-N}(k-N)\phi(x-k), \quad j = 0, 1, \dots, 2N+1, \tag{3.12a}$$

$$\phi_j^r(x) = \phi(x-j), \quad j = 2N+2, \dots, 2^r n - 2N - 2, \tag{3.12b}$$

$$\phi_j^r(x) = \phi_{2^r n - j}^r(2^r n - x), \quad j = 2^r n - 2N - 1, \dots, 2^r n, \tag{3.12c}$$

where the sequence $\{L_k : k = -N, \dots, N+1\} \subset \pi_{2N+1}$ is given by (2.2). Examples are plotted in Figure 4.

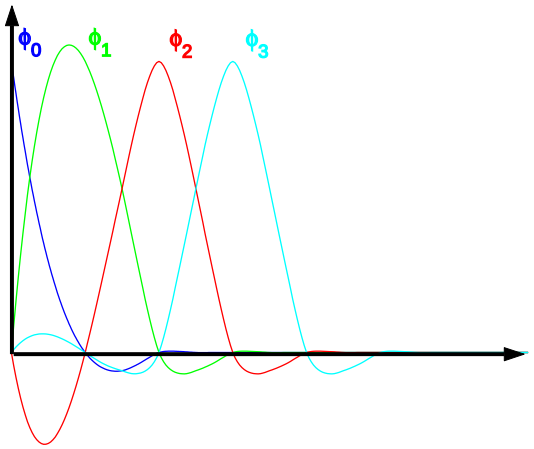


FIG. 4. *Illustration of definition* (3.12) *with* $N = 1$.

Using

$$\phi\left(N + \frac{3}{2} + k\right) = 0, \quad k = 0, 1, 2, \ldots, \tag{3.13}$$

and

$$\phi\left(-N - \frac{3}{2} - k\right) = 0, \quad k = 0, 1, 2, \ldots, \tag{3.14}$$

which follow from (2.16) and (2.21), we prove the following useful properties of the sequence $\{\phi_j^{(r)}\}$.

PROPOSITION 3.1. *For given nonnegative integers* N *and* r*, let* n *be an integer with* $n \geq 4N + 2$*. Then the sequence* $\{\phi_j^r\}$*, as defined by* (3.12)*, satisfies*

$$\phi_j^r(x) = \phi(x - j), \qquad x \in [2N, 2^r n - 2N], \quad j = 0, 1, \ldots, 2^r n, \tag{3.15}$$

$$\phi_j^r(x) = L_{j-N}(x - N), \quad x \in [0, 1], \quad j = 0, \ldots, 2N + 1, \tag{3.16}$$

$$\phi_j^r\left(\frac{k}{2}\right) = \begin{cases} L_{j-N}(\frac{k}{2} - N), & k = 0, \ldots, 2N + 1, \\ \phi(\frac{k}{2} - j), & k = 2N, \ldots, \end{cases} \qquad j = 0, \ldots, 2N + 1. \tag{3.17}$$

Proof. For $j = 0, \ldots, 2^r n - 2N - 2$, (3.15) follows from the definition (3.12) and the finite support (2.12) of ϕ. It therefore suffices to prove (3.15) for $j = 2^r n - 2N - 1, \ldots, 2^r n$. But then, from (3.12c), and (3.12a),

$$\phi_j^r(x) = \phi_{2^r n - j}^r(2^r n - x) = \phi(-x + j) = \phi(x - j),$$

by virtue of (2.14).

To prove (3.16), suppose $x \in [0, 1]$ and $j \in \{0, 1, \ldots, 2N+1\}$. Then (3.12a), (2.3),

and (2.12) yield

$$\phi_j^r(x) = \sum_{k=-2N}^{2N+1} L_{j-N}(k-N)\phi(x-k)$$
$$= \sum_k L_{j-N}(k-N)\phi(x-k) = L_{j-N}(x-N),$$

by virtue of (2.13).

For the proof of (3.17), we first let $k \in \{0,1,\dots,2N+1\}$ and $j \in \{0,1,\dots,2N+1\}$. Then, using consecutively (3.12a), (2.3), (3.14), (2.12), and (2.13), we get

$$\phi_j^r\left(\frac{k}{2}\right) = \phi\left(\frac{k}{2}-j\right) + \sum_{\ell=-2N}^{-1} L_{j-N}(\ell-N)\phi\left(\frac{k}{2}-\ell\right) \tag{3.18}$$
$$= \sum_{\ell=-2N}^{2N+1} L_{j-N}(\ell-N)\phi\left(\frac{k}{2}-\ell\right)$$
$$= \sum_{\ell=-2N}^{3N+1} L_{j-N}(\ell-N)\phi\left(\frac{k}{2}-\ell\right)$$
$$= \sum_{\ell} L_{j-N}(\ell-N)\phi\left(\frac{k}{2}-\ell\right)$$
$$= L_{j-N}\left(\frac{k}{2}-N\right). \tag{3.19}$$

Next, for $k \in \{2N+2,\dots\}$ and $j \in \{0,1,\dots,2N+1\}$, we see that (3.18) holds again, and the bottom part of (3.17) therefore follows since $\phi(\frac{k}{2}-\ell) = 0,\ \ell = -2N,\dots,-1$, by virtue of (2.11) and (3.13).

To complete the proof of (3.17), it remains to show that for $j \in \{0,1,\dots,2N+1\}$, we have

$$\phi\left(\frac{k}{2}-j\right) = L_{j-N}\left(\frac{k}{2}-N\right) \quad \text{for} \quad k = 2N \text{ and } k = 2N+1. \tag{3.20}$$

For $k = 2N$, the property (3.20) is a consequence of (2.11) and (2.3), whereas for $k = 2N+1$ we use (2.15) and (2.6b) to deduce that

$$\phi\left(\frac{k}{2}-j\right) = a_{2N+1-2j} = L_{j-N}\left(\frac{1}{2}\right) = L_{j-N}\left(\frac{k}{2}-N\right). \qquad \square$$

Remark. Observe from (3.15) and (3.12a), (3.12b), together with the support properties (3.2) and (2.12), that

$$\phi_j^r(x) = \phi(x-j), \quad x \in [2N, 2^r n], \quad j = 0,1,\dots,2^r n - 2N - 2. \tag{3.21}$$

3.2. The refinability of the sequence $\{\phi_j^r\}$. Analogous to the bi-infinite case, our proof in section 3.3 below of the convergence of the subdivision scheme will depend on the refinability of the sequence $\{\phi_j^r\}$, as proved in this section.

THEOREM 3.2. *The sequence $\{\phi_j^r\}$, as defined in* (3.12)*, satisfies the properties* (3.1)–(3.5).

Proof. The properties (3.1)–(3.3) are immediate consequences of (2.11) and (2.12).

To prove (3.4), we choose $p \in \pi_{2N+1}$ and assume first that $x \in [0, 2N)$. Using (3.2), (3.12a), (3.12b), (2.4), (2.12), and (2.13) consecutively, one has

$$\begin{aligned}\sum_{j=0}^{2^r n} p(j)\phi_j^r(x) &= \sum_{j=0}^{4N} p(j)\phi_j^r(x) \\ &= \sum_{j=0}^{4N} p(j)\phi(x-j) + \sum_{k=-2N}^{-1}\left[\sum_{j=0}^{2N+1} p(j)L_{j-N}(k-N)\right]\phi(x-k) \\ &= \sum_{j=-2N}^{4N} p(j)\phi(x-j) = \sum_{j} p(j)\phi(x-j) = p(x).\end{aligned}$$

For $x \in [2N, 2^r n - 2N]$, the property (3.4) follows from (3.15) and (3.11). Similar arguments establish polynomial reproduction for $x \in (2^r n - 2N, 2^r n]$.

To prove the refinability (3.5), assume first that $j \in \{0, 1, \ldots, 2N+1\}$ and $x \in [0, 2N)$. Then (3.12a) and (2.3) imply

$$\phi_j^r(x) = \sum_{k=-2N}^{2N+1} L_{j-N}(k-N)\phi(x-k).$$

Now, use the refinement equation (2.20), together with the support property (2.12), to obtain

$$\phi_j^r(x) = \sum_{\ell=-2N}^{6N}\left[\sum_{k=-2N}^{2N+1} L_{j-N}(k-N)\phi\left(\frac{\ell}{2}-k\right)\right]\phi(2x-\ell). \tag{3.22}$$

For the first part of the sum in (3.22) we get, from (2.11), (3.13), (3.14), the polynomial reproduction (2.13), and definition (3.12a),

$$\begin{aligned}&\sum_{\ell=-2N}^{2N+1}\left[\sum_{k=-2N}^{2N+1} L_{j-N}(k-N)\phi\left(\frac{\ell}{2}-k\right)\right]\phi(2x-\ell) \\ &= \sum_{\ell=-2N}^{2N+1}\left[\sum_{k} L_{j-k}(k-N)\phi\left(\frac{\ell}{2}-k\right)\right]\phi(2x-\ell) \\ &= \sum_{\ell=-2N}^{2N+1} L_{j-N}\left(\frac{\ell}{2}-N\right)\phi(2x-\ell) \\ &= \sum_{\ell=-2N}^{-1} L_{j-N}\left(\frac{\ell}{2}-N\right)\phi(2x-\ell) + \sum_{\ell=0}^{2N+1} L_{j-N}\left(\frac{\ell}{2}-N\right)\phi_\ell^{r+1}(2x) \\ &\qquad - \sum_{k=-2N}^{-1}\left[\sum_{\ell=0}^{2N+1} L_{j-N}\left(\frac{\ell}{2}-N\right)L_{\ell-N}(k-N)\right]\phi(2x-k) \\ &= \sum_{\ell=0}^{2N+1} L_{j-N}\left(\frac{\ell}{2}-N\right)\phi_\ell^{r+1}(2x) \;=\; \sum_{\ell=0}^{2N+1}\phi_j^r\left(\frac{\ell}{2}\right)\phi_\ell^{r+1}(2x),\end{aligned} \tag{3.23}$$

having also used the polynomial reproduction (2.4) and (3.17).

For the remaining part of the sum in (3.22), we use the interpolatory properties (2.3) and (2.11), as well as (3.13) and (3.17), to obtain

$$\begin{aligned}
&\sum_{\ell=2N+2}^{6N}\left[\sum_{k=-2N}^{2N+1} L_{j-N}(k-N)\phi\left(\frac{\ell}{2}-k\right)\right]\phi(2x-\ell)\\
&\quad=\sum_{\ell=2N+2}^{6N}\phi\left(\frac{\ell}{2}-j\right)\phi(2x-\ell)\\
&\quad=\sum_{\ell=2N+2}^{6N}\phi_j^r\left(\frac{\ell}{2}\right)\phi_\ell^{r+1}(2x)=\sum_{\ell=2N+2}^{2^{r+1}n}\phi_j^r\left(\frac{\ell}{2}\right)\phi_\ell^{r+1}(2x), \qquad (3.24)
\end{aligned}$$

where we also used the definition (3.12b) and the inequality $n \geq 4N+2$. Combining (3.22), (3.23), and (3.24) then yields the result (3.5) for $j \in \{0,1,\ldots,2N+1\}$ and $x \in [0,2N)$.

Next, for $j \in \{0,1,\ldots,2N+1\}$ and $x \in [2N, 2N+1+j)$, we use Proposition 3.1, Theorem 2.1, and (3.13) to get

$$\begin{aligned}
\phi_j^r(x)=\phi(x-j)&=\sum_k \phi\left(\frac{k}{2}-j\right)\phi(2x-k)\\
&=\sum_{k=2N}^{4N+2j+2}\phi\left(\frac{k}{2}-j\right)\phi(2x-k)\\
&=\sum_{k=2N}^{4N+2j+2}\phi_j^r\left(\frac{k}{2}\right)\phi_k^{r+1}(2x)\\
&=\sum_{k=0}^{2^{r+1}n}\phi_j^r\left(\frac{k}{2}\right)\phi_k^{r+1}(2x),
\end{aligned}$$

since $n \geq 4N+2$, thereby establishing (3.5) for this subcase.

For $j \in \{0,1,\ldots,2N+1\}$ and $x \in [2N+1+j, 2^r n]$, we deduce from (3.2), (3.17), (2.11), and (3.13) that

$$\begin{aligned}
\sum_{k=0}^{2^{r+1}n}\phi_j^r\left(\frac{k}{2}\right)\phi_k^{r+1}(2x)&=\sum_{k=2N+2+2j}^{4N+1+2j}\phi_j^r\left(\frac{k}{2}\right)\phi_k^{r+1}(2x)\\
&=\sum_{k=2N+2+2j}^{4N+1+2j}\phi\left(\frac{k}{2}-j\right)\phi_k^{r+1}(2x) \;=\; 0=\phi_j^r(x),
\end{aligned}$$

by virtue of the top part of (3.2). Hence we have established equation (3.5) for all $j \in \{0,1,\ldots,2N+1\}$.

Now consider the case $j \in \{2N+2,\ldots,2^r n-2N-2\}$ and $x \in (-2N-1+j,\ 2N+1+j)$. We use definition (3.12b), (2.20), (2.12), as well as (3.13) and the finite

support property (3.2), to find that

$$\begin{aligned}\phi_j^r(x) = \phi(x-j) &= \sum_{k=-2N-1+2j}^{2N+1+2j} \phi\left(\frac{k}{2}-j\right)\phi(2x-k)\\ &= \sum_{k=-2N-1+2j}^{2N+1+2j} \phi\left(\frac{k}{2}-j\right)\phi_k^{r+1}(2x)\\ &= \sum_{k=0}^{2^{r+1}n} \phi\left(\frac{k}{2}-j\right)\phi_k^{r+1}(2x)\\ &= \sum_{k=0}^{2^{r+1}n} \phi_j^r\left(\frac{k}{2}\right)\phi_k^{r+1}(2x),\end{aligned}$$

based on the inequalities $-2N-1+2j \geq 2N+3$ and $2N+1+2j \leq 2^{r+1}n-2N-3$.

For $j \in \{2N+2,\ldots,2^r n-2N-2\}$ and $x \in [0,-2N-1+j]$, we use (3.2), (3.12b), (2.11), and (3.14) to get

$$\sum_{k=0}^{2^{r+1}n} \phi_j^r\left(\frac{k}{2}\right)\phi_k^{r+1}(2x) \;=\; \sum_{k=0}^{-2N-2+2j} \phi\left(\frac{k}{2}-j\right)\phi_k^{r+1}(2x) \;=\; 0 \;=\phi_j^r(x).$$

Similarly, for $j \in \{2N+2,\ldots,2^r n-2N-2\}$ and $x \in [2N+1+j,\ 2^r n]$, with (3.13), we have

$$\sum_{k=0}^{2^{r+1}n} \phi_j^r\left(\frac{k}{2}\right)\phi_k^{r+1}(2x) \;=\; \sum_{k=2N+2+2j}^{2^{r+1}n} \phi\left(\frac{k}{2}-j\right)\phi_k^{r+1}(2x) \;=\; 0 \;=\phi_j^r(x).$$

Hence (3.5) also holds for $j \in \{2N+2,\ldots,2^r n-2N-2\}$.

Finally, let $j \in \{2^r n-2N-1,\ldots,2^r n\}$ and $x \in [0,2^r n]$. Then (3.12c), together with the fact that (3.5) holds for $j \in \{0,1,\ldots,2N+1\}$, gives

$$\begin{aligned}\phi_j^r(x) &= \phi_{2^r n-j}^r(2^r n-x)\\ &= \sum_{k=0}^{2^{r+1}n} \phi_{2^r n-j}^r\left(\frac{k}{2}\right)\phi_k^{r+1}\left(2^{r+1}n-2x\right)\\ &= \sum_{k=0}^{2^{r+1}n} \phi_{2^r n-j}^r\left(2^r n-\frac{k}{2}\right)\phi_{2^{r+1}n-k}^{r+1}(2^{r+1}n-2x)\\ &= \sum_{k=0}^{2^{r+1}n} \phi_j^r\left(\frac{k}{2}\right)\phi_{2^{r+1}n-k}^{r+1}(2^{r+1}n-2x).\end{aligned}$$

We claim that

$$\phi_{2^{r+1}n-k}^{r+1}(2^{r+1}n-2x) = \phi_k^{r+1}(2x), \quad x \in [0,2^r n], \quad k=0,1,\ldots,2^{r+1}n, \tag{3.25}$$

which, if true, completes the proof of the theorem. Definition (3.12c) implies that (3.25) is true for $k \in \{0,1,\ldots,2N+1\} \cup \{2^r n-2N-1,\ldots,2^r n\}$, whereas if $k \in \{2N+2,\ldots,2^r n-2N-2\}$, we find that (3.12b) and (2.14) yield, for $x \in [0,2^r n]$,

$$\phi_{2^{r+1}n-k}^{r+1}(2^{r+1}n-2x) = \phi(k-2x) = \phi(2x-k) = \phi_k^{r+1}(2x). \qquad \square$$

3.3. Convergence of the modified subdivision scheme. We now prove the analogue of Theorem 2.2 for *finite* subdivision sequences.

THEOREM 3.3. *For each initial sequence $c = \{c_j\} \in M_0$, the subdivision scheme* (3.6)–(3.7) *converges to the function*

$$g(x) = \sum_{j=0}^{n} c_j \phi_j(x), \quad x \in [0, n], \tag{3.26}$$

with $\phi_j = \phi_j^0$ defined as in (3.12)*, in the sense that*

$$c_k^{(r)} = g\left(\frac{k}{2^r}\right), \quad k = 0, 1, \ldots, 2^r n, \quad r = 0, 1, \ldots. \tag{3.27}$$

Proof. Repeatedly using (3.5), (3.6), (3.7), and eventually (3.3), for $k \in \{0, 1, \ldots, 2^r n\}$ and $r \in \{0, 1, \ldots\}$, we obtain

$$\begin{aligned} g\left(\frac{k}{2^r}\right) &= \sum_{j=0}^{n} c_j \phi_j^0\left(\frac{k}{2^r}\right) \\ &= \sum_{j=0}^{n} c_j \sum_{\ell=0}^{2n} \phi_j^0\left(\frac{\ell}{2}\right) \phi_\ell^1\left(\frac{k}{2^{r-1}}\right) \\ &= \sum_{\ell=0}^{2n} c_\ell^{(1)} \phi_\ell^1\left(\frac{k}{2^{r-1}}\right) \\ &\;\;\vdots \\ &= \sum_{\ell=0}^{2^r n} c_\ell^{(r)} \phi_\ell^r(k) = c_k^{(r)}. \end{aligned}$$

□

3.4. An explicit formulation. We derive an explicit formulation for the subdivision scheme (3.6)–(3.7).

Let $c \in M_r$. Then, from the subdivision operator definition (3.6), we obtain

$$(S_r c)_{2j+1} = \sum_{k=0}^{2^r n} \phi_k^r\left(j + \tfrac{1}{2}\right) c_k, \quad j = 0, 1, \ldots, 2^r n - 1, \tag{3.28}$$

and thus, using also (3.12b), (3.14), and the top of (3.17), we get

$$(S_r c)_{2j+1} = \sum_{k=0}^{2N+1} L_{k-N}\left(j + \tfrac{1}{2} - N\right) c_k, \quad j = 0, 1, \ldots, N. \tag{3.29}$$

Next, we claim that

$$(S_r c)_{2j+1} = \sum_{k=-N+j}^{N+1+j} L_{k-j}\left(\tfrac{1}{2}\right) c_k, \quad j = N+1, \ldots, 2^r n - 2N - 2. \tag{3.30}$$

Indeed, if $j \in \{N+1, \ldots, 2N-1\}$, then (3.28), (3.2), (3.17), (3.12b), (2.15), and (2.6c) yield

$$(S_r c)_{2j+1} = \sum_{k=0}^{4N} \phi\left(j + \tfrac{1}{2} - k\right) c_k = \sum_{k=-N+j}^{N+1+j} a_{2j+1-2k} c_k,$$

and (3.30) then follows from (2.6b) for that range of j. Similarly, if $j \in \{2N, \ldots, 2^r n - 2N - 2\}$, we additionally use (3.15) to get

$$(S_r c)_{2j+1} = \sum_{k=0}^{2^r n} \phi\left(j + \tfrac{1}{2} - k\right) c_k = \sum_{k=-N+j}^{N+1+j} a_{2j+1-2k} c_k,$$

which, together with (2.6b), then proves (3.30).

For $j \in \{2^r n - 2N - 1, \ldots, 2^r n - 1\}$, we first note the symmetry

$$\phi_k^r(x) = \phi_{2^r n-k}^r(2^r n - x), \quad x \in [0, 2^r n], \quad k = 0, 1, \ldots, 2^r n, \tag{3.31}$$

which follows from (3.12c), and the fact that, for $k \in \{2N + 2, \ldots, 2^r n - 2N - 2\}$, (3.12b) and (2.14) give

$$\phi_{2^r n-k}^r(2^r n - x) = \phi(k - x) = \phi(x - k) = \phi_k^r(x).$$

Thus, from (3.31),

$$\begin{aligned} \phi_k^r\left(j + \tfrac{1}{2}\right) = \phi_{2^r n-k}^r\left((2^r n - 1 - j) + \tfrac{1}{2}\right), \qquad & k = 0, 1, \ldots, 2^r n, \\ & j = 2^r n - 2N - 1, \ldots, 2^r n - 1. \end{aligned} \tag{3.32}$$

Combining (3.29), (3.30), (3.32), and using (3.9), we find that the subdivision scheme (3.6)–(3.7) has, for a given initial sequence $c^{(0)} = c \in M_0$, the explicit formulation

$$\left.\begin{aligned} c_{2j}^{(r+1)} &= c_j^{(r)}, && j = 0, 1, \ldots, 2^r n, \\ c_{2j+1}^{(r+1)} &= \sum_{k=0}^{2^r n} a_{j,k}^{(r)} c_k^{(r)}, && j = 0, 1, \ldots 2^r n - 1, \end{aligned}\right\} \quad r = 0, 1, \ldots, \tag{3.33}$$

where

$$a_{j,k}^{(r)} = \begin{cases} L_{k-N}\left(j + \frac{1}{2} - N\right), & k = 0, 1, \ldots, 2N + 1, \\ 0, & k = 2N + 2, \ldots, 2^r n, \end{cases} \quad \begin{array}{l} j = 0, 1, \ldots, N - 1, \\ \text{(if } N \geq 1\text{)}, \end{array} \tag{3.34}$$

$$a_{j,k}^{(r)} = \begin{cases} L_{k-j}(\frac{1}{2}), & k = -N + j, \ldots, N + 1 + j, \\ 0, & k \in [0, -N - 1 + j] \cup [N + 2 + j, 2^r n], \end{cases} \quad j = N, \ldots, 2^r n - N - 1, \tag{3.35}$$

$$a_{j,k}^{(r)} = a_{2^r n-1-j,2^r n-k}^{(r)}, \quad k = 0, 1, \ldots, 2^r n, j = 2^r n - N, \ldots, 2^r n - 1. \tag{3.36}$$

Explicit formulations of (3.34) and (3.35) are now obtained by using a calculation similar to the one which yielded the top of (2.8). We find that, for $k = 0, 1, \ldots, 2N+1$ and $j = 0, 1, \ldots, N - 1$,

$$L_{k-N}(j + \tfrac{1}{2} - N) = \frac{(-1)^{j+k}}{2^{4N+1}} \frac{1}{2j + 1 - 2k} \frac{(2j + 1)!(4N + 1 - 2j)!}{(2N - j)!(2N + 1 - k)!j!k!}, \tag{3.37}$$

whereas for $k = -N + j, \dots, N + 1 + j$ and $j = N, \dots, 2^r n - N - 1$,

$$L_{k-j}(\tfrac{1}{2}) = \frac{N+1}{2^{4N+1}} \binom{2N+1}{N} \frac{(-1)^{j+k}}{2j+1-2k} \binom{2N+1}{N+1+j-k}. \tag{3.38}$$

For example, if $N = 1$, (3.34) and (3.38) give

$$a_{0,k}^{(r)} = \begin{cases} \frac{5}{16}, & k = 0, \\ \frac{15}{16}, & k = 1, \\ -\frac{5}{16}, & k = 2, \\ \frac{1}{16}, & k = 3, \\ 0, & k = 4, \dots, 2^r n. \end{cases}$$

4. Interpolation wavelets on an interval. We show how the refinable sequence $\{\phi_j^r : j = 0, 1, \dots, 2^r n,\ r = 0, 1, \dots\}$ of section 3 can be used to explicitly construct interpolation wavelets on an interval. Our definition in (4.10) below coincides, in an inner region bounded away from the endpoints, with the definition of interpolation wavelets on $\mathbb{R}$ as given in, e.g., [7, equation (1.11)], [22, p. 300], and [27, p. 193].

4.1. Decomposition based on interpolation. Let the integers N and $n \geq 4N + 2$ be as in section 3, and let R be a given positive integer. We define the linear space sequence $\{V_r : r = 0, 1, \dots, R\}$ by

$$V_r = \operatorname{span}\{\phi_j^r(2^{r-R}\,\bullet) : j = 0, 1, \dots, 2^r n\}, \quad r = 0, 1, \dots, R, \tag{4.1}$$

and the linear operator sequence $\{P_r : r = 0, 1, \dots, R\}$ for $P_r : C[0, 2^R n] \to V_r$ by

$$(P_r f)(x) = \sum_{j=0}^{2^r n} f(2^{R-r} j)\phi_j^r(2^{r-R} x), \quad x \in [0, 2^R n], \quad r = 0, 1, \dots, R. \tag{4.2}$$

It follows from (3.3) that P_r is an interpolation operator, which means that, for each $f \in C[0, 2^R n]$,

$$(P_r f)(2^{R-r} j) = f(2^{R-r} j), \quad j = 0, \dots, 2^r n, \quad r = 0, 1, \dots, R. \tag{4.3}$$

Also, a proof based on (3.3) shows that P_r is a projection on V_r. Thus,

$$P_r f = f, \quad f \in V_r. \tag{4.4}$$

Furthermore, (3.4) gives

$$\sum_{j=0}^{2^r n} p(2^{R-r} j)\phi_j^r(2^{r-R} x) = p(x), \quad x \in [0, 2^R n], \quad p \in \pi_{2N+1}, \quad r = 0, 1, \dots, R, \tag{4.5}$$

by virtue of which

$$\pi_{2N+1} \subset V_r, \quad r = 0, 1, \dots, R, \tag{4.6}$$

where here π_{2N+1} is restricted to the interval $[0, 2^R n]$.

Since (3.5) yields the refinement equation

$$\phi_j^r(2^{r-R}x) = \sum_{k=0}^{2^{r+1}n} \phi_j^r\left(\frac{k}{2}\right)\phi_k^{r+1}(2^{r+1-R}x), \quad x \in [0, 2^R n], \quad j = 0, 1, \ldots, 2^r n, \quad r = 0, \ldots, R-1, \tag{4.7}$$

we have the nesting property

$$V_r \subset V_{r+1}, \quad r = 0, 1, \ldots, R-1. \tag{4.8}$$

Based on (3.3) one can prove that for each fixed r the set $\{\phi_j^r(2^{r-R}\bullet) : j = 0, 1, \ldots, 2^r n\}$ is linearly independent on $[0, 2^R n]$, and thus, from (4.1), we have

$$\dim V_r = 2^r n + 1, \quad r = 0, 1, \ldots, R. \tag{4.9}$$

Now define the sequence $\{\psi_j^r\} = \{\psi_j^r : j = 0, 1, \ldots, 2^r n - 1,\ r = 0, 1, \ldots, R-1\}$ by

$$\psi_j^r(x) = \phi_{2j+1}^{r+1}(x), \quad x \in [0, 2^R n], \quad j = 0, 1, \ldots, 2^r n - 1, \quad r = 0, 1, \ldots, R-1, \tag{4.10}$$

with corresponding linear spaces

$$W_r = \text{ span}\{\psi_j^r(2^{r+1-R}\bullet) : j = 0, 1, \ldots, 2^r n - 1\}, \quad r = 0, 1, \ldots, R-1. \tag{4.11}$$

From (4.10) and (4.1) it follows that

$$W_r \subset V_{r+1}, \quad r = 0, 1, \ldots, R-1. \tag{4.12}$$

If U, V, and W are linear spaces, we use the direct sum notation $U = V \oplus W$ to denote the fact that, for each $f \in V$, there exist $g \in V$ and $h \in W$ such that $f = g + h$, and with g and h uniquely determined by f.

The following direct sum decomposition result holds.

THEOREM 4.1.

$$V_{r+1} = V_r \oplus W_r, \quad r = 0, 1, \ldots, R-1, \tag{4.13}$$

with V_r and W_r defined by (4.1) *and* (4.11).

To prove Theorem 4.1, we first introduce the linear spaces U_r and X_r, where

$$U_r = \{f - P_r f : f \in V_{r+1}\}, \quad r = 0, 1, \ldots, R-1, \tag{4.14}$$

with P_r defined by (4.2), and

$$X_r = \{f \in V_{r+1} : f(2^{R-r}j) = 0, \quad j = 0, 1, \ldots, 2^r n\}, \quad r = 0, 1, \ldots, R-1, \tag{4.15}$$

in terms of which the following preliminary result holds.

PROPOSITION 4.2. *The linear spaces V_r, W_r, U_r, and X_r, as defined by* (4.1), (4.11), (4.14), *and* (4.15), *satisfy*

$$W_r = U_r = X_r, \quad r = 0, 1, \ldots, R-1; \tag{4.16}$$

$$V_r \cap W_r = \{0\}, \quad r = 0, 1, \ldots, R-1. \tag{4.17}$$

Proof. First, we show that $U_r = X_r$. If $g \in U_r$, then (4.2) and (4.3) yield $g(2^{R-r}j) = 0,\ j = 0, \ldots, 2^r n$, i.e., $g \in X_r$, so that $U_r \subset X_r$. If $g \in X_r$, then (4.2) and (4.15) imply $P_r g = 0$; hence, $g = f - P_r f$ with $f = g$, and thus, since also $f \in X_r \subset V_{r+1}$ from (4.15), we have from (4.14) that $g \in U_r$. Consequently, $X_r \subset U_r$.

Next, we prove that $W_r \subset X_r$. If $g \in W_r$, there exists a coefficient sequence $\{c_0, \ldots, c_{2^r n-1}\}$ such that $g(x) = \sum_{\ell=0}^{2^r n-1} c_\ell \psi_\ell^r(2^{r+1-R}x),\ x \in [0, 2^R n]$. But then, from (4.10), we have

$$g(2^{R-r}j) = \sum_{\ell=0}^{2^r n-1} c_\ell \phi_{2\ell+1}^{r+1}(2j) = 0, \quad j \in \{0, 1, \ldots, 2^r n\},$$

by virtue of (3.3). Definition (4.15) then implies that $g \in X_r$. So, $W_r \subset X_r$.

We now show that $U_r \subset W_r$, thereby completing the proof of (4.16). Suppose therefore that $g \in U_r$, so that, from (4.14) and (4.1), there exists a coefficient sequence $\{c_j : j = 0, 1, \ldots, 2^{r+1}n\}$ such that the function $f \in V_{r+1}$ given by

$$f(x) = \sum_{j=0}^{2^{r+1}n} c_j \phi_j^{r+1}(2^{r+1-R}x), \quad x \in [0, 2^R n], \tag{4.18}$$

satisfies the equation

$$g(x) = f(x) - (P_r f)(x), \quad x \in [0, 2^R n]. \tag{4.19}$$

However, from (4.2), (4.18), (3.3), and (3.5), we obtain

$$\begin{aligned}
(P_r f)(x) &= \sum_{\ell=0}^{2^{r+1}n} \left[\sum_{k=0}^{2^r n} \phi_\ell^{r+1}(2k) \phi_k^r(2^{r-R}x) \right] c_\ell \\
&= \sum_{\ell=0}^{2^r n} c_{2\ell} \phi_\ell^r(2^{r-R}x) \\
&= \sum_{\ell=0}^{2^r n} c_{2\ell} \sum_{k=0}^{2^{r+1}n} \phi_\ell^r\left(\frac{k}{2}\right) \phi_k^{r+1}(2^{r+1-R}x) \\
&= \sum_{\ell=0}^{2^r n} c_{2\ell} \phi_{2\ell}^{r+1}(2^{r+1-R}x) \\
&\quad + \sum_{\ell=0}^{2^r n} c_{2\ell} \sum_{k=0}^{2^r n-1} \phi_\ell^r\left(k + \tfrac{1}{2}\right) \phi_{2k+1}^{r+1}(2^{r+1-R}x).
\end{aligned} \tag{4.20}$$

Also, from (4.18),

$$f(x) = \sum_{j=0}^{2^r n} c_{2j} \phi_{2j}^{r+1}(2^{r+1-R}x) + \sum_{j=0}^{2^r n-1} c_{2j+1} \phi_{2j+1}^{r+1}(2^{r+1-R}x). \tag{4.21}$$

Combining (4.19), (4.20), and (4.21) yields

$$g(x) = \sum_{j=0}^{2^r n-1} \left[c_{2j+1} - \sum_{k=0}^{2^r n} c_{2k} \phi_k^r\left(j + \tfrac{1}{2}\right) \right] \phi_{2j+1}^{r+1}(2^{r+1-R}x), \tag{4.22}$$

which, together with (4.11), (4.10), implies that $g \in W_r$. Thus $U_r \subset W_r$, thereby completing our proof of (4.16).

Finally, to prove (4.17), suppose $f \in V_r \cap W_r$. Then (4.4) and (4.2) imply that, for $x \in [0, 2^R n]$,

$$f(x) = (P_r f)(x) = \sum_{j=0}^{2^r n} f(2^{R-r} j)\phi_j^r(2^{r-R}x) = 0,$$

after noting from (4.16) that $f \in X_r$ and then using the definition (4.15). □

We can now prove Theorem 4.1.

Proof. Let $r \in \{0, 1, \ldots, R-1\}$ be fixed. Take any $f \in V_{r+1}$ and define $g = P_r f$ and $h = f - P_r f$. Then (4.2) implies that $g \in V_r$, whereas (4.14) and (4.16) imply that $h \in W_r$. We have therefore shown that there exist functions $g \in V_r$ and $h \in W_r$ such that $f = g + h$. It remains to be proven that g and h are uniquely determined by f. But, if $g_0 \in V_r$ and $h_0 \in W_r$ were such that $f = g_0 + h_0$, then $u = g_0 - g \in V_r$ and $v = h - h_0 \in W_r$, with $u = v$. Thus $u \in V_r \cap W_r$ and $v \in V_r \cap W_r$, and the desired uniqueness result follows from (4.17). □

From (4.13) one concludes that $\dim W_r = \dim V_{r+1} - \dim V_r$, so that (4.9) leads to

$$\dim W_r = 2^r n, \quad r = 0, 1, \ldots, R-1. \tag{4.23}$$

Hence, based on the definition (4.11), we conclude that the set $\{\psi_j^r(2^{r+1-R}\,\bullet) : j = 0, 1, \ldots, 2^r n - 1\}$ is linearly independent on $[0, 2^R n]$ and therefore is a basis for W_r.

We have therefore established, for each fixed $r \in \{0, 1, \ldots, R-1\}$, an *interpolation wavelet basis* $\{\psi_j^r(2^{r+1-R}\,\bullet) : j = 0, 1, \ldots, 2^r n - 1\}$ for the *interpolation wavelet space* W_r. The elements of the sequence $\{\psi_j^r\}$ are called *interpolation wavelets*. Examples for $N = 1$ are plotted in Figure 5.

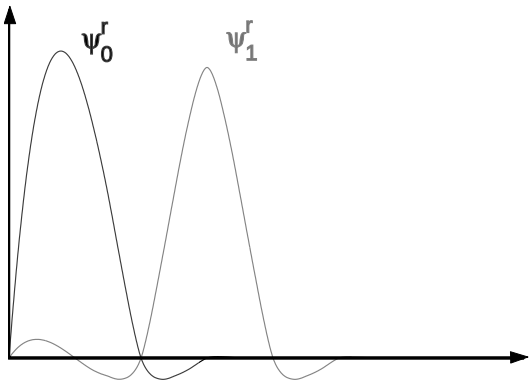


FIG. 5. *Boundary interpolation wavelets with* $N = 1$.

In the next section, we derive the corresponding decomposition and reconstruction algorithms.

4.2. Decomposition and reconstruction algorithms. To obtain the decomposition algorithm, let $r \in \{0, 1, \ldots, R-1\}$ be fixed, and suppose $f_{r+1} \in V_{r+1}$ is given by

$$f_{r+1}(x) = \sum_{j=0}^{2^{r+1}n} c_j^{(r+1)} \phi_j^{r+1}(2^{r+1-R}x), \quad x \in [0, 2^R n]. \tag{4.24}$$

According to Theorem 4.1, and for the bases $\{\phi_j^r\}$ and $\{\psi_j^r\}$ of V_r and W_r, we know that there exist unique coefficient sequences $\{c_j^{(r)} : j = 0, 1, \ldots, 2^r n\}$ and $\{d_j^{(r)} : j = 0, 1, \ldots, 2^r n - 1\}$ such that the functions $f_r \in V_r$ and $g_r \in W_r$ defined by

$$f_r(x) = \sum_{j=0}^{2^r n} c_j^{(r)} \phi_j^r(2^{r-R}x), \quad x \in [0, 2^R n], \tag{4.25}$$

and

$$g_r(x) = \sum_{j=0}^{2^r n-1} d_j^{(r)} \psi_j^r(2^{r+1-R}x), \quad x \in [0, 2^R n], \tag{4.26}$$

satisfy

$$f_{r+1} = f_r + g_r. \tag{4.27}$$

In particular, observe that we then have the *interpolation wavelet decomposition*

$$f_R = f_0 + \sum_{j=0}^{R-1} g_j.$$

Moreover,

$$f_r = P_r f_{r+1} \tag{4.28}$$

and

$$g_r = f_{r+1} - P_r f_{r+1}. \tag{4.29}$$

The coefficients $\{d_j^{(r)}\}$ in (4.26) are called the *interpolation wavelet coefficients*.

Using (4.24), (4.10), and (4.29), and the argument which led to (4.22), we obtain, for all $x \in [0, 2^R n]$,

$$\sum_{j=0}^{2^r n-1} d_j^{(r)} \psi_j^r(2^{r+1-R}x) = \sum_{j=0}^{2^r n-1} \left[c_{2j+1}^{(r+1)} - \sum_{k=0}^{2^r n} c_{2k}^{(r+1)} \phi_k^r\left(j + \tfrac{1}{2}\right) \right] \psi_j^r(2^{r+1-R}x).$$

Since $\{\psi_j^r(2^{r+1-R}\,\bullet) : j = 0, 1, \ldots, 2^r n - 1\}$ is a linearly independent set on $[0, 2^R n]$, we have

$$d_j^{(r)} = c_{2j+1}^{(r+1)} - \sum_{k=0}^{2^r n} c_{2k}^{(r+1)} \phi_k^r\left(j + \tfrac{1}{2}\right), \quad j = 0, 1, \ldots, 2^r n - 1. \tag{4.30}$$

Next, using (4.25), (4.28), (4.2), and (3.3) we get, for all $x \in [0, 2^R n]$,

$$\begin{aligned} \sum_{j=0}^{2^r n} c_j^{(r)} \phi_j^r(2^{r-R}x) = f_r(x) &= (P_r f_{r+1})(x) \\ &= \sum_{j=0}^{2^r n} \left[\sum_{\ell=0}^{2^{r+1} n} c_\ell^{(r+1)} \phi_\ell^{r+1}(2j) \right] \phi_j^r(2^{r-R}x) \\ &= \sum_{j=0}^{2^r n} c_{2j}^{(r+1)} \phi_j^r(2^{r-R}x). \end{aligned}$$

Since the set $\{\phi_j^r(2^{r-R}\bullet) : j = 0, 1, \ldots, 2^r n\}$ is linearly independent on $[0, 2^R n]$, we deduce that

$$c_j^{(r)} = c_{2j}^{(r+1)}, \quad j = 0, 1, \ldots, 2^r n. \tag{4.31}$$

Now observe from (3.28), (3.7), and (3.33) that

$$\phi_k^r(j + \tfrac{1}{2}) = a_{j,k}^{(r)}, \quad k = 0, 1, \ldots, 2^r n, \quad j = 0, 1, \ldots, 2^r n - 1,$$

which, together with (4.30) and (4.31), then yield, for a given data sequence $\{c_j^{(R)} : j = 0, 1, \ldots, 2^R n\}$, the following interpolation wavelet algorithms.

Decomposition algorithm:

$$\left.\begin{aligned} c_j^{(r)} &= c_{2j}^{(r+1)}, && j = 0, 1, \ldots, 2^r n, \\ d_j^{(r)} &= c_{2j+1}^{(r+1)} - \sum_{k=0}^{2^r n} a_{j,k}^{(r)} c_{2k}^{(r+1)}, && j = 0, 1, \ldots, 2^r n - 1, \end{aligned}\right\} r = R-1, R-2, \ldots, 0. \tag{4.32}$$

Reconstruction algorithm:

$$\left.\begin{aligned} c_{2j}^{(r+1)} &= c_j^{(r)}, && j = 0, 1, \ldots, 2^r n, \\ c_{2j+1}^{(r+1)} &= d_j^{(r)} + \sum_{k=0}^{2^r n} a_{j,k}^{(r)} c_k^{(r)}, && j = 0, 1, \ldots, 2^r n - 1, \end{aligned}\right\} r = 0, 1, \ldots, R-1. \tag{4.33}$$

Here the coefficient sequence $\{a_{j,k}^{(r)} : k = 0, 1, \ldots, 2^r n,\ j = 0, 1, \ldots, 2^r n - 1,\ r = 0, 1, \ldots, R-1\}$ is defined by (3.34), (3.35), (3.36), with explicit formulations in (3.37) and (3.38).

Suppose $f \in C[0, 2^R n]$, with the integers R and n suitably chosen. The sequence $\{f_r : r = R-1, R-2, \ldots, 0\}$ is then defined by (4.28), with $f_R = f$, whereas the sequence $\{g_r : r = R-1, R-2, \ldots, 0\}$ is defined by (4.29). The coefficient sequences $\{c_j^{(r)} : j = 0, 1, \ldots, 2^r n,\ r = R, R-1, \ldots, 0\}$ and $\{d_j^{(r)} : j = 0, 1, \ldots, 2^r n - 1,\ r = R-1, R-2, \ldots, 0\}$ are computed recursively by means of (4.32). In particular, observe that since (4.2) and (4.25) yield

$$\sum_{j=0}^{2^R n} c_j^{(R)} \phi_j^R(x) = f_R(x) = (P_R f)(x) = \sum_{j=0}^{2^R n} f(j) \phi_j^R(x),$$

we have $c_j^{(R)} = f(j),\ j = 0, 1, \ldots, 2^R n$, and thus the interpolation wavelet decomposition algorithm (4.32) can, in this context, be rewritten as

$$\left.\begin{aligned} c_j^{(r)} &= f(2^{R-r} j), && j = 0, 1, \ldots, 2^r n, \\ d_j^{(r)} &= f(2^{R-r-1}(2j+1)) - \sum_{k=0}^{2^r n} a_{j,k}^{(r)} f(2^{R-r} k), && j = 0, 1, \ldots, 2^r n - 1, \end{aligned}\right\} \tag{4.34}$$

$$r = R-1, R-2, \ldots, 0.$$

The reconstruction is then performed by means of (4.33).

At this stage, it is of interest to point out the following relationship between the interpolation wavelet procedure (4.34), (4.33) and the interpolatory subdivision scheme (3.33).

After the decomposition with (4.34) has been performed, suppose that we set the interpolation wavelet coefficients

$$d_j^{(r)} = 0, \quad j = 0, 1, \ldots, 2^r n - 1, \quad r = 0, 1, \ldots, R - 1, \tag{4.35}$$

at each successive step of the reconstruction phase (4.33). The final reconstructed (and smoothed) function is then

$$f_R(x) = \sum_{j=0}^{2^R n} c_j^{(R)} \phi_j^R(x), \quad x \in [0, 2^R n], \tag{4.36}$$

with, as is clear from the top parts of (4.33) and (4.34), and (3.3),

$$f_R(2^R k) = c_{2^R k}^{(R)} = f(2^R k), \quad k = 0, 1, \ldots, n.$$

Now observe that the zero values (4.35) substituted into the bottom part of (4.33) yield precisely the bottom part of the subdivision formula (3.33).

Hence the interpolation wavelet decomposition and reconstruction scheme described above is equivalent to the following interpolatory subdivision scheme:

Choose $c = c^{(0)} = f(2^R k), \ \ k = 0, 1, \ldots,$ and apply the interpolatory subdivision scheme (3.33). According to Theorem 3.3, this scheme converges to the limit curve

$$g(x) = \sum_{j=0}^{n} f(2^R j) \phi_j^0(x), \quad x \in [0, n]. \tag{4.37}$$

Then, using the refinement equation (3.5), as well as (3.6)–(3.7), we get

$$\begin{aligned} g(x) &= \sum_{j=0}^{n} c_j^{(0)} \phi_j^0(x) \\ &= \sum_{j=0}^{n} c_j^{(0)} \sum_{\ell=0}^{2n} \phi_j^0 \left(\frac{\ell}{2} \right) \phi_j^1(2x) \\ &= \sum_{\ell=0}^{2n} c_\ell^{(1)} \phi_j^1(2x) \\ &\ \ \vdots \\ &= \sum_{\ell=0}^{2^R n} c_\ell^{(R)} \phi_j^R(2^R x) = f_R(2^R x) \end{aligned}$$

from (4.36). Hence,

$$g(x) = f_R(2^R x), \quad x \in [0, n]. \tag{4.38}$$

The subdivision approach therefore has the significant advantage of providing an efficient iterative procedure for the construction of the function f_R in (4.36).

Finally, suppose in the decomposition procedure defined by the algorithm (4.34) we have, for some given polynomial $p \in \pi_{2N+1}$, and for a fixed $r \in \{R-1, R-2, \ldots, 0\}$, that there exists an integer $\mu \in [0, 2^r n]$ such that

$$f(x) = p(x), \quad x \in [\alpha, \beta],$$

where

$$\alpha = \max\{0, 2^{R-r}(\mu - 2N)\}, \qquad \beta = \min\{2^R n, 2^{R-r}(\mu + 2N)\}.$$

Then, using also (4.34), (4.30), (3.2), and (3.4), we find

$$\begin{aligned} d_\mu^{(r)} &= p(2^{R-r-1}(2\mu+1)) - \sum_{k=\max\{0,\mu-2N\}}^{k=\min\{2^r n,\mu+2N\}} p(2^{R-r}k)\phi_k^r\left(\mu + \tfrac{1}{2}\right) \\ &= p(2^{R-r-1}(2\mu+1)) - \sum_{k=0}^{2^r n} p(2^{R-r}k)\phi_k^r\left(\mu + \tfrac{1}{2}\right) = 0. \end{aligned}$$

Hence the interpolation wavelet coefficients $d_j^{(r)}$ corresponding, according to (3.2), to those regions in $[0, 2^R n]$ where the function f has polynomial-like behavior can be expected to be small relative to the interpolation wavelet coefficients $d_j^{(r)}$ corresponding to those regions in $[0, 2^R n]$ where f exhibit nonpolynomial-like behavior.

For example, choose the function f in (4.34) as the cubic cardinal B-spline, with an arbitrary choice of the integers R and n. The B-spline consists of four cubic polynomial pieces joined together in such a way that the second derivative of the B-spline is continuous, while the third derivative has jump discontinuities at the nodes. Now choose $N = 1$ to ensure, by the above argument, that the wavelet coefficients $d_j^{(r)}$, $r = R-1, R-2, \ldots, 0$, are zero in the regions where f is identical to a polynomial and nonzero in the regions where the third derivative has a jump discontinuity.

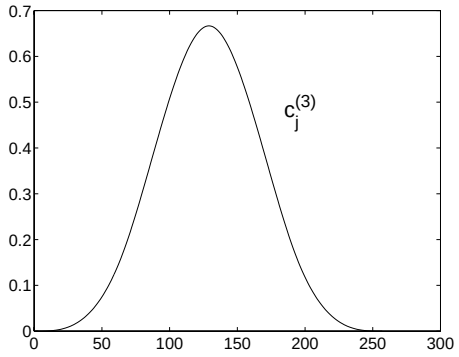
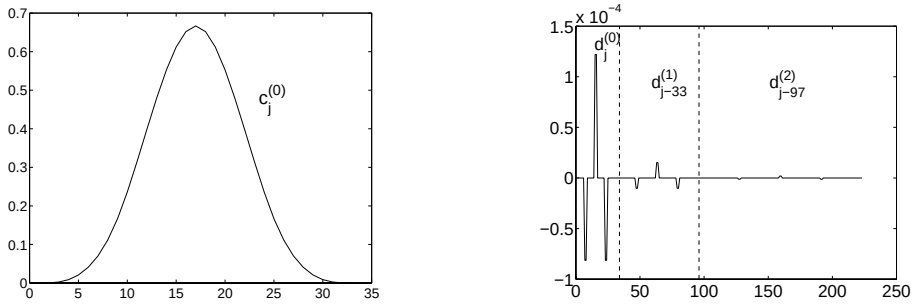


FIG. 6. *Cubic B-spline, sampled at* 257 *equally spaced points.*

In Figure 6, we have chosen $R = 3$ and $n = 32$ and plotted the sequence of sampled values $c_j^{(3)} = f(j)$ versus their indices. Figure 7 shows the result of the interpolation wavelet decomposition algorithm (4.34), where the resulting coefficients are also plotted versus their indices.

We observe that the average values $c_j^{(0)}$, as shown in Figure 7(a), consist of precisely every eighth value of the original B-spline. What is of more interest is the set



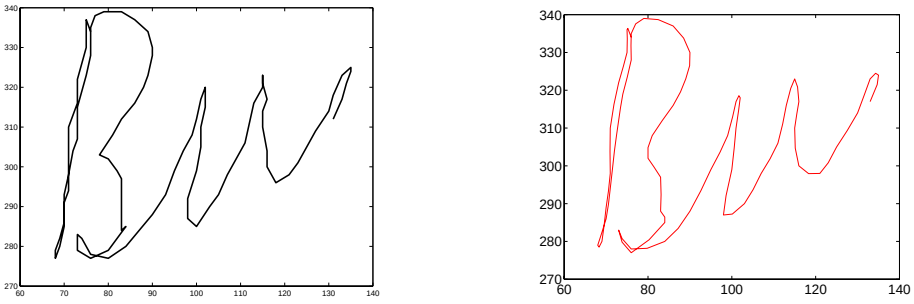
(a) *Average component.* (b) *Detail components.*

FIG. 7. *Third level decomposition of the B-spline of Figure* 6.

of detail components shown in Figure 7(b). Note that, as expected, the interpolation wavelet coefficients $d_j^{(r)}$, $r = 2, 1, 0$, are nonzero only where the B-spline has jump discontinuities in its third derivative.

5. Examples. In this section we illustrate the interpolation wavelets by applying them to two practical problems. Of course we do not claim that the schemes described above are more efficient in practice than any of the alternatives. This would require a detailed study, which we leave for future consideration. These examples merely serve to illustrate the theory developed above. Note in particular how the interpolation wavelet decomposition on an interval avoids any edge artifacts.

5.1. Signature smoothing. Figure 8(a) shows part of a signature that was captured by a digitized tablet. One clearly sees the quantization effect of the underlying grid of this particular tablet. Almost all applications require that the signature should be smoothed. We therefore applied a single level of the interpolation wavelet decomposition (4.34) with $N = 1$. All the detail coefficients were set equal to zero, and a single reconstruction step was performed using the zero detail coefficients. The result is the smoothed signature in Figure 8(b). Note that we follow standard practice by displaying only the discrete coefficients $c_j^{(R)}$, connected by straight line segments. This is different from the smooth curve given by (4.36).

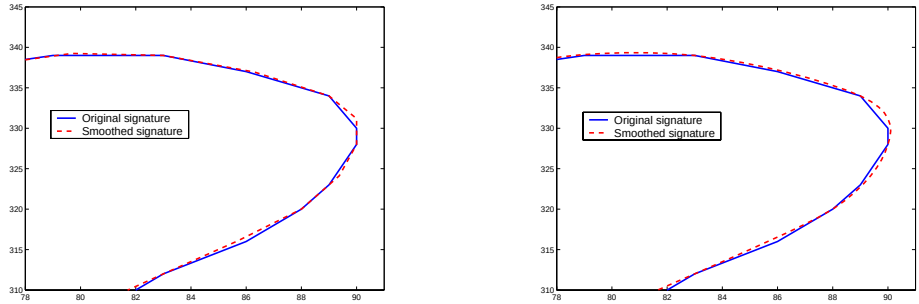


(a) *Original signature sample.* (b) *Smoothed signature.*

FIG. 8. *Illustration of smoothing with interpolation wavelet decomposition.*

A closer look at the original and reconstructed signature is given in Figure 9. Note from Figure 9(a) how every other data point stays the same (due to the interpolation property). The remaining points, however, are calculated in such a way that the

result is a smoother signature. In fact, this procedure is exactly the same as if we discarded every other data point of the original signature and then applied one step of the subdivision scheme with $N = 1$ to the result, as described in section 4. As mentioned above we plot only the data points connected by straight lines, and not the smooth curve f_R described by (4.36). An efficient way of calculating f_R is by subdivision: according to the argument leading from (4.36) to (4.38), each step of the subdivision scheme doubles the number of points on the curve f_R. The result of two more subdivision steps is shown in Figure 9(b). Note how the reconstructed curve has become noticeably smoother.



(a) *A single reconstruction step.* (b) *After two more subdivision steps.*

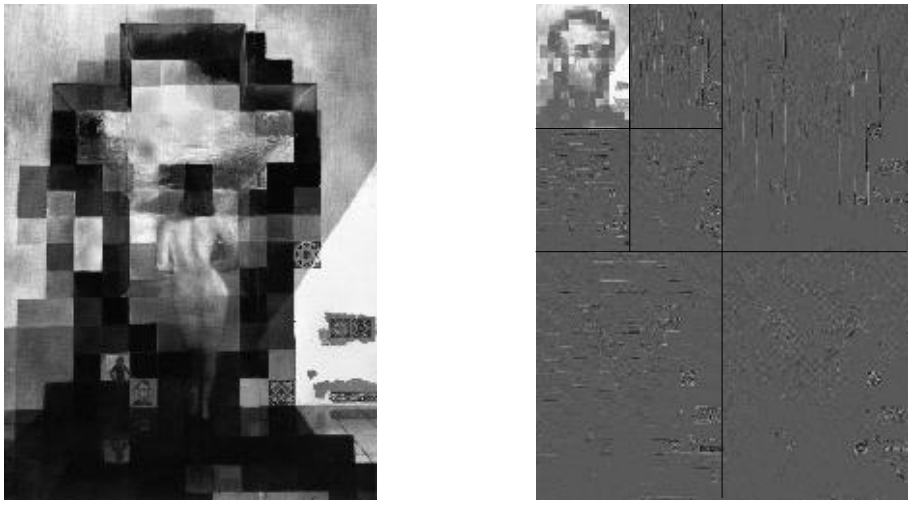
FIG. 9. *Magnification of original and smoothed signature.*

Note that the main problem in this case is that the data points themselves are corrupted by noise. Insisting that the data points are interpolated is therefore not the best way to proceed. In this case noninterpolatory schemes such as Lane–Riesenfeld [20] (see also [23, Chapter 2]) might prove beneficial. Even in this less than ideal situation, the interpolation wavelets provide a surprisingly good smoothing, with no artificial edge effects.

5.2. Two-dimensional interpolation wavelet decomposition. For our second example we use the well-known painting by Salvador Dali, *Gala Contemplating the Mediterranean*, Figure 10(a), the original of which can be seen in the Dali museum in Figueres, Spain. Although painted in 1976, it is a beautiful illustration of Dali's awareness of images on different scales, in this case, a portrait of Abraham Lincoln, the 16th president of the United States of America, and Gala, Dali's wife, looking out to sea.

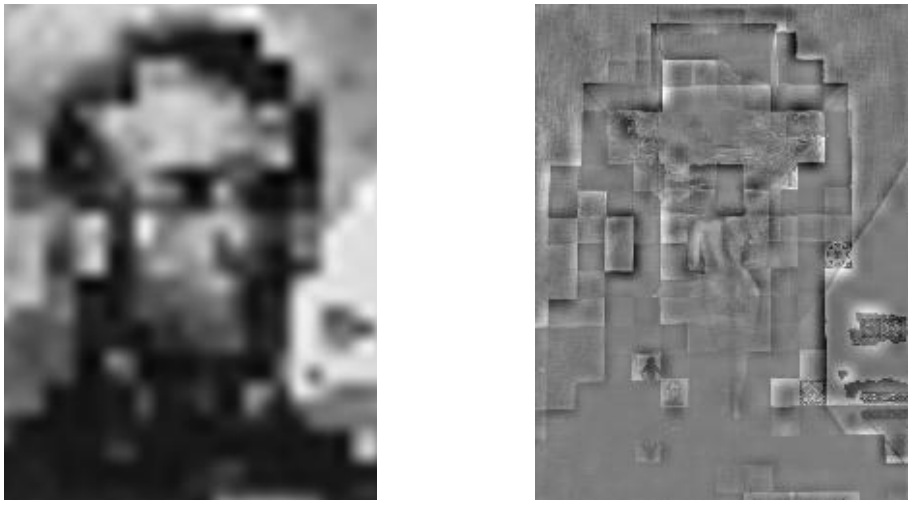
Constructing a two-dimensional tensor product from the interpolation wavelet (4.10) (see, e.g., [5, section 6.4]) allows us to perform a two-dimensional decomposition of Dali's painting, as illustrated in Figure 10(b). Note how clearly the image of Abraham Lincoln is captured by the average component, displayed in the top left-hand corner of Figure 10(b). The remaining part of the figure consists of the various detail components.

Now we set the detail components equal to zero, and then reconstruct according to a two-dimensional reconstruction algorithm also based on the tensor product interpolation wavelet. The result, as shown in Figure 11(a), is the image of Abraham Lincoln interpolated back onto the original grid of Figure 10(a). If we set the average component equal to zero, and then reconstruct, we obtain the image in Figure 11(b), which contains all the areas of sharp transition which are absent from Figure 11(a).



(a) *Original image.* (b) *Decomposition of image.*

FIG. 10. *Illustration of two-dimensional interpolation wavelet decomposition.*



(a) *Detail removed.* (b) *Average removed.*

FIG. 11. *Reconstructed images.*

Acknowledgment. We are grateful to Willy Hereman for numerous suggestions and comments during the preparation of this manuscript.

REFERENCES

[1] F. ARÀNDIGA, R. DONAT, AND A. HARTEN, *Multiresolution based on weighted averages of the hat function* I: *Linear reconstruction techniques*, SIAM J. Numer. Anal., 36 (1998), pp. 160–203.

[2] S. BERTOLUZZA AND G. NALDI, *Some remarks on wavelet interpolation*, Comput. Appl. Math., 13 (1994), pp. 13–32.

[3] S. BERTOLUZZA AND G. NALDI, *A wavelet collocation method for the numerical solution of*

partial differential equations, Appl. Comput. Harmon. Anal., 3 (1996), pp. 1–9.
[4] G. Beylkin and N. Saito, *Multiresolutional representations using the auto-correlation functions of compactly supported wavelets*, IEEE Trans. Signal Process., 41 (1993), pp. 3584–3590.
[5] C. K. Chui, *Wavelets: A Mathematical Tool for Signal Analysis*, SIAM, Philadelphia, 1997.
[6] C. K. Chui and J. M. de Villiers, *Spline-wavelets with arbitrary knots on a bounded interval: Orthogonal decomposition and computational algorithms*, Commun. Appl. Anal., 2 (1998), pp. 457–486.
[7] C. K. Chui and C. Li, *Dyadic affine decompositions and functional wavelet transforms*, SIAM J. Math. Anal., 27 (1996), pp. 865–890.
[8] C. K. Chui and E. Quak, *Wavelets on a bounded interval*, in Numerical Methods of Approximation Theory, Vol. 6, D. Braess and L. L. Schumaker, eds., Birkhäuser-Verlag, Basel, 1992, pp. 53–75.
[9] A. Cohen, I. Daubechies, and P. Vial, *Wavelets on the interval and fast wavelet transforms*, Appl. Comput. Harmon. Anal., 1 (1993), pp. 54–81.
[10] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.
[11] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.
[12] G. Deslauriers and S. Dubuc, *Symmetric iterative interpolation processes*, Constr. Approx., 5 (1989), pp. 49–68.
[13] D. L. Donoho, *Interpolating Wavelet Transforms*, tech. report, Stanford University, Stanford, CA, 1992.
[14] D. L. Donoho, *Smooth wavelet decompositions with blocky coefficient kernels*, in Recent Advances in Wavelet Analysis, L. L. Schumaker and G. Webb, eds., Wavelet Anal. Appl. 3, Academic Press, Boston, 1994, pp. 259–308.
[15] S. Dubuc, *Interpolation through an iterative scheme*, J. Math. Anal. Appl., 114 (1986), pp. 185–204.
[16] K. M. Goosen, *Subdivision, Interpolation and Splines*, master's thesis, University of Stellenbosch, South Africa, 2000.
[17] A. Harten, *Discrete multi-resolution analysis and generalized wavelets*, Appl. Numer. Math., 12 (1993), pp. 153–192.
[18] A. Harten, *Multiresolution representation of data: A general framework*, SIAM J. Numer. Anal., 33 (1996), pp. 1205–1256.
[19] D. E. Knuth, *Digital Typography*, CSLI Publications, Stanford, CA, 1999.
[20] J. M. Lane and R. F. Riesenfeld, *A theoretical development for the computer generation of piecewise polynomial surfaces*, IEEE Trans. Pattern Anal. Machine Intelligence, 2 (1980), pp. 34–46.
[21] S. L. Lee, A. Sharma, and H. H. Tan, *Spline interpolation and wavelet construction*, Appl. Comput. Harmon. Anal., 5 (1998), pp. 249–276.
[22] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed., Academic Press, London, 1999.
[23] C. A. Micchelli, *Mathematical Aspects of Geometric Modeling*, SIAM, Philadelphia, 1995.
[24] C. A. Micchelli, *Interpolatory subdivision schemes and wavelets*, J. Approx. Theory, 86 (1996), pp. 41–71.
[25] E. Quak and N. Weyrich, *Decomposition and reconstruction algorithms for spline wavelets on a bounded interval*, Appl. Comput. Harmon. Anal., 1 (1994), pp. 217–231.
[26] P. Schröder and W. Sweldens, *Building Your Own Wavelets at Home*, Tech. Report IMI 1995:5, Department of Mathematics, University of South Carolina, 1995.
[27] W. Sweldens, *The lifting scheme: A custom-design construction of biorthogonal wavelets*, Appl. Comput. Harmon. Anal., 3 (1996), pp. 186–200.

SIAM J. MATH. ANAL.
Vol. 35, No. 2, pp. 453–491

COMPETING SPECIES NEAR A DEGENERATE LIMIT*

V. HUTSON†, Y. LOU‡, K. MISCHAIKOW§, AND P. POLÁČIK¶

Abstract. We consider a competitive reaction-diffusion model of two species in a bounded domain which are identical in all aspects except for their birth rates, which differ by a function g. Under a fairly weak hypothesis, the semitrivial solutions always exist. Our analysis provides a description of the stability of these solutions as a function of the diffusion rate μ and the difference between the birth rates g. In the case in which the magnitude of g is small we provide a fairly complete characterization of the stability in terms of the zeros of a single function. In particular, we are able to show that for any fixed number n, one can choose the difference function g from an open set of possibilities in such a way that the stability of the semitrivial solutions changes at least n times as the diffusion μ is varied over $(0, \infty)$. This result allows us to make conclusions concerning the existence of coexistence states. Furthermore, we show that under these hypotheses, coexistence states are unique if they exist.

The biological implication is that there is a delicate balance between resource utilization and dispersal rates which can have a dramatic impact with regards to extinction. Furthermore, we show that there is no optimal form of resource utilization. To be more precise, given a fixed diffusion rate and a particular spatially dependent utilization of resources which are expressed in terms of the birth rate, there always exists a birth rate, which on average is the same but differs pointwise, which allows the corresponding species to invade.

Key words. competing species, dispersal rate, spatial heterogeneity, reaction-diffusion systems

AMS subject classifications. 35K57, 92B40

DOI. 10.1137/S0036141002402189

1. Introduction.

We study the semilinear parabolic system

$$u_t = \mu \Delta u + u(\alpha(x) - u - v), \tag{1.1a}$$

$$v_t = \mu \Delta v + v(\beta(x) - u - v) \tag{1.1b}$$

in $\Omega \times (0, \infty)$, where Ω is a bounded region in $\mathbb{R}^N$ with smooth boundary $\partial\Omega$. On $\partial\Omega \times (0, \infty)$, we impose the homogeneous Neumann boundary condition

$$\frac{\partial u}{\partial n} = \frac{\partial v}{\partial n} = 0, \tag{1.1c}$$

where n is the outward unit normal vector on $\partial\Omega$. The diffusion rate μ is a positive constant, while the intrinsic growth rates $\alpha(x), \beta(x)$ are nonconstant functions in Ω.

*Received by the editors February 5, 2002; accepted for publication (in revised form) December 23, 2002; published electronically July 18, 2003. The visits of the first, second, and fourth authors to Georgia Tech were supported by the Center for Dynamical Systems and Nonlinear Studies at Georgia Tech. The visit of the fourth author to Ohio State University was supported by the Math Research Institute at Ohio State University.

http://www.siam.org/journals/sima/35-2/40218.html

†Department of Applied Mathematics, Sheffield University, Sheffield, S3 7RH, United Kingdom (V.Hutson@sheffield.ac.uk). This author was supported by NATO grant 930149.

‡Department of Mathematics, Ohio State University, Columbus, OH 43210 (lou@math.ohio-state.edu). This author was supported by NSF grant DMS-9801609 and the Ohio State University Seed Grant.

§Center for Dynamical Systems and Nonlinear Studies, School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (mischaik@math.gatech.edu). This author was supported by NSF grants DMS-9807395 and DMS-0107395.

¶School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (polacik@math.umn.edu). This author was supported by VEGA grant 1/7677/20.

To motivate this work, consider first the single equation

$$u_t = \mu\Delta u + u(\alpha(x) - u) \qquad \text{in } \Omega \times (0,\infty), \qquad \frac{\partial u}{\partial n} = 0 \qquad \text{on } \partial\Omega \times (0,\infty). \tag{1.2}$$

If $\int_\Omega \alpha > 0$ and α is Hölder continuous in $\overline{\Omega}$, then it is well known that there exists a unique positive equilibrium solution $\tilde{u}$ satisfying

$$\mu\Delta\tilde{u} + \tilde{u}(\alpha - \tilde{u}) = 0 \qquad \text{in } \ \Omega, \qquad \frac{\partial \tilde{u}}{\partial n} = 0 \qquad \text{on } \ \partial\Omega. \tag{1.3}$$

Furthermore, $\tilde{u}$ depends smoothly on μ, is asymptotically stable, and is the global attractor for the set of nontrivial, nonnegative initial conditions. In a slight abuse of language we shall refer to a solution with this property as a *global attractor*. When it is important to express the dependence on μ we will write $\tilde{u}(x,\mu)$. Equations of this form have attracted considerable attention in biology, where u might represent the density of an organism. The fact that $\tilde{u}$ is a global attractor can be interpreted as a statement that the species survives.

One of the fundamental driving forces in evolution is competition. An elementary, but not unreasonable, approach to understanding the relative fitness of two such organisms is to couple two equations of the form of (1.2) in a competitive manner. Using the classical Lotka–Volterra interaction terms we are led to the system

$$\begin{aligned} u_t &= \mu\Delta u + u(\alpha(x) - u - bv), \\ v_t &= \nu\Delta v + v(\beta(x) - cu - v), \end{aligned} \tag{1.4}$$

assuming Neumann boundary conditions again. In a suitable ordering, (1.4) defines a strongly monotone semiflow on the cone of continuous vector functions with both components positive. This has strong implications for the global dynamics. In particular, combining general results on monotone semiflows with special features of (1.4), one can prove that either there is a coexistence steady state (that is, a steady state with both components positive) or else for each solution in the cone one of the components tends to 0 as $t \to \infty$ (see [He, HSW]). Our goal here is to gain an understanding of the relationship between diffusion rates, spatial heterogeneity, and competitive coupling. In the tradition of bifurcation theory, this suggests looking for a system which is an organizing center for these parameters.

With this in mind consider the following degenerate version of (1.4):

$$\begin{aligned} u_t &= \mu\Delta u + u(\alpha(x) - u - v), \\ v_t &= \mu\Delta v + v(\alpha(x) - u - v). \end{aligned} \tag{1.5}$$

Obviously in this system u and v play an identical role, and hence for each fixed μ there is a set of nonnegative equilibria $\{(s\tilde{u}, (1-s)\tilde{u}) \mid 0 \le s \le 1\}$. Furthermore, this set of equilibria is the global attractor for the set of nontrivial nonnegative initial conditions. Thus, the ultimate goal would be to provide a complete unfolding of the degenerate system in the directions of the general systems taking the form of (1.4). This would provide us with detailed information concerning the existence, multiplicity, and stability of the equilibria, along with an understanding of the global dynamics.

Unfortunately, we are far from attaining this goal. However, this paper can be viewed as a natural addition to a series of work [HLMV, DHMP, HLM, HMP] in which we have obtained partial results. In particular, in [DHMP] the case of $b = c = 1$ and $\beta = \alpha$, α being a nonconstant function, was analyzed. Without loss

of generality assume that $\nu > \mu$ (otherwise (1.5) is recovered). The result, under no further hypotheses, is that $(\tilde{u}, 0)$ is the global attractor for the set of positive initial conditions. Observe that this implies that there are no other equilibria in the cone of positive functions. Biologically this suggests that if the two species interact identically with the environment, then the slower diffuser always survives and the faster diffuser is always driven to extinction. As is shown in [HMP], corresponding results are not always true if, in addition, one assumes that the reproductive rate α is periodic in time.

To gain an understanding of the importance of this phenomena as compared to the effects of relative strengths of competition, the case of $\nu > \mu$, $b > 1$, $c < 1$, and $\alpha = \beta$ was considered in [HLM]. Here we wish to consider similar questions; however, our focus is on the interaction between diffusion rates and the form of the heterogeneity of the environment. This justifies the assumption that $b = c = 1$. Furthermore, to make progress on this problem we have assumed that $\mu = \nu$, which results in system (1.1).

This system was studied in [HLMV], motivated by the following biological question. Consider a species with spatially dependent reproductive rate $\beta(x)$. The form of β may be regarded as reflecting the manner in which the resources are utilized. Suppose now that there is a mutation leading to another phenotype with a spatially distinct reproductive rate $\alpha(x)$ but otherwise identical. Typically, the initial population of the mutant species will be very small. Problem (1.1) may be taken to be a model representing the first stage in speciation, with different species differing only in the spatial dependence of their reproductive rate. A well-known example is the different beak size in "Darwin's finches" [G]. There are two key questions. First, under what circumstances will this mutant invade? Second, if it does invade, will it go to fixation, i.e., force extinction of the original phenotype, or will there be coexistence, which one would regard as a speciation event? Mathematically this leads to the system (1.1). The analysis of such a system when the difference between α and β is small occupies a major portion of the present paper. Our results go some way towards clarifying the question of the existence of a coexistence state and its dependence on the size of the difference between α and β and on the value of the diffusion coefficient. In a number of situations we are able to give a complete description of the global dynamics. We shall return to a discussion of the implications of our results at the end of this section.

Turning now to a description of the results, we start by recalling a few conclusions from [HLMV]. Let $\tilde{v}$ be a positive solution of

$$\mu\Delta\tilde{v} + \tilde{v}(\beta - \tilde{v}) = 0 \quad \text{in } \Omega, \qquad \frac{\partial\tilde{v}}{\partial n} = 0 \quad \text{on } \partial\Omega. \tag{1.6}$$

As was remarked earlier, if $\int_\Omega \beta > 0$, and β is Hölder continuous in $\overline{\Omega}$, then $\tilde{v}$ is uniquely determined and depends smoothly on μ. Observe that the solutions $\tilde{u}$ of (1.2) and $\tilde{v}$ of (1.6) define equilibria $(\tilde{u}, 0)$ and $(0, \tilde{v})$ of (1.1).

It was shown in [HLMV] that if $\int_\Omega \alpha > \int_\Omega \beta > 0$, then for large enough μ, $(\tilde{u}, 0)$ is a global attractor. Thus, in particular, $(\tilde{u}, 0)$ is asymptotically stable and $(0, \tilde{v})$ is unstable. From the biological perspective this implies that the species u always drives the species v to extinction, no matter what the initial data may be. On the other hand, [HLMV] also demonstrates that if $\alpha_+ - \beta_+$ (α_+ denotes the positive part of α) changes sign, then for small enough μ both semitrivial states $(\tilde{u}, 0)$ and $(0, \tilde{v})$ of (1.1) are unstable. This in turn implies that there is at least one stable coexistence state of (1.1), i.e., an equilibrium (u, v) of (1.1) with $u, v > 0$ on $\overline{\Omega}$.

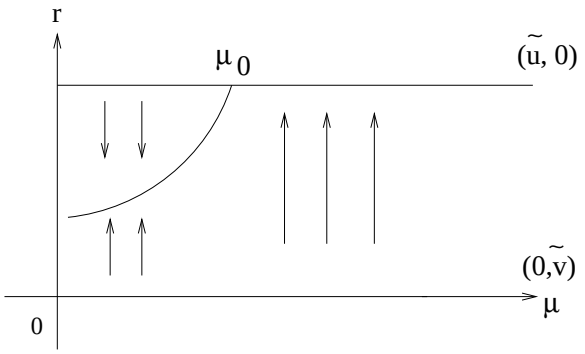


FIG. 1. *Expected bifurcation diagram of coexistence states of* (1.1): *for* $\mu > \mu_0$, $(\tilde{u}, 0)$ *is the global attractor of* (1.1); *for* $\mu < \mu_0$, (1.1) *has a coexistence state which is stable and possibly the global attractor of* (1.1). *In this and all other figures, the vertical axis is* $r = \|u\|/(\|u\| + \|v\|)$, *where* $\|\cdot\|$ *is the* L^2 *norm.*

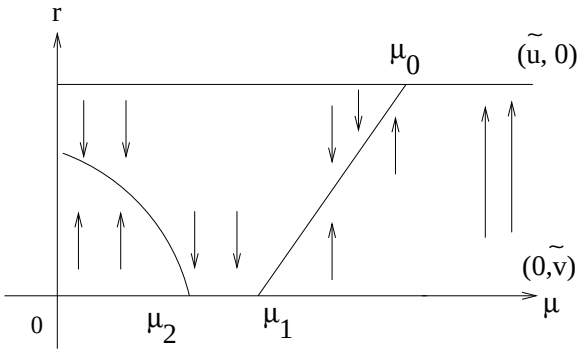


FIG. 2. *Another possible bifurcation diagram suggested by computation.*

The simplest interpretation of these results suggests the following scenario. As μ decreases from a large value, a branch of coexistence states of (1.1) bifurcates from $(\tilde{u}, 0)$ at some value μ_0 and remains in the interior of the positive cone for all $\mu < \mu_0$; see Figure 1. This would suggest some type of monotone relation between the stability of $(\tilde{u}, 0)$ and the rate of diffusion. Surprisingly, according to [HLMV], numerical computations show that this simple situation is not always the case. For reasonable choices of α and β, the branch of coexistence states bifurcating from $(\tilde{u}, 0)$ at μ_0 can reach $(0, \tilde{v})$, which becomes globally attracting for some range of μ as μ decreases. Eventually another branch of coexistence states of (1.1) bifurcates from $(0, \tilde{v})$ and remains in the positive cone for the rest of μ; see Figure 2.

It is worthwhile comparing these figures. In Figure 1, the species u can always invade when rare, that is, $(0, \tilde{v})$ is unstable for any μ; in Figure 2, the species u cannot invade when rare if $\mu \in (\mu_2, \mu_1)$, and indeed $(0, \tilde{v})$ is the global attractor of (1.1) for this range of μ. This is surprising because for the above-mentioned numerical computations species u was chosen to have better than average reproductive rate resource utilization than species v; i.e., it was assumed that $\int_\Omega \alpha > \int_\Omega \beta$. These numerical results also show that there need not be a monotone relation between stability (in the case of $(0, \tilde{v})$) and the rate of diffusion. Furthermore, the coexistent

states can appear and disappear as a function of the diffusion rate.

These observations suggest that the task of characterizing the equilibria for (1.1) will not be trivial. In fact, recall that, given the degenerate equation (1.5) and letting μ vary over $(0,\infty)$, one has a two-dimensional set of degenerate equilibria. Setting

$$\alpha(x) := \beta(x) + \tau g(x), \tag{1.7}$$

and treating both τ and μ as parameters, one might be inclined to conclude that the structure of the set of possible equilibria and their stability obtained from an arbitrary perturbation would be beyond a simple description. Remarkably, as we shall demonstrate, this is not the case.

Using (1.7) we rewrite (1.1) as

$$u_t = \mu\Delta u + u\big[\beta(x) + \tau g(x) - u - v\big], \tag{1.8a}$$

$$v_t = \mu\Delta v + v\big[\beta(x) - u - v\big], \tag{1.8b}$$

$$\frac{\partial u}{\partial n} = \frac{\partial v}{\partial n} = 0, \tag{1.8c}$$

where $\tau > 0$. The following assumption guarantees the existence of semitrivial equilibria for all diffusion rates.

A1. β is a C^1 nonconstant function on $\overline{\Omega}$ and $\int_\Omega \beta(x)\,dx > 0$.

A2. g is a C^1 function on $\overline{\Omega}$ satisfying $\int_\Omega g(x)\,dx \geq 0$.

On occasion we will make use of the slightly stronger condition:

A2$^+$. g is a C^1 function on $\overline{\Omega}$ satisfying $\int_\Omega g(x)\,dx > 0$.

Observe that if $g(x) > 0$ for all $x \in \Omega$, then $\alpha(x) > \beta(x)$. As was mentioned above, this case was studied in [HLMV]. Here we concentrate on the case when g changes sign. Although not assumed explicitly, this property is a consequence of the hypotheses of most of our theorems.

We remark that our regularity requirement on β, g can be relaxed at several places. For the sake of simplicity, however, we assume $\beta, g \in C^1(\overline{\Omega})$ throughout the paper.

Though unmotivated at the moment, it will soon become clear that the following function plays an essential role in all our analyses. Let $G : [0,\infty) \to \mathbb{R}$ be defined by

$$G(\mu) = \begin{cases} \displaystyle\int_\Omega g(x)\tilde{v}^2(x,\mu)\,dx & \text{if } \mu > 0, \\ \displaystyle\int_\Omega g(x)\beta_+^2(x)\,dx & \text{if } \mu = 0. \end{cases} \tag{1.9}$$

The fact that $\tilde{v}^2(x,\mu)$ is a smooth function of μ and [HLMV, Theorem 4.1] implies that G is differentiable on $(0,\infty)$ and continuous at 0. Note in particular that, under condition A2$^+$, G can assume a nonpositive value only if g changes sign.

We begin our study by analyzing the stability of $(0,\tilde{v})$ and confirm the numerical indication that the stability can change, and indeed can do so more than once, as the diffusion rate μ is varied.

THEOREM 1.1. *Suppose that* A1 *and* A2$^+$ *hold. If* $G(0) < 0$*, then there exists a unique* $\tau_0 > 0$ *such that*

(i) $\tau > \tau_0$ *implies that* $(0,\tilde{v})$ *is unstable for any* μ;

(ii) $\tau < \tau_0$ *implies that* $(0,\tilde{v})$ *changes stability*[1] *at least once as* μ *varies from zero to infinity. It changes stability at least twice, provided the following additional*

[1]We say a steady state changes stability at μ_0 if for $\mu \approx \mu_0$ it is stable on one side of μ_0 and unstable on the other side of μ_0.

condition is satisfied:

$$g > 0,\ \beta > 0 \quad \text{on a nonempty set} \quad \Omega^+ \subset \Omega. \tag{1.10}$$

As mentioned earlier, the instability of $(0, \tilde{v})$ means that the species u with low density can successfully invade. Theorem 1.1 qualitatively illustrates how the invasion of species relies on its diffusion rate and the difference between its intrinsic growth rate and that of its competitor. The invasion of species has always been an active and important subject in biology, and we refer to [SK, CC] and references therein for some recent biological and mathematical developments.

Theorem 1.1 immediately raises two questions:

1. What are the values of μ where $(0, \tilde{v})$ changes stability?
2. If there are coexistence states bifurcating from $(0, \tilde{v})$ or $(\tilde{u}, 0)$, what can be said about their stability and how does this influence the global dynamics?

In general, these are hard questions to answer, but for small τ we shall give fairly complete answers, and shall obtain some partial understanding when τ is large.

For small τ, the roots of G approximate the values of μ where $(0, \tilde{v})$ and $(\tilde{u}, 0)$ change stability. Stated differently, if $(0, \tilde{v})$ or $(\tilde{u}, 0)$ changes stability at $\mu = \mu_\tau$, then as $\tau \to 0+$, either $\mu_\tau \to 0$ or $\mu_\tau \to \bar{\mu}$ with $G(\bar{\mu}) = 0$. With some further minor assumptions on G the converse is also true. The precise statement is given in the following theorem (see Figure 3).

THEOREM 1.2. *Assume* A1 *and* A2. *Let* $G^{-1}(0) = \{\mu_1 < \cdots < \mu_k\}$. *Furthermore, assume that* $G'(\mu_i) \neq 0$ *for every* $1 \le i \le k$. *Fix* $\eta \in (0, \mu_1)$. *Then there exists some* $\tau_0 > 0$ *and functions* $\mu_i^*(\tau), \mu_{i,*}(\tau) : (0, \tau_0) \to (\eta, \infty)$, $i = 1, \dots, k$, *such that*

$$\mu_{1,*}(\tau) < \mu_1^*(\tau) < \mu_{2,*}(\tau) < \mu_2^*(\tau) < \cdots < \mu_{k,*}(\tau) < \mu_k^*(\tau), \tag{1.11a}$$

$$\lim_{\tau \to 0} \mu_{i,*}(\tau) = \lim_{\tau \to 0} \mu_i^*(\tau) = \mu_i, \qquad 1 \le i \le k, \tag{1.11b}$$

and for $\mu \in [\eta, \mu_k^*(\tau))$, (1.8) *has a coexistence state if and only if* $\mu \in \bigcup_{i=1}^k (\mu_{i,*}(\tau), \mu_i^*(\tau))$. *Moreover, for* $\mu \ge \eta$, *any coexistence state of* (1.8) *is the global attractor of* (1.8).

Under the stricter assumption A2$^+$, $\mu_k^*(\tau)$ *can be chosen such that* (1.8) *has no coexistence states for* $\mu > \mu_k^*(\tau)$, *and* $(\tilde{u}, 0)$ *is the global attractor.*

We remark that A2 is not needed for the first statement of the theorem. If g is strictly positive, then $G^{-1}(0) = \emptyset$, and hence the theorem is vacuously true. A more subtle point is the assumption that $G'(\mu_i) \neq 0$. Recall that μ enters into the definition of G indirectly, via $\tilde{v}(x, \mu)$; thus it is neither evident that this is a generic condition nor that G can have arbitrarily many roots. We clarify these issues in the next proposition.

Let $C^1(\overline{\Omega})$ be equipped with the standard norm:

$$\|g\| = \sup_{x \in \Omega} (|g(x)| + |Dg(x)|).$$

Let U be an open set in $C^1(\overline{\Omega})$. We say that a statement holds for generic $g \in U$ if the set of functions g for which the statement holds contains the intersection of countably many open and dense subsets of U. Such an intersection is dense by Baire's theorem.

PROPOSITION 1.3.

(i) *Let* β *satisfy* A1. *Then for generic* $g \in C^1(\overline{\Omega})$, 0 *is a regular value of the function* G; *that is,* $G'(\mu) \neq 0$ *whenever* $G(\mu) = 0$.

(ii) *For generic* $\beta \in U = \{\beta \in C^1(\overline{\Omega}) : \int_\Omega \beta > 0\}$ *the following holds: given any nonnegative integer* k, *there exists* $g \in U$ *such that* G *has at least* k *nondegenerate zeros.*

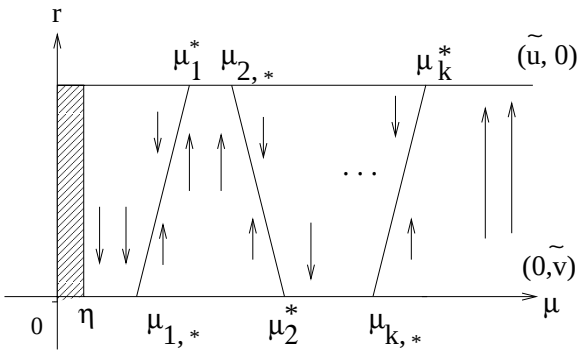


FIG. 3. *Bifurcation diagram of coexistence states of* (1.8) *when* τ *is small and* μ *is bounded away from zero.*

Theorem 1.2 handles the case where τ is small and μ is bounded away from zero. What happens when both τ and μ are small? As was indicated earlier, for fixed τ, both $(\tilde{u}, 0)$ and $(0, \tilde{v})$ are unstable if μ is small enough. In view of Figure 3, for fixed τ, either $(\tilde{u}, 0)$ or $(0, \tilde{v})$ will change stability at least once when $\mu \in (0, \eta]$. The following result gives a characterization of this bifurcation point when τ and μ are both small.

THEOREM 1.4. *Suppose that* A1 *and* A2 *hold and that* $\beta > 0$ *in* $\bar{\Omega}$. *Denote by* μ_1 *the smallest positive root of* G *and assume that* $G'(\mu_1) \neq 0$. *Let* $\mu_1^*(\tau)$ *and* $\mu_{1,*}(\tau)$ *be as in Theorem* 1.2.

(i) *If* $G(0) < 0$, *there exists* $\tau_0 > 0$ *such that for every* $\tau \leq \tau_0$, $(\tilde{u}, 0)$ *is unstable for* $\mu < \mu_1^*(\tau)$. *Furthermore, there exists a unique* $\mu_0(\tau) \in (0, \mu_{1,*}(\tau))$ *such that* $(0, \tilde{v})$ *is unstable for* $\mu < \mu_0(\tau)$ *and stable for* $\mu \in (\mu_0(\tau), \mu_{1,*}(\tau))$ *(see Figure* 4a*). Moreover,* $\mu_0(\tau)$ *satisfies*

$$\lim_{\tau \to 0} \frac{\tau}{\mu_0(\tau)} = \inf_{\left\{\psi \in H^1 : \int_\Omega g\beta^2\psi^2 > 0\right\}} \frac{\int_\Omega \beta^2 |\nabla\psi|^2}{\int_\Omega g\beta^2\psi^2} > 0. \tag{1.12}$$

(ii) *If* $G(0) > 0$, *there exists* $\tau_0 > 0$ *such that for every* $\tau \leq \tau_0$, $(0, \tilde{v})$ *is unstable for* $\mu < \mu_1^*(\tau)$. *Furthermore, there exists a unique* $\mu_0(\tau) \in (0, \mu_{1,*}(\tau))$ *such that* $(\tilde{u}, 0)$ *is unstable for* $\mu < \mu_0(\tau)$ *and stable for* $\mu \in (\mu_0(\tau), \mu_{1,*}(\tau))$ *(see Figure* 4b*). Moreover,* $\mu_0(\tau)$ *satisfies*

$$\lim_{\tau \to 0} \frac{\tau}{\mu_0(\tau)} = \inf_{\left\{\psi \in H^1 : \int_\Omega g\beta^2\psi^2 < 0\right\}} \frac{\int_\Omega \beta^2 |\nabla\psi|^2}{-\int_\Omega g\beta^2\psi^2} > 0. \tag{1.13}$$

From Figure 4a we see that a branch of coexistence states bifurcate from $(0, \tilde{v})$ at $\mu = \mu_0(\tau)$. We suspect that if $G(0) < 0$ and $\tau \ll 1$, for $0 < \mu < \mu_0(\tau)$, (1.8) has a unique coexistence state and it is the global attractor of (1.8). We also believe that for $\mu \in (\mu_0, \mu_{1,*}(\tau))$, $(0, \tilde{v})$ is the global attractor of (1.8). However, these questions remain open.

Again, motivated by biology the following question is natural. Given a dispersal rate μ, is there a birth rate $\beta(x)$ that is "optimal" in the sense that an invading mutant with intrinsic growth rate $\beta(x) + \tau g(x)$ necessarily dies out? Mathematically, this is equivalent to asking that $(0, \tilde{v}(x, \mu))$ be stable for all sufficiently small τ, no matter how g is chosen. In order not to bias the result, e.g., by allowing one phenotype to have a higher reproductive rate at every point, it is reasonable to impose an additional

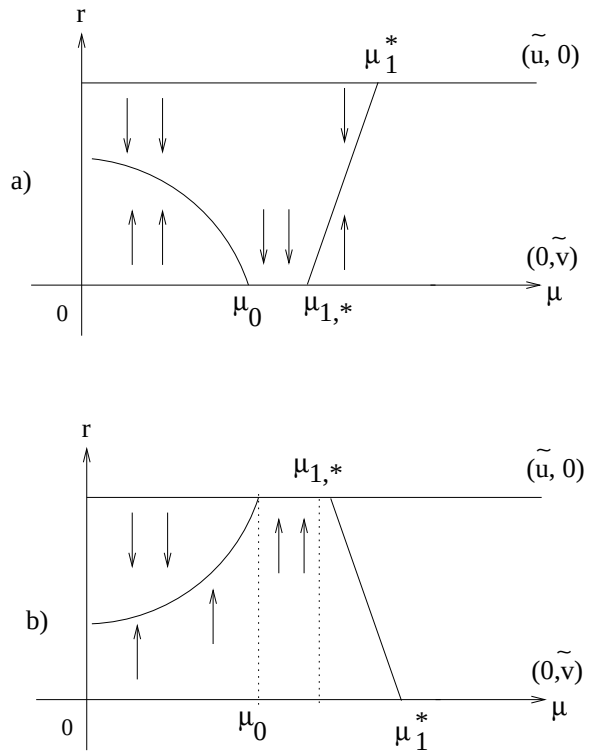


FIG. 4. *Bifurcation diagrams of coexistence states for* $\tau \ll 1$, $\mu < \mu_{1,*}$ *and* (a) $\int_\Omega g\beta^2 < 0$, (b) $\int_\Omega g\beta^2 > 0$.

global "fairness" assumption $\int_\Omega g(x)\,dx = 0$. However, Theorem 1.2 indicates that the stability of $(0, \tilde{v}(x,\mu))$ is determined by the sign of $G(\mu)$. Therefore, if $(0, \tilde{v}(x,\mu))$ is stable under a perturbation in the direction of $g(x)$, then it is unstable under a perturbation in the opposite direction $-g(x)$. In particular, given a particular environment and dispersal rate, there is no optimal birth rate.

The instability of $(0, \tilde{v}(x,\mu))$ indicates that a new mutant can easily invade and suggests the possibility of coexistence. This leads to the following question: for any fixed μ, can one always find g such that there is a coexistence for all small $\tau > 0$? Interestingly, the answer depends on the relation of μ to a single value μ^* depending only on β and Ω. Of course, as we know from the previous results, if the coexistence is to hold for all small τ, then g must be chosen such that $G(\mu) = \int_\Omega \tilde{v}^2(\cdot,\mu)g = 0$.

THEOREM 1.5. *Assume* A1. *There exists a unique* $\mu^* > 0$, *depending only on* β *and* Ω, *with the following properties:*

(i) *If* $\mu > \mu^*$, (1.8) *has no coexistence state for any* g *satisfying* A2, *provided* $0 < \tau < \tau_0$, *where* $\tau_0 = \tau_0(g)$.

(ii) *If* $\mu < \mu^*$, *then there exists a nonempty open subset* U *of*

$$\left\{ g \in C^1(\overline{\Omega}) : \int_\Omega \tilde{v}^2(\cdot,\mu)g = 0 \text{ and } \int_\Omega g > 0 \right\}$$

(with the topology induced from $C^1(\overline{\Omega})$*) such that* (1.8) *has a coexistence state that is the global attractor of* (1.8), *provided* $g \in U$ *and* $0 < \tau < \tau_0$, *where* $\tau_0 = \tau_0(g)$.

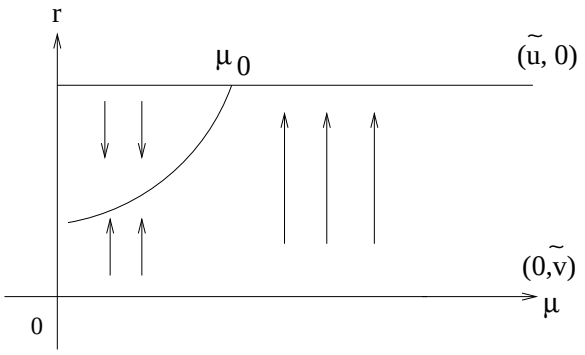


FIG. 5. *Bifurcation diagram of coexistence states of* (1.8) *when* $\tau \gg 1$.

Thus $\mu = \mu^*$ is a critical value for the diffusion rate above which if an invasion can occur, then the invading mutant necessarily goes to fixation. Below μ^*, invasions leading to either fixations or coexistence are possible, depending on the choice of g.

From the point of view of biology, a possible objection to Theorem 1.5 is that the requirement $\int_\Omega \tilde{v}^2(\cdot,\mu)g = 0$ is not generic, and therefore statement (ii) is not applicable. However, stability properties of coexistence equilibria, examined in detail in section 5, guarantee their persistence. Thus if one chooses τ small but bounded away from zero, then coexistence occurs for nonempty open sets of functions g (not restricted to U) and diffusion rates μ. Furthermore, if one thinks of the change in the birth rate as being caused by a mutation, then it makes sense to think of τ as having some finite, though possibly small, size.

We now turn to the case of large τ. In this setting, assuming A2$^+$, it is not hard to show that $(0,\tilde{v})$ is unstable for any μ. On the other hand, $(\tilde{u},0)$ changes stability at least once as μ varies from ∞ to 0. This prompts the following question: given $\tau \gg 1$, what is the range of values of μ for which $(\tilde{u},0)$ changes stability? To this end, set

$$\Omega^+ = \{x \in \Omega : g(x) > 0\}, \quad \Omega^- = \{x \in \Omega : g(x) < 0\}, \quad \Gamma = \{x \in \bar{\Omega} : g(x) = 0\}.$$

In the next theorem we shall assume that the closure of Ω^- is a nonempty subset of Ω and that $\beta > 0$ on $\bar{\Omega}$. This guarantees that there is a unique value $\mu_0 > 0$ such that the linear problem

$$\mu_0 \Delta\phi + \beta(x)\phi = 0 \quad \text{in } \Omega^-, \qquad \phi > 0 \quad \text{in } \Omega^-, \qquad \phi = 0 \quad \text{on } \Gamma \tag{1.14}$$

has a solution.

THEOREM 1.6 (see Figure 5). *Assume* A1, A2$^+$, *that the closure of* Ω^- *is a nonempty subset of* Ω, $g \in C^2(\overline{\Omega})$, $\nabla g \neq 0$ *on* Γ, *and that* $\beta > 0$ *on* $\bar{\Omega}$. *Let* μ_0 *be the value introduced above. Then the following statements hold true:*

(i) *There exists* $\tau_0 > 0$ *such that if* $\tau \geq \tau_0$, *then* $(0,\tilde{v})$ *is unstable for any* $\mu > 0$.

(ii) *For any* $\epsilon > 0$, *there exists* $\tau_1 = \tau_1(\epsilon)$ *such that if* $\tau \geq \tau_1$, *then* $(\tilde{u},0)$ *is unstable for* $\mu \leq \mu_0 - \epsilon$ (*this implies that* (1.8) *has a stable coexistence state for* $\mu \leq \mu_0 - \epsilon$), *and* $(\tilde{u},0)$ *is the global attractor of* (1.8) *for* $\mu \geq \mu_0 + \epsilon$.

To have an overview of the results presented here, it may be helpful to the reader to refer to Figure 6, which is sketched from computed results. This shows a sequence of bifurcation curves for various values of τ. This sketch suggests of course a great deal more than we have proved.

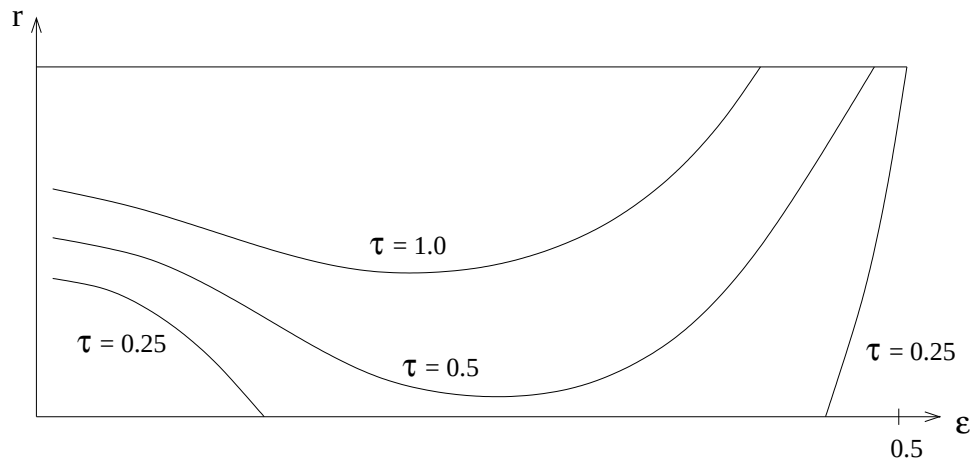


FIG. 6. *Sketch based on computation with* $\beta(x) = 1-\cos(\pi x)$, $g(x) = 0.25+\cos(\pi x)$, $\Omega = (0,1)$, *and* $\epsilon = \sqrt{\mu}$.

We conclude with some remarks on the implications for the biological considerations which motivate this model. The first issue highlighted by the analysis, and one which we suggest is very striking, is the following. We have shown that for a large class of functions g, and for small τ, representing small variations of the phenotype, the stability of the two phenotypes varies with diffusion in a complex manner, and one which appears to be highly unintuitive. In particular, there is no monotonicity with respect to the diffusion rate μ. As is shown in Figure 3 (and proved in Theorem 1.2 and Proposition 1.3) the stability may change back and forth several times as μ changes. Furthermore, unless μ is large, this has no relation to the total resource utilizations, $\int_\Omega \beta$ and $\int_\Omega(\beta + \tau g)$. This is surprising. From the observer's point of view, this suggests that without careful measurement and elaborate experiments, it is totally unpredictable which species will survive.

A second comment is that mutation leads to multiple opportunities for coexistence and thus potentially for speciation. For small τ, which is more realistic biologically, this only happens for narrow ranges of μ, but as τ increases, the ranges widen. Nonetheless, it suggests that there is no surprise in finding a large range of coexisting phenotypes which differ only in one, sometimes small, manner, which is the precise manner in which they utilize the resources of the environment. Indeed, Theorem 1.5 ensures that, given μ, there is a large class of functions for which there will be (stable) coexistence.

Finally, we make a comment on the role of the diffusion μ as one of the bifurcation parameters. This appears to be a relevant mathematical tool for modeling situations when changes in the environment or mutation affect the diffusion rates of the species. One can think of the direct effects caused by changes in climate (temperature, rainfall) or indirect effects of changes in the biotic environment (resources). The bifurcation problem is appropriate in the study of long-term changes—those that occur at a much slower rate than the growth of the species. If they occur on a comparable time scale, a diffusion problem with time-dependent diffusion coefficients may be a more appropriate model.

This paper is organized as follows: In section 2 we summarize basic material concerning local and global stability of equilibria. In section 3 we prove Theorem 1.1.

We study the $\tau \ll 1$ case in section 4, where Theorems 1.2–1.5 are established. In section 5, we consider the $\tau \gg 1$ case and prove Theorem 1.6. Some tedious computations needed in the proof of Theorem 1.2 are carried out in Appendix A. Appendix B is devoted to the proof of Proposition 1.3.

2. Preliminaries. In this section we summarize basic properties of equilibria of (1.8) related to their stability. We start by making more precise the statements given in the introduction regarding the scalar equation

$$\mu\Delta\tilde{u} + \tilde{u}(\beta - \tilde{u}) = 0 \quad \text{in } \Omega, \qquad \frac{\partial\tilde{u}}{\partial n} = 0 \quad \text{on } \partial\Omega. \tag{2.1}$$

It is well known that for each $\beta \in C(\overline{\Omega})$ with $\int_\Omega \beta > 0$, there exists a unique positive solution $\tilde{u} \in H^2(\Omega)$ of this problem; it is a classical solution if β is Hölder continuous. It is further known (see, e.g., [He, sect. III.28]) that $\tilde{u}$ is the global attractor for the corresponding parabolic problem: the solution of (1.2) with any nonnegative nonzero initial condition $u_0 \in C(\overline{\Omega})$ converges as $t \to \infty$ to $\tilde{u}$. Also $\tilde{u}$ is linearly stable, which is to say that all eigenvalues of the linearized operator $-\Delta - (\beta - 2\tilde{u})$, under Neumann boundary condition, are positive. In particular $\tilde{u}$ is a nondegenerate equilibrium; hence, by the implicit function theorem, it depends analytically (as a $W^{2,p}(\Omega)$-valued function for any $p > 1$) on $\mu \in (0, \infty)$ and $\beta \in C(\overline{\Omega})$.

At several places below we shall use asymptotic behavior of the positive solution $\tilde{v} = \tilde{v}(\cdot, \mu)$ of (2.1) when $\mu \to 0$ or ∞. The following properties have been proved in [HLMV]:

$$\tilde{v} \to \beta_+ \quad \text{in } L^\infty(\Omega) \qquad \text{as } \mu \to 0+, \tag{2.2a}$$

$$\tilde{v} \to \frac{1}{|\Omega|}\int_\Omega \beta \quad \text{in } L^\infty(\Omega) \quad \text{as } \mu \to \infty. \tag{2.2b}$$

Let us now turn to the system (1.8). By standard theory (see, e.g., [L, H1]), it defines a smooth dynamical system on

$$\mathcal{X} := C(\overline{\Omega}) \times C(\overline{\Omega}).$$

We understand the notions of stability and asymptotic stability of equilibria of (1.8) with respect to the topology of $\mathcal{X}$. We restrict our attention to physically relevant solutions, that is, solutions with nonnegative initial conditions. They are positive for all times by the maximum principle. We say an equilibrium (u_e, v_e) is the *global attractor* if it is stable and for each nontrivial $(u_0, v_0) \in \mathcal{X}$ with $u_0 \geq 0$, $v_0 \geq 0$ one has $(u(\cdot,t), v(\cdot,t)) \to_{\mathcal{X}} (u_e, v_e)$ as $t \to \infty$, where $(u(\cdot,t), v(\cdot,t))$ is the solution of (1.8) with the initial conditions

$$u(\cdot, 0) = u_0, \quad v(\cdot, 0) = v_0.$$

An equilibrium (u_e, v_e) with both components positive is called a *coexistence state* (or coexistence equilibrium); (u_e, v_e) is a *semitrivial equilibrium* if one component is positive and the other one is zero.

Let us now assume hypothesis A1 to be satisfied. Then system (1.8) has two semitrivial equilibria $(\tilde{u}, 0)$ and $(0, \tilde{v})$ for each $\mu > 0$ and each τ sufficiently small (for each $\tau > 0$ if also A2 is satisfied). They are given by the unique solutions of the corresponding scalar equations.

Due to the competitive Lotka–Volterra structure of the system, knowledge of equilibria and their stability is in some cases sufficient for complete understanding of the global dynamics of (1.8). We recall a few results to that effect (see [He, Chap. IV]):

(a) If there is no coexistence state, then one of the semitrivial equilibria is unstable and the other one is the global attractor.
(b) If there is a unique coexistence state and it is stable, then it is the global attractor (in particular, both semitrivial equilibria are unstable).
(c) If all coexistence states are asymptotically stable, then there is at most one of them, so either (a) or (b) applies.

Let us now discuss in some detail the linearized stability of an equilibrium (u, v). We thus consider the eigenvalue problem

$$\mu\Delta\varphi + (\beta + \tau g - 2u - v)\varphi + (-u)\psi = -\lambda\varphi \quad \text{in } \Omega, \tag{2.3a}$$

$$\mu\Delta\psi + (-v)\varphi + (\beta - u - 2v)\psi = -\lambda\psi \quad \text{in } \Omega, \tag{2.3b}$$

$$\frac{\partial\varphi}{\partial n} = \frac{\partial\psi}{\partial n} = 0 \qquad \text{on } \partial\Omega. \tag{2.3c}$$

It is well known (see, e.g., [He]) that if (u, v) is a coexistence state, one can put this eigenvalue problem in the context of spectral theory of compact strongly positive operators with respect to the order cone

$$\mathcal{C} = \{(\phi, \psi) \in \mathcal{X} : \phi \geq 0,\ \psi \leq 0\}.$$

In particular, using the Krein–Rutman theorem [De, He], one can show that (2.3) has an eigenvalue λ (called *the principal eigenvalue of* (2.3)), which has the following properties: it is real, algebraically simple, and all other eigenvalues have their real part greater than λ. Moreover, λ corresponds to an eigenfunction in the interior of $\mathcal{C}$, and it is the only eigenvalue with an eigenfunction in $\mathcal{C}$. The linearized stability criterion for (u, v) can be expressed in terms of the principal eigenvalue: (u, v) is asymptotically stable if $\lambda > 0$; it is unstable if $\lambda < 0$.

When (u, v) is a semitrivial equilibrium, for example, $(u, v) = (u, 0)$, then (2.3) simplifies to a triangular system

$$\mu\Delta\varphi + (\beta + \tau g - 2u)\varphi + (-u)\psi = -\lambda\varphi \quad \text{in } \Omega, \tag{2.4a}$$

$$\mu\Delta\psi + (\beta - u)\psi = -\lambda\psi \quad \text{in } \Omega, \tag{2.4b}$$

$$\frac{\partial\varphi}{\partial n} = 0 \qquad \text{on } \partial\Omega, \tag{2.4c}$$

$$\frac{\partial\psi}{\partial n} = 0 \qquad \text{on } \partial\Omega. \tag{2.4d}$$

Again one can examine the eigenvalues using a suitable positive operator (see [He]). This time, however, such a positive operator is not strongly positive (the reason is that the second equation decouples), and one gets weaker conclusions from the general theory. Nonetheless, employing the triangular structure, one can still establish the existence of a principal eigenvalue, that is, a simple real eigenvalue which is smaller than the real part of any other eigenvalue. Specifically, the principal eigenvalue coincides with the principal eigenvalue of the scalar eigenvalue problem (2.4b), (2.4d) (see [HMP, Lem. 3.2] for the proof of this fact; the corresponding eigenfunction for the system is $(0, -\psi) \in \mathcal{C}$, where $\psi > 0$ is the principal eigenfunction of (2.4b), (2.4d)).

Similarly, if $(u, v) = (0, v)$, then the principal eigenvalue of (2.3) coincides with the principal eigenvalue of the scalar problem

$$\mu\Delta\varphi + (\beta + \tau g - v)\varphi = -\lambda\varphi \quad \text{in } \Omega, \qquad \frac{\partial\varphi}{\partial n} = 0 \quad \text{on } \partial\Omega. \tag{2.5}$$

We remark that since the principal eigenvalue is always simple, it inherits the smoothness properties of the data in the problem. In particular, we consider below the principal eigenvalue of (2.5), $v = \tilde{v}(\cdot, \mu)$ being the positive solution of (1.6). As remarked above, $v(\cdot, \mu)$ is analytic in μ as a $W^{2,p}(\Omega)$-valued function (for any $p > 1$). Therefore, by standard analytic perturbation theory (see [K1]), the principal eigenvalue of (2.5) is an analytic function of $\tau > 0$ and $\mu > 0$.

3. Invasion of new species. As pointed out in the introduction, biologically it is important to understand the invasion of new species with low density. Mathematically, the invasion of species is determined by the stability of $(0, \tilde{v})$. This section is devoted to the study of this, and in particular we are interested in discovering conditions under which $(0, \tilde{v})$ will change stability twice (or more) as μ varies from ∞ to 0. The principal aim is to prove Theorem 1.1.

Recall that $\tilde{v} = \tilde{v}(\cdot, \mu) > 0$ is the unique positive solution of (1.6). As mentioned in section 2, for the stability of $(0, \tilde{v})$, it suffices to determine the sign of the principal eigenvalue, denoted by $\lambda_1 = \lambda_1(\mu, \tau)$, of the linear eigenvalue problem

$$\mu\Delta\varphi + (\beta + \tau g - \tilde{v})\varphi = -\lambda\varphi \quad \text{in } \Omega, \qquad \frac{\partial\varphi}{\partial n}\Big|_{\partial\Omega} = 0. \tag{3.1}$$

For any $\mu > 0$, set

$$C(\mu) := \inf_{\left\{\phi \in H^1(\Omega) : \int_\Omega g\tilde{v}^2\phi^2 > 0\right\}} \frac{\int_\Omega \tilde{v}^2|\nabla\phi|^2}{\int_\Omega g\tilde{v}^2\phi^2}. \tag{3.2}$$

LEMMA 3.1. *The following holds under the standing assumption $\tau > 0$:*

$$\lambda_1 > 0 \Leftrightarrow \tau < \mu C(\mu); \tag{3.3a}$$

$$\lambda_1 = 0 \Leftrightarrow \tau = \mu C(\mu); \tag{3.3b}$$

$$\lambda_1 < 0 \Leftrightarrow \tau > \mu C(\mu). \tag{3.3c}$$

Proof. Let $\varphi_1 > 0$ be an eigenfunction corresponding to the principal eigenvalue λ_1 of (3.1). Set $\psi = \varphi_1/\tilde{v}$. It is straightforward to see that $\psi > 0$ satisfies

$$\mu\nabla\cdot(\tilde{v}^2\nabla\psi) + \tau g\tilde{v}^2\psi = -\lambda_1\tilde{v}^2\psi \qquad \text{in } \Omega, \tag{3.4a}$$

$$\frac{\partial\psi}{\partial n}\Big|_{\partial\Omega} = 0. \tag{3.4b}$$

Suppose first that $\int_\Omega g\tilde{v}^2 > 0$. Dividing (3.4a) by ψ and integrating in Ω we have

$$\lambda_1\int_\Omega \tilde{v}^2 = -\left[\mu\int_\Omega \frac{\tilde{v}^2}{\psi^2}\,|\nabla\psi|^2 + \tau\int_\Omega g\tilde{v}^2\right] < 0, \tag{3.5}$$

so that $\lambda_1 < 0$. On the other hand, it is obvious (by taking $\phi = 1$ in (3.2)) that $C(\mu) = 0$. This proves (3.3) when $\int_\Omega g\tilde{v}^2 > 0$. The case $\int_\Omega g\tilde{v}^2 = 0$ can be handled in a similar way.

Next assume $\int_\Omega g\tilde{v}^2 < 0$. By the variational characterization (3.2) we know (see [F]) that $C(\mu) > 0$, and there exists $\tilde{\psi} > 0$ such that $\tilde{\psi}$ satisfies

$$\nabla\cdot(\tilde{v}^2\nabla\tilde{\psi}) + C(\mu)g\tilde{v}^2\tilde{\psi} = 0 \qquad \text{in } \Omega, \tag{3.6a}$$

$$\frac{\partial\tilde{\psi}}{\partial n}\Big|_{\partial\Omega} = 0. \tag{3.6b}$$

Notice that $\lambda_1 = \lambda_1(\mu, \tau)$, as a function of τ, is concave [K1]. It is easy to see that $\lambda_1(\mu, 0) = 0$ from (3.4) (with corresponding eigenfunction $\psi = 1$), and it follows from (3.6) that $\lambda_1(\mu, \mu C(\mu)) = 0$. Therefore $\lambda_1(\mu, \tau) > 0$ for $0 < \tau < \mu C(\mu)$, and $\lambda_1(\mu, \tau) < 0$ for $\tau > \mu C(\mu)$. This completes the proof of Lemma 3.1. □

Proof of Theorem 1.1. The assumptions of the theorem and asymptotic properties (2.2) yield

$$\lim_{\mu \to 0+} G(\mu) = \int_\Omega g\beta_+^2 < 0, \tag{3.7a}$$

$$\lim_{\mu \to +\infty} G(\mu) = \left(\int_\Omega g\right)\left(\int_\Omega \beta/|\Omega|\right)^2 > 0. \tag{3.7b}$$

Hence $G(\mu) = 0$ has at least one positive root. Let $\underline{\mu} \le \overline{\mu}$ denote the smallest and largest positive root of G, respectively. Recall that $C(\mu) > 0$ if $G(\mu) < 0$, and $C(\mu) = 0$ if $G(\mu) \ge 0$ (see the proof of Lemma 3.1). This ensures that $C(\mu) = 0$ for $\mu \ge \bar{\mu}$, and $C(\mu) > 0$ for $0 < \mu < \underline{\mu}$.

Choose $\phi = 1/\tilde{v}$ in (3.2): since $\int_\Omega g > 0$ and $\tilde{v} \to \beta_+$ uniformly as $\mu \to 0$, we see that

$$0 \le \mu C(\mu) \le \frac{-\mu \int_\Omega \nabla\left(\frac{1}{\tilde{v}}\right)\nabla\tilde{v}}{\int_\Omega g} = \frac{\mu \int_\Omega \frac{1}{\tilde{v}}\,\Delta\tilde{v}}{\int_\Omega g} = \frac{\int_\Omega(\tilde{v} - \beta)}{\int_\Omega g} \to \frac{\int_\Omega(\beta_+ - \beta)}{\int_\Omega g}$$

as $\mu \to 0$. This ensures that $\mu C(\mu)$ is bounded. Moreover, if (1.10) is satisfied, then

$$\mu C(\mu) \to 0 \qquad \text{as } \mu \to 0+\,. \tag{3.8}$$

Indeed, choose a smooth nonzero function ϕ with compact support in Ω^+. Then for all small $\mu > 0$, we have

$$\int_\Omega g\tilde{v}^2\phi^2 = \int_{\Omega^+} g\tilde{v}^2\phi^2 \approx \int_{\Omega^+} g\beta^2\phi^2 > 0$$

(so that ϕ is an admissible test function in (3.2)) and

$$C(\mu) \le \frac{\int_{\Omega^+} \tilde{v}^2|\nabla\phi|^2}{\int_{\Omega^+} g\tilde{v}^2\phi^2} \to \frac{\int_{\Omega^+} \beta^2|\nabla\phi|^2}{\int_{\Omega^+} g\beta^2\phi^2};$$

hence $\mu C(\mu) \to 0$.

We verify that

$$\tau_0 := \sup_{0<\mu<\infty} \mu C(\mu) \tag{3.9}$$

has the properties stated in the theorem. If $\tau > \tau_0$, then $\tau > \mu C(\mu)$ for any $\mu \in (0, \infty)$. By Lemma 3.1, this implies that $(0, \tilde{v})$ is unstable for any μ. On the other hand, if $0 < \tau < \tau_0$, then $\tau - \mu C(\mu)$ has at least one root and, by (3.8), it has at least two roots if condition (1.10) is satisfied. We claim that $\mu \mapsto \tau - \mu C(\mu)$ actually changes sign, and it does so at least twice if (1.10) is satisfied. By Lemma 3.1, this proves that $(0, \tilde{v})$ changes stability, and it does so at least twice if (1.10) holds.

To prove the claim, it suffices to exclude the possibility that $\mu C(\mu) \equiv \tau$ in some interval of μ, say $[\mu_1, \mu_2]$. We argue by the contradiction: if this is the case, by

Lemma 3.1 we see that $\lambda_1(\mu,\tau)\equiv 0$ for $\mu\in[\mu_1,\mu_2]$. Since λ_1 is an analytic function of μ (see section 2), $\lambda_1\equiv 0$ for any $\mu>0$. However, (2.2b), (3.5), and A2$^+$ imply that $\lambda_1(\mu,\tau)<0$ for large values of μ—a contradiction. □

The above proof shows that if $\mu>\overline{\mu}$, then for $\tau>0$ we have $\tau>\mu C(\mu)$, and hence $(0,\tilde{v})$ is unstable. For future reference we state this in a more precise form as the following.

COROLLARY 3.2. *Assume* A1, A2$^+$ *are satisfied. Further assume that for some* $\overline{\mu}>0$ *one has* $G(\mu)\neq 0$ *for* $\mu>\overline{\mu}$ *(hence* $G(\mu)>0$ *for* $\mu>\overline{\mu}$*, by* (3.7b)*). Then for any* $\mu>\overline{\mu}$ *and* $\tau>0$ *the semitrivial equilibrium* $(0,\tilde{v})$ *is unstable.*

4. Two similar competing species. In this section we consider the case τ positive but small; i.e., the two competing species are very similar. In subsection 4.1 we shall discuss the coexistence states and the dynamics of (1.8) for $\tau\ll 1$ and μ bounded away from zero. Theorem 1.2 will be proved in this subsection. In subsection 4.2 we shall establish Theorem 1.4, which covers the remaining case $\tau\ll 1$ and $\mu\ll 1$. Finally in subsection 4.3 we shall address some biological questions and prove Theorem 1.5. As can be seen later, Theorem 1.5 is supplementary to the results of subsection 4.1 as it gives more detailed information on the bifurcation diagram of coexistence states of (1.8).

4.1. Dynamics of (1.8) for $\tau\ll 1$ and μ bounded away from zero. The main purpose of this subsection is to prove Theorem 1.2. As will become clear later, Theorem 1.2 is a consequence of the following result.

THEOREM 4.1. *Assume* A1. *For any* $\tilde{\mu}>0$ *the following statements hold true:*

(i) *If* $G(\tilde{\mu})\neq 0$, *then there exists* $\epsilon>0$ *such that for* $\mu\in(\tilde{\mu}-\epsilon,\tilde{\mu}+\epsilon)$ *and* $\tau\in(0,\epsilon)$ *problem* (1.8) *has no coexistence states.*

(ii) *If* $G(\tilde{\mu})=0$ *and* $G'(\tilde{\mu})\neq 0$, *then for any sufficiently small* $\epsilon>0$, *there exists* $\tilde{\tau}=\tilde{\tau}(\epsilon)>0$ *with the following property. For every* $\tau\in(0,\tilde{\tau})$, *there exist* $\mu_*<\mu^*$ *with* $\mu_*,\mu^*\in(\tilde{\mu}-\epsilon,\tilde{\mu}+\epsilon)$ *such that for* $\mu\in[\tilde{\mu}-\epsilon,\tilde{\mu}+\epsilon]$, (1.8) *has a coexistence state if and only if* $\mu\in(\mu_*,\mu^*)$*; moreover, any coexistence state, if it exists, is the global attractor of* (1.8).

A crucial step in the proof of the theorem is a local bifurcation analysis. For $\tau\approx 0$, we look for triples (u,v,μ) that satisfy

$$\begin{aligned}
&\mu\Delta u+u(\beta(x)+\tau g(x)-u-v)=0, && x\in\Omega,\\
&\mu\Delta v+v(\beta(x)-u-v)=0, && x\in\Omega,\\
&\frac{\partial u}{\partial n}=\frac{\partial v}{\partial n}=0 \quad \text{on } \partial\Omega
\end{aligned} \tag{4.1}$$

and that are close to the curve $\Upsilon_{\tilde{\mu}}\times\{\tilde{\mu}\}$, where

$$\Upsilon_\mu:=\{(s\tilde{v}(\cdot,\mu),(1-s)\tilde{v}(\cdot,\mu)):s\in[0,1]\}.$$

Note that for any μ, $\Upsilon_\mu\times\{\mu\}$ is a curve of solutions of (4.1) for $\tau=0$. Also, for any small τ, (4.1) has the semitrivial solutions

$$\begin{aligned}
&(0,\tilde{v}(\cdot,\mu),\mu) \quad \text{(independent of } \tau),\\
&(\tilde{u}(\cdot,\mu,\tau),0,\mu),
\end{aligned} \tag{4.2}$$

where $\tilde{u}(\cdot,\mu,\tau)$ is the positive solution of (1.3) with $\alpha(x):=\beta(x)+\tau g(x)$.

For the functional analytic framework of the local analysis we introduce the following spaces:

$$Y = L^p(\Omega) \times L^p(\Omega),$$
$$X = \left\{ (y,z) \in W^{2,p}(\Omega) \times W^{2,p}(\Omega) : \frac{\partial y}{\partial n} = \frac{\partial z}{\partial n} = 0 \text{ on } \partial\Omega \right\},$$
$$X_2 = \left\{ (y,z) \in X : \int_\Omega (y(x) - z(x))\tilde{v}(x,\tilde{\mu})\, dx = 0 \right\},$$

where $p > N/2$ (so that $W^{2,p}(\Omega) \hookrightarrow C(\overline{\Omega})$).

PROPOSITION 4.2. *Let the hypotheses of Theorem* 4.1 *be satisfied. Then there exist a neighborhood U of the curve $\Upsilon_{\tilde{\mu}} \times \{\tilde{\mu}\}$ in $X \times (0,\infty)$ and $\delta > 0$ with the following properties:*

(i) *If $G(\tilde{\mu}) \neq 0$, then for $\tau \in (0,\delta)$ there are no solutions of* (4.1) *in U other than the semitrivial solutions* (4.2).

(ii) *If $G(\tilde{\mu}) = 0$ and $G'(\tilde{\mu}) \neq 0$, then for $\tau \in (0,\delta)$ the set of solutions of* (4.1) *in U consists of the semitrivial solutions and of the set $\Xi \cap U$, where Ξ is a smooth curve given by*

$$\Xi = \{(u(\tau,s), v(\tau,s), \mu(\tau,s)) : \ -\delta \le s \le 1+\delta\}. \tag{4.3}$$

Here $(\tau,s) \mapsto (u(\tau,s), v(\tau,s)) \in X$ and $(\tau,s) \mapsto \mu(\tau,s) \in (0,\infty)$ are smooth functions on $[0,\delta) \times (-\delta, 1+\delta)$ satisfying the following relations:

$$(u(\tau,0), v(\tau,0)) = (0, \tilde{v}(\cdot, \mu(\tau,0))), \tag{4.4}$$
$$(u(\tau,1), v(\tau,1)) = (\tilde{u}(\cdot, \mu(\tau,1), \tau), 0), \tag{4.5}$$
$$(u(0,s), v(0,s), \mu(0,s)) = (s\tilde{v}(\cdot,\tilde{\mu}), (1-s)\tilde{v}(\cdot,\tilde{\mu}), \tilde{\mu}). \tag{4.6}$$

In other words, a branch of coexistence states bifurcates from the branch of semitrivial equilibria $(\tilde{u}, 0)$ at $\mu = \mu(\tau,1)$ and meets the other branch of semitrivial equilibria $(0,\tilde{v})$ at $\mu = \mu(\tau,0)$. For $\tau = 0$ the branch coincides with with $\Upsilon_{\tilde{\mu}}$.

Note that from (4.6) it follows that the functions $u(\tau,s)$, $v(\tau,s)$, $\mu(\tau,s)$ have the following expansions for $-\delta \le s \le 1+\delta$ and $\tau \to 0$:

$$\begin{aligned} u(\tau,s) &= s\tilde{v}(\cdot,\tilde{\mu}) + \tau u_1(s) + O(\tau^2), \\ v(\tau,s) &= (1-s)\tilde{v}(\cdot,\tilde{\mu}) + \tau v_1(s) + O(\tau^2), \\ \mu(\tau,s) &= \tilde{\mu} + \tau\tilde{\mu}_1(s) + O(\tau^2), \end{aligned} \tag{4.7}$$

where $(u_1, v_1) \in X$ and $\tilde{\mu}_1 \in (0,\infty)$ are smooth functions of s. We will use these expansions below.

Proof of Proposition 4.2. For any triple (u,v,μ) near $\Upsilon_{\tilde{\mu}} \times \{\tilde{\mu}\}$, (u,v) can be written in a unique way as

$$(u,v) = (s\tilde{v}(\cdot,\mu), (1-s)\tilde{v}(\cdot,\mu)) + (y,z), \tag{4.8}$$

where $s \in \mathbb{R}$ and $(y,z) \in X_2$, and they are in or near $[0,1]$ and $\{(0,0)\}$, respectively. We shall thus look for solutions of (4.1) in this form. For that, an explicit expression of s and (y,z) in (4.8) will be useful. Writing (4.8) as

$$(u, v - \tilde{v}(\cdot,\mu)) = s(\tilde{v}(\cdot,\mu), -\tilde{v}(\cdot,\mu)) + (y,z),$$

we find s and (y,z) from

$$(y,z) = Q(\mu)(u, v-\tilde{v}(\cdot,\mu)),$$
$$s(\tilde{v}(\cdot,\mu), -\tilde{v}(\cdot,\mu)) = (I-Q(\mu))(u, v-\tilde{v}(\cdot,\mu)),$$

where I is the identity on X and $Q(\mu)$ is the projection of X onto X_2 along the subspace

$$X_1(\mu) := \operatorname{span}\{(\tilde{v}(\cdot,\mu), -\tilde{v}(\cdot,\mu))\}.$$

(Note that $X_1(\mu)$ is a complement of X_2 in X for $\mu \approx \tilde{\mu}$.) In particular, we find the following values of s and (y,z) for semitrivial equilibria:

$$(0,\tilde{v}(\cdot,\mu)) = (0,\tilde{v}(\cdot,\mu)) + (0,0) \quad (\text{i.e., } s=0,\ (y,z)=(0,0)), \tag{4.9a}$$
$$(\tilde{u}(\cdot,\mu,\tau),0) = (s\tilde{v}(\cdot,\mu),(1-s)\tilde{v}(\cdot,\mu)) + (\eta(\tau,\mu),\zeta(\tau,\mu)),\ \text{with } s=\sigma(\tau,\mu), \tag{4.9b}$$

where (η,ζ) and σ are smooth functions of (τ,μ) taking values in X_2 and $\mathbb{R}$, respectively. Clearly,

$$\sigma(0,\mu) = 1, \qquad (\eta(0,\mu),\zeta(0,\mu)) = (0,0), \tag{4.10}$$

as $\tilde{u}(\cdot,\mu,0) = v(\cdot,\mu)$.

For a small $\delta > 0$ let H be the map on

$$X \times (-\delta,\delta) \times (-\delta, 1+\delta) \times (\tilde{\mu}-\delta, \tilde{\mu}+\delta)$$

defined by

$$H(y,z,\tau,s,\mu) = \begin{bmatrix} \mu\Delta y - (y+z)s\tilde{v}(\cdot,\mu) + (\beta - \tilde{v}(\cdot,\mu))y - (y+z)y + \tau g s\tilde{v}(\cdot,\mu) + \tau g y \\ \mu\Delta z - (y+z)(1-s)\tilde{v}(\cdot,\mu) + (\beta-\tilde{v}(\cdot,\mu))z - (y+z)z \end{bmatrix}.$$

Note that, since $X \hookrightarrow C(\overline{\Omega}) \times C(\overline{\Omega}) \hookrightarrow Y$, H is well defined and smooth (in fact polynomial) as a Y-valued map. To find solutions of (4.1), we need to solve the equation

$$H(y,z,\tau,s,\mu) = 0, \tag{4.11}$$

with $(y,z) \in X_2$. It will be useful, however, to examine properties of $H(y,z,\tau,s,\mu)$ for $(y,z) \in X$. From the form of the solutions of (4.1) mentioned above (see (4.2) and the text preceding it) and by (4.9), we have

$$H(0,0,0,s,\mu) \equiv 0 \quad (s \in (-\delta, 1+\delta),\ \mu \in (\tilde{\mu}-\delta,\tilde{\mu}+\delta)), \tag{4.12a}$$
$$H(0,0,\tau,0,\mu) \equiv 0 \quad (\tau \in (-\delta,\delta),\ \mu \in (\tilde{\mu}-\delta,\tilde{\mu}+\delta)), \tag{4.12b}$$
$$H(\eta(\tau,\mu),\zeta(\tau,\mu),\tau,\sigma(\tau,\mu),\mu) \equiv 0 \quad (\tau \in (-\delta,\delta),\ \mu \in (\tilde{\mu}-\delta,\tilde{\mu}+\delta)). \tag{4.12c}$$

Define

$$L(s,\mu) := D_{(y,z)}H(0,0,0,s,\mu) \in \mathcal{L}(X,Y).$$

It is a standard consequence of the compactness of the embedding $W^{2,p}(\Omega) \hookrightarrow L^p(\Omega)$ that $L(s,\mu)$ is a Fredholm operator of zero index. By the definition of H, the vector $(\tilde{v}(\cdot,\mu), -\tilde{v}(\cdot,\mu))$ is in the kernel of $L(s,\mu)$. Put differently, $(\tilde{v}(\cdot,\mu), -\tilde{v}(\cdot,\mu))$ is an

eigenfunction corresponding to the eigenvalue 0 of $L(s,\mu)$, when $L(s,\mu)$ is viewed as an operator on Y with domain X. Since $\tilde{v}(\cdot,\mu) > 0$, zero must be a simple eigenvalue (cf. section 2) and we have

$$\ker L(s,\mu) = \operatorname{span}\{(\tilde{v}(\cdot,\mu), -\tilde{v}(\cdot,\mu))\} = X_1(\mu).$$

Let $P(s,\mu)$ be the continuous linear projection of Y onto $X_1(\mu)$ along the range of $L(s,\mu)$ (the range $R(L(s,\mu))$ is a closed subspace of Y of codimension one). We can write $P(s,\mu)$ explicitly as follows:

$$P(s,\mu)(y,z) = \frac{(1-s)\int_\Omega \tilde{v}(\cdot,\mu)y - s\int_\Omega \tilde{v}(\cdot,\mu)z}{\int_\Omega \tilde{v}^2(\cdot,\mu)} \left(\tilde{v}(\cdot,\mu), -\tilde{v}(\cdot,\mu)\right). \tag{4.13}$$

To verify this formula, one needs to show that

$$R(P(s,\mu)) = X_1(\mu), \quad (P(s,\mu))^2 = P(s,\mu), \quad \text{and } P(s,\mu)L(s,\mu) = 0.$$

The first property is obvious; the other two follow from a straightforward computation, which is left to the reader. Formula (4.13) in particular implies

$$Y_2(s,\mu) := R(L(s,\mu)) = \left\{(y,z) \in Y : (1-s)\int_\Omega \tilde{v}(\cdot,\mu)y - s\int_\Omega \tilde{v}(\cdot,\mu)z = 0\right\}.$$

Also note that $(s,\mu) \mapsto P(s,\mu)$ is smooth (in the operator norm). Following the Lyapunov–Schmidt scenario, we now consider the system

$$P(s,\mu)H(y,z,\tau,s,\mu) = 0, \tag{4.14a}$$

$$(I - P(s,\mu))H(y,z,\tau,s,\mu) = 0, \tag{4.14b}$$

where $(y,z) \in X_2$ and I is the identity on Y. If μ is sufficiently close to $\tilde{\mu}$ (and we make δ small enough for that to hold for all $\mu \in (\tilde{\mu}-\delta, \tilde{\mu}+\delta)$), then

$$\ker(L(s,\mu)) \cap X_2 = \{0\}.$$

It follows that $L(s,\mu)$ is an isomorphism of X_2 onto $Y_2(s,\mu)$. By the implicit function theorem, we can thus solve (4.14b) for (y,z), which leads to the following conclusion. There exist $\delta_1 > 0$, a neighborhood V of $(0,0) \in X_2$, and a smooth function

$$(\tau,s,\mu) \mapsto (y(\tau,s,\mu), z(\tau,s,\mu)) : (-\delta_1,\delta_1) \times (-\delta_1, 1+\delta_1) \times (\tilde{\mu}-\delta_1, \tilde{\mu}+\delta_1) \to X_2$$

such that $(y(0,s,\mu), z(0,s,\mu)) = (0,0)$ and $(y,z,\tau,s,\mu) \in V \times (-\delta_1,\delta_1) \times (-\delta_1, 1+\delta_1) \times (\tilde{\mu}-\delta_1, \tilde{\mu}+\delta_1)$ satisfies (4.11) if and only if $(y,z) = (y(\tau,s,\mu), z(\tau,s,\mu))$ and (τ,s,μ) solves the bifurcation equation

$$P(s,\mu)H(y(\tau,s,\mu), z(\tau,s,\mu), \tau, s, \mu) = 0.$$

By (4.12b), (4.12c), y and z necessarily satisfy

$$\begin{aligned} (y(\tau,0,\mu), z(\tau,0,\mu)) &= (0,0), \\ (y(\tau,\sigma(\tau,\mu),\mu), z(\tau,\sigma(\tau,\mu),\mu)) &= (\eta(\tau,\mu), \zeta(\tau,\mu)). \end{aligned} \tag{4.15}$$

Now, defining $\xi(\tau,s,\mu)$ by

$$\xi(\tau,s,\mu)\left(\tilde{v}(\cdot,\mu), -\tilde{v}(\cdot,\mu)\right) = P(s,\mu)H(y(\tau,s,\mu), z(\tau,s,\mu), \tau, s, \mu),$$

the bifurcation equation is equivalent to

$$\xi(\tau, s, \mu) = 0. \tag{4.16}$$

We immediately have the following solutions of (4.16):

$$\xi(0, s, \mu) \equiv \xi(\tau, 0, \mu) \equiv \xi(\tau, \sigma(\tau, \mu), \mu) \equiv 0. \tag{4.17}$$

These identities hold because for each of the indicated values of $(\tau, s, \mu) \in (-\delta_1, \delta_1) \times (-\delta_1, 1 + \delta_1) \times (\tilde{\mu} - \delta_1, \tilde{\mu} + \delta_1)$, there is a solution $(y, z) \in V$ of (4.11); see (4.12) (we make δ_1 smaller, if necessary, so that the solutions are indeed contained in V). Recall in passing that, from these solutions, the triples $(\tau, 0, \mu)$ and $(\tau, \sigma(\tau, \mu), \mu)$ correspond to semitrivial equilibria of (4.1); see (4.9), (4.15).

It follows from (4.17) that

$$\xi(\tau, s, \mu) \equiv s(\sigma(\tau, \mu) - s)\tau\xi_1(\tau, s, \mu)$$

for some smooth function $\xi_1(\tau, s, \mu)$. Solutions of (4.16) different from (4.17) are found by solving

$$\xi_1(\tau, s, \mu) = 0. \tag{4.18}$$

Observe that

$$\partial_\tau \xi(0, s, \mu) \equiv s(1 - s)\xi_1(0, s, \mu),$$

as $\sigma(0, \mu) \equiv 1$. The derivative on the left-hand side is computed from

$$\begin{aligned}\big(\tilde{v}(\cdot, \mu), -\tilde{v}(\cdot, \mu)\big)\partial_\tau \xi(0, s, \mu) &= \partial_\tau (P(s, \mu)H(y(\tau, s, \mu), z(\tau, s, \mu), \tau, s, \mu)) \,\big|_{\tau=0} \\ &= P(s, \mu)H_\tau(0, 0, 0, s, \mu) + P(s, \mu)L(s, \mu)(y_\tau, z_\tau) \\ &= P(s, \mu)H_\tau(0, 0, 0, s, \mu)\end{aligned}$$

(recall that $R(L(s, \mu)) = \ker P(s, \mu)$). Using (4.13), we find

$$P(s, \mu)H_\tau(0, 0, 0, s, \mu) = P(s, \mu)\big(sg\tilde{v}(\cdot, \mu), 0\big) = \frac{s(1 - s)\int_\Omega g\tilde{v}^2}{\int_\Omega \tilde{v}^2}\,\big(\tilde{v}(\cdot, \mu), -\tilde{v}(\cdot, \mu)\big).$$

Thus

$$\partial_\tau \xi(0, s, \mu) = \frac{s(1 - s)G(\mu)}{\int_\Omega \tilde{v}^2},$$

i.e.,

$$\xi_1(0, s, \mu) = \frac{G(\mu)}{\int_\Omega \tilde{v}^2}. \tag{4.19}$$

To complete the proof, consider first the case $G(\tilde{\mu}) \neq 0$. Making δ_1 yet smaller, if necessary, we infer from (4.19) that (4.18) has no solution in $(-\delta_1, \delta_1) \times (-\delta_1, 1 + \delta_1) \times (\tilde{\mu} - \delta_1, \tilde{\mu} + \delta_1)$. This implies statement (i) of Proposition 4.2.

Now assume $G(\tilde{\mu}) = 0$. Then

$$\partial_\mu \xi_1(0, s, \mu) \,\big|_{\mu=\tilde{\mu}} = \frac{G'(\tilde{\mu})}{\int_\Omega \tilde{v}^2(x, \tilde{\mu})}.$$

If $G'(\tilde{\mu}) \neq 0$, the implicit function theorem implies that for some $\delta_2 > 0$, all solutions of (4.18) in $(-\delta_2, \delta_2) \times (-\delta_2, 1+\delta_2) \times (\tilde{\mu}-\delta_2, \tilde{\mu}+\delta_2)$ are given by

$$\mu = m(\tau, s), \qquad \tau \in (-\delta_2, \delta_2), \quad s \in (-\delta_2, 1+\delta_2),$$

where $m(\tau, s)$ is a smooth function satisfying $m(0, s) \equiv \tilde{\mu}$. Thus, in addition to the immediate solutions (4.17), the bifurcation equation (4.16) has the family of solutions

$$\{(\tau, s, m(\tau, s)) : \tau \in (-\delta_2, \delta_2),\ s \in (-\delta_2, 1+\delta_2)\}. \tag{4.20}$$

In this family, the point $(\tau, 0, m(\tau, 0))$ is also contained in the set of solutions found in (4.17), and it corresponds to the semitrivial solution $(0, \tilde{v}(\cdot, \mu))$ of (4.1) with $\mu = m(\tau, 0)$ (see the remarks following (4.17)). Next we look for points $(\tau, s, m(\tau, s))$ in the family corresponding to the semitrivial equilibria $(\tilde{u}(\cdot, \mu, \tau), 0)$. Referring to the remarks following (4.17) again, we see that s and τ are found from the equation

$$s = \sigma(\tau, m(\tau, s)). \tag{4.21}$$

Since $\sigma(0, \mu) \equiv 1$, for $\tau \approx 0$ there is a unique solution $s = \bar{s}(\tau)$ of (4.21), and it depends smoothly on τ. Hence for each fixed $\tau \approx 0$, $(\tau, \bar{s}(\tau), m(\tau, \bar{s}(\tau)))$ is a point contained in the family (4.20) which corresponds to the semitrivial solution $(\tilde{u}(\cdot, \mu, \tau), 0)$ of (4.1) with $\mu = m(\tau, \bar{s}(\tau))$.

Using the scaled variable $\tilde{s} = s\bar{s}(\tau)$, we now define

$$\begin{aligned} u(\tau, s) &= \tilde{s}\tilde{v}(\cdot, m(\tau, \tilde{s})) + y(\tau, \tilde{s}, m(\tau, \tilde{s})), \\ v(\tau, s) &= (1-\tilde{s})\tilde{v}(\cdot, m(\tau, \tilde{s})) + z(\tau, \tilde{s}, m(\tau, \tilde{s})), \\ \mu(\tau, s) &= m(\tau, \tilde{s}) \quad (\tilde{s} = s\bar{s}(\tau)). \end{aligned}$$

Clearly, these are smooth functions of $(\tau, s) \in (-\delta, \delta) \times (-\delta, 1+\delta)$ if δ is sufficiently small, and $(u(\tau, s), v(\tau, s))$ is a solution of (4.1) for $\mu = m(\tau, s)$. By construction, these solutions, together with the semitrivial equilibria, contain all solutions of (4.1) in a small neighborhood of $\Upsilon_{\tilde{\mu}} \times \{\tilde{\mu}\}$ for $\tau \in (-\delta, \delta)$. The relations $(y(0, s, \mu), z(0, s, \mu)) = (0, 0)$, $m(0, s) = \tilde{\mu}$, and $\bar{s}(0) = 1$ imply (4.6). The correspondences between the solutions $(\tau, \bar{s}(\tau), m(\tau, \bar{s}(\tau)))$, $(\tau, 0, \mu(0, \tau))$ of (4.16) and the semitrivial equilibria, as discussed above, imply (4.4), (4.5). This completes the proof. □

The next lemma shows that by the local analysis we have found all coexistence states.

LEMMA 4.3. *Let the hypotheses of Theorem* 4.1 *be satisfied. Then given any neighborhood U of the curve $\Upsilon_{\tilde{\mu}} \times \{\tilde{\mu}\}$ in $X \times (0, \infty)$ there exists $\delta > 0$ such that for $\tau \in (-\delta, \delta)$ all solutions (u, v, μ) of* (4.1) *with $u, v \geq 0$ and $|\mu - \tilde{\mu}| \leq \delta$ are contained in U.*

Proof. The conclusion follows once we prove the following claim. If $\tau_j \to 0+$, $\mu_j \to \tilde{\mu}$, and (u_j, v_j, μ_j) is a sequence of solutions of (4.1) with $\tau = \tau_j$, $\mu = \mu_j$ such that $u_j, v_j \geq 0$, then (u_j, v_j) approaches the curve $\Upsilon_{\tilde{\mu}}$.

By the maximum principle, we have the following a priori bound on the nonnegative solutions:

$$u_j \leq \sup(\beta(x) + \tau g(x)), \quad v_j \leq \sup \beta(x).$$

Since $\tau_j \to 0+$, $\mu_j \to \tilde{\mu}$, by standard elliptic estimates (see [GT]), passing to a

subsequence we may assume that $(u_j, v_j) \to (\hat{u}, \hat{v})$ in X, where $\hat{u},\ \hat{v} \geq 0$ in $\bar{\Omega}$, and

$$\tilde{\mu}\Delta\hat{u} + \hat{u}(\beta - \hat{u} - \hat{v}) = 0 \quad \text{in} \quad \Omega, \tag{4.22a}$$

$$\tilde{\mu}\Delta\hat{v} + \hat{v}(\beta - \hat{u} - \hat{v}) = 0 \quad \text{in} \quad \Omega, \tag{4.22b}$$

$$\frac{\partial\hat{u}}{\partial n} = \frac{\partial\hat{v}}{\partial n} = 0 \quad \text{on} \quad \partial\Omega. \tag{4.22c}$$

Since any solution $(\hat{u}, \hat{v})$ of (4.22) with $\hat{u},\ \hat{v} \geq 0$ is contained on the curve $\Upsilon_{\tilde{\mu}}$, the claim follows. □

In the remainder of the subsection, we simplify the notation by writing

$$\tilde{v} = \tilde{v}(x, \tilde{\mu}).$$

For other values of μ we keep the notation $\tilde{v}(x, \mu)$.

The next crucial step is the stability of coexistence states on the curve Ξ: Let

$$(u, v, \mu) = (u(\tau, s), v(\tau, s), \mu(\tau, s)) \in \Xi,$$

and consider the corresponding linear eigenvalue problem (2.3). When $\tau = 0$, we have $(u, v) = (s\tilde{v}, (1 - s)\tilde{v})$ and (2.3) has an eigenvalue $\lambda = 0$, the corresponding eigenfunction being $(\tilde{v}, -\tilde{v})$. Since $\tilde{v} > 0$, $\lambda = 0$ is the principal eigenvalue (see section 2); in particular, it is (algebraically) simple and all other eigenvalues are positive. By standard spectral perturbation theory [K2], for $|\tau| \ll 1$, (2.3) has a unique eigenvalue, denoted by $\lambda(\tau, s)$ such that $\lim_{\tau\to 0} \lambda(\tau, s) = 0$, and all the other eigenvalues of (2.3) are positive and uniformly bounded away from zero for any $s \in [0, 1]$ and small positive τ. Hence the sign of $\lambda(\tau, s)$ determines the stability of coexistence states on Ξ. Note that

$$\lambda(\tau, 0) = \lambda(\tau, 1) = 0 \tag{4.23}$$

for $\mu(\tau, 0)$ and $\mu(\tau, 1)$ are bifurcation points (points of intersections of Ξ with the branches of semitrivial solutions). It is not hard to check that $\lambda_\tau(0, s) = 0$ (see Appendix A) and Proposition 4.4 below gives a formula for $\lambda_{\tau\tau}(0, s)$; the proof is a straightforward but tedious computational exercise and is given in Appendix A. For the formulation we introduce some notation. Let $H = L^2(\Omega)$ be the usual Hilbert space with norm $\|\cdot\|$ and inner product $(\cdot, \cdot)$. Take the linear subspace spanned by $\tilde{v}$ to be Θ and let $\Theta^\perp$ be its orthogonal complement. Denote the domain and kernel of a linear operator $\mathcal{L}$ by $\mathrm{dom}(\mathcal{L})$ and $\ker(\mathcal{L})$, respectively, so $\mathcal{L} : \mathrm{dom}(\mathcal{L}) \to H$. Consider the formally self-adjoint operator $\tilde{\mu}\Delta + \beta - \tilde{v}$ and define in the standard manner the self-adjoint operator on H corresponding to zero Neumann boundary conditions. We thus have the self-adjoint operators

$$\mathcal{L} = \tilde{\mu}\Delta + \beta - \tilde{v},$$
$$\mathcal{L} - \tilde{v} = \tilde{\mu}\Delta + \beta - 2\tilde{v}.$$

Since the principal eigenvalue of $\mathcal{L}$ is 0 (the eigenfunction is $\tilde{v}$), it is straightforward to show that $\mathcal{L} - \tilde{v}$ has the bounded inverse $(\mathcal{L} - \tilde{v})^{-1}$ on H, and we define $\mathcal{L}^{-1}$ on $\Theta^\perp$ by setting $\mathcal{L}^{-1}\phi = \psi$ if and only if $\mathcal{L}\psi = \phi$ and $\phi, \psi \in \Theta^\perp$.

PROPOSITION 4.4. *For $0 \le s \le 1$ and $0 < \tau \ll 1$, the following statements hold:*
(i)

$$\lambda(\tau, s) = s(1-s)\tau^2 \left\{ \frac{2}{\int_\Omega \tilde{v}^2} \int_\Omega g\tilde{v} \left[(\mathcal{L} - \tilde{v})^{-1} - \mathcal{L}^{-1} \right] g\tilde{v} + C_1(\tau, s)\tau \right\},$$

where $C_1(\tau, s)$ is some constant uniformly bounded for $s \in [0,1]$ and $\tau \ll 1$.
(ii) *With $\tilde{\mu}_1$ as in* (4.7), *one has*

$$\tilde{\mu}_1(s) = \frac{1}{G'(\tilde{\mu})} \int_\Omega g\tilde{v} \left[2s(\mathcal{L} - \tilde{v})^{-1}(g\tilde{v}) + (1-2s)\mathcal{L}^{-1}(g\tilde{v}) \right].$$

Note that, by our assumption, $G(\tilde{\mu}) = \int g\tilde{v}^2 = 0$, so that $\mathcal{L}^{-1}(g\tilde{v})$ is well defined.

LEMMA 4.5. *The following holds for any nontrivial $\varphi \in \Theta^\perp$:*

$$\int_\Omega \varphi \left[(\mathcal{L} - \tilde{v})^{-1} - \mathcal{L}^{-1} \right] \varphi > 0. \tag{4.24}$$

Proof. For $\tau > 0$ let

$$h(\tau) = \int_\Omega \varphi (\mathcal{L} - \tau\tilde{v})^{-1} \varphi. \tag{4.25}$$

Since $\mathcal{L} - \tau\tilde{v}$ is invertible for $\tau > 0$, h is well defined. We claim that h is strictly increasing. To prove this set $\Phi = (\mathcal{L} - \tau\tilde{v})^{-1}\varphi$. It is easy to check that

$$\frac{\partial \Phi}{\partial \tau} = (\mathcal{L} - \tau\tilde{v})^{-1}(\tilde{v}\Phi). \tag{4.26}$$

Then

$$\begin{aligned} \frac{dh}{d\tau} &= \int_\Omega \varphi \frac{\partial \Phi}{\partial \tau} = \int_\Omega \varphi (\mathcal{L} - \tau\tilde{v})^{-1}(\tilde{v}\Phi) \\ &= \int_\Omega (\mathcal{L} - \tau\tilde{v})^{-1}\varphi \cdot \tilde{v}\Phi \qquad \text{(self-adjointness of } (\mathcal{L} - \tau\tilde{v})^{-1}) \\ &= \int_\Omega \tilde{v}\Phi^2 > 0. \end{aligned} \tag{4.27}$$

The last inequality in (4.27) is strict since $\Phi \not\equiv 0$. In particular, we have $h(1) > \lim_{\tau \to 0+} h(\tau)$. In the following we show that

$$\lim_{\tau \to 0+} h(\tau) = \int_\Omega \varphi \mathcal{L}^{-1} \varphi \tag{4.28}$$

for every $\varphi \in \Theta^\perp$, from which (4.24) follows.

To prove (4.28), let $Mu := \tilde{v}u$ for $u \in H$. We always assume that $\tau > 0$ is small and that C_i are strictly positive constants independent of τ. Note that $\mathcal{L}\,\mathrm{dom}(\mathcal{L}) = \Theta^\perp$, $\mathcal{L}^{-1}\Theta^\perp \sqsubset \Theta^\perp$, and

$$\|\mathcal{L}^{-1}\| \le C_1, \tag{4.29a}$$

$$(-\mathcal{L}\phi, \phi) \ge C_2\|\phi\|^2, \qquad \phi \in \Theta^\perp \cap \mathrm{dom}(\mathcal{L}), \tag{4.29b}$$

$$\|M\| \le C_3, \tag{4.29c}$$

$$(Mu, u) \ge C_4\|u\|^2, \qquad u \in H. \tag{4.29d}$$

Consider the equation

$$(\mathcal{L}-\tau M)u=\varphi,\qquad \varphi\in\Theta^{\perp}. \tag{4.30}$$

We claim that for the solution u of (4.30) (that is, $u=(\mathcal{L}-\tau M)^{-1}\varphi$) we have

$$\|u\|\le C_5\|\varphi\|. \tag{4.31}$$

To prove this assertion, put $u=a\tilde{v}+\phi$, where $a\in\mathbb{R}$, $\phi\in\Theta^{\perp}$. Substituting in (4.30), we have

$$\mathcal{L}\phi-\tau aM\tilde{v}-\tau M\phi=\varphi. \tag{4.32}$$

Take the inner product with $\tilde{v}$ and use (4.29c) and (4.29d):

$$|a|=|(M\phi,\tilde{v})/(M\tilde{v},\tilde{v})|\le C_6\|\phi\|. \tag{4.33}$$

Take the inner product of (4.32) with $-\phi$ and use (4.29b), (4.29c), and (4.33), obtaining

$$\begin{aligned} C_2\|\phi\|^2 &\le(-\mathcal{L}\phi,\phi)=-(\varphi,\phi)-\tau a(M\tilde{v},\phi)-\tau(M\phi,\phi)\\ &\le\|\varphi\|\|\phi\|+\tau(C_6+1)C_3\|\phi\|^2, \end{aligned} \tag{4.34}$$

which implies that

$$\|\phi\|\le C_7\|\varphi\| \tag{4.35}$$

if τ is small enough. Estimates (4.33) and (4.35) prove claim (4.31).

For u given by (4.30) set

$$w=\frac{1}{\tau}\left(u-\mathcal{L}^{-1}\varphi+\frac{(M\mathcal{L}^{-1}\varphi,\tilde{v})}{(M\tilde{v},\tilde{v})}\,\tilde{v}\right). \tag{4.36}$$

It is easy to check that

$$(\mathcal{L}-\tau M)w=\varphi_1, \tag{4.37}$$

where

$$\varphi_1=M\mathcal{L}^{-1}\varphi-\frac{(M\mathcal{L}^{-1}\varphi,\tilde{v})}{(M\tilde{v},\tilde{v})}\,M\tilde{v}\in\Theta^{\perp}. \tag{4.38}$$

Obviously $\|\varphi_1\|\le C_8\|\varphi\|$, so using (4.31) on (4.37) with φ_1, w replacing φ, u, respectively, we find that $\|w\|\le C_5\|\varphi_1\|\le C_9\|\varphi\|$.

Finally, for any $\varphi\in\Theta^{\perp}$, from (4.36)

$$h(\tau)=(\varphi,u)=(\varphi,\mathcal{L}^{-1}\varphi)+\tau(\varphi,w). \tag{4.39}$$

Since $\|w\|$ is uniformly bounded, we deduce that, as required, $\lim_{\tau\to0+}h(\tau)=(\varphi,\mathcal{L}^{-1}\varphi)$. This proves Lemma 4.5. □

Proof of Theorem 4.1. By Proposition 4.2 and Lemma 4.3, for $\mu\approx\tilde{\mu}$ and $\tau\approx0$, all coexistence states lie on the branch Ξ. On the branch we have

$$\mu(\tau,s)=\tilde{\mu}+\tau\tilde{\mu}_1(s)+\tau^2H(\tau,s)$$

for some smooth function H. By Proposition 4.4 and Lemma 4.5, $\tilde{\mu}_1$ is a nonconstant affine function, and thus for small τ, $\mu(\tau,\cdot)$ is strictly monotone. It follows that the first statement of the theorem holds with

$$\mu_* = \min_{s\in[0,1]} \mu(\tau,s), \qquad \mu_* = \max_{s\in[0,1]} \mu(\tau,s)$$

and that the coexistence state is unique for each fixed $\mu \in (\mu_*,\mu^*)$. By Proposition 4.4 and Lemma 4.5, we also have $\lambda(\tau,s) > 0$ for small τ, and thus the coexistence state is stable. As noted in section 2, the uniqueness and stability of the coexistence state implies that it is the global attractor. The theorem is proved. □

We end this subsection by applying Theorem 4.1 to the proof of Theorem 1.2.

Proof of Theorem 1.2. In view of Theorem 4.1 and Lemma 4.3, by the standard compactness argument, the dynamics and the structures of the coexistence state are clear for values of μ in any compact subset of $(0,\infty)$. To complete the proof of Theorem 1.2, it suffices to show that under the stronger assumption A2$^+$, $(\tilde{u},0)$ is the global attractor of (1.8) for $\mu > \mu_k^*(\tau)$. By Corollary 3.2 and Theorem 4.1, we see that $(0,\tilde{v})$ is unstable for $\mu > \mu_{k,*}(\tau)$. Therefore in view of property (a) of system (1.8) (see section 2), it suffices to show that (1.8) has no coexistence state for $\mu \geq \mu_k^*(\tau)$. By Lemma 4.3 and Theorem 4.1, it suffices to prove this for sufficiently large μ and positive bounded τ. To this end, we argue by contradiction. Suppose that there exist sequences τ_j uniformly bounded and $\mu_j \to \infty$ such that (1.8) has coexistence states $\{(u_j,v_j)\}_{j=1}^{\infty}$ with $(\tau,\mu) = (\tau_j,\mu_j)$. Set $\hat{u}_j = u_j/\|u_j\|_\infty$, $\hat{v}_j = v_j/\|v_j\|_\infty$. It is easy to check that $\hat{u}_j, \hat{v}_j$ satisfy

$$\mu_j \Delta \hat{u}_j + \hat{u}_j(\beta + \tau_j g - u_j - v_j) = 0 \qquad \text{in} \quad \Omega, \tag{4.40a}$$

$$\mu_j \Delta \hat{v}_j + \hat{v}_j(\beta - u_j - v_j) = 0 \qquad \text{in} \quad \Omega, \tag{4.40b}$$

$$\frac{\partial \hat{u}_j}{\partial n} = \frac{\partial \hat{v}_j}{\partial n} = 0 \qquad \text{on} \quad \partial\Omega. \tag{4.40c}$$

Since τ_j is uniformly bounded, by standard elliptic estimates (see [GT]) and passing to a subsequence we may assume that $(\hat{u}_j,\hat{v}_j) \to (\hat{u},\hat{v})$ in $C^2(\overline{\Omega})$. Since $\mu_j \to \infty$ and $\|\hat{u}_j\|_\infty = \|\hat{v}_j\|_\infty = 1$, it is easy to see that $\hat{u} \equiv 1$ and $\hat{v} \equiv 1$. That is, $(\hat{u}_j,\hat{v}_j) \to (1,1)$ uniformly. By Lemma A.1, $\int_\Omega g u_j v_j = 0$ for all j, i.e., $\int_\Omega g \hat{u}_j \hat{v}_j = 0$. Passing to the limit we get $\int_\Omega g = 0$, which contradicts our assumption $\int_\Omega g > 0$. This completes the proof of Theorem 1.2. □

4.2. Dynamics of (1.8) for $\tau \ll 1$ and $\mu < \mu_{1,*}$. The goal of this subsection is to establish Theorem 1.4. We shall consider the case $\int_\Omega g\beta^2 < 0$ only since the case $\int_\Omega g\beta^2 > 0$ is very similar. In this subsection $\tilde{v}$ stands for $\tilde{v}(\cdot,\mu)$, and $\tilde{v}_\mu$ stands for $\frac{\partial \tilde{v}}{\partial \mu}(\cdot,\mu)$.

Proof of Theorem 1.4. We first consider the stability of $(0,\tilde{v})$. Recall that $C(\mu)$ is defined as in (3.2), $C(\mu) > 0$ if $G(\mu) < 0$, and $C(\mu) = 0$ if $G(\mu) \geq 0$. Since $C(\mu)$ is the principal eigenvalue, it is a smooth function of μ. By Lemma 3.1, it suffices to solve the equation $\tau = \mu C(\mu)$ for τ, μ small.

We claim that the following hold:

$$\lim_{\mu\to 0+} C(\mu) = C^* := \inf_{\left\{\psi\in H^1 : \int_\Omega g\beta^2\psi^2 > 0\right\}} \frac{\int_\Omega \beta^2 |\nabla\psi|^2}{\int_\Omega g\beta^2\psi^2}, \tag{4.41a}$$

$$\lim_{\mu\to 0+} \mu C'(\mu) = 0. \tag{4.41b}$$

Note that $C^* > 0$ because $\int_\Omega g\beta^2 < 0$. Assuming (4.41), we prove (i) of Theorem 1.4 (the proof of (ii) is quite analogous). Set $F(\tau,\mu) = \tau - \mu C(\mu)$. Since $\frac{\partial F}{\partial \mu} \to -C^*$ as $\mu \to 0+$, by the implicit function theorem we see that there exists $\eta_1, \eta_2 > 0$ such that for every $0 < \tau < \eta_1$, there exists a unique $\mu_0 = \mu_0(\tau) \in (0, \eta_2)$ such that $\tau > \mu C(\mu)$ when $0 < \mu < \mu_0$, $\tau = \mu_0 C(\mu_0)$, and $\tau < \mu C(\mu)$ when $\mu_0 < \mu \le \eta_2$. By Lemma 3.1, this proves that $(0, \tilde{v})$ is unstable for $\mu < \mu_0$ and stable for $\mu_0 < \mu \le \eta_2$. Choose $\eta = \eta_2$ in Theorem 1.2. From the proof of Theorem 1.2 we see that $(0, \tilde{v})$ does not changes its stability for $\eta_2 \le \mu < \mu_{1,*}$ and $\tau < \tau_0(\eta_2)$. Hence for $\tau < \min\{\eta_1, \tau_0(\eta_2)\}$, $(0, \tilde{v})$ is unstable for $\mu < \mu_0$ and stable for $\mu_0 < \mu < \mu_{1,*}$.

To prove (1.12), observe that as $\lim_{\tau\to 0+} \mu_0(\tau) = 0$, we have by (4.41a)

$$\frac{\tau}{\mu_0(\tau)} = C(\mu_0(\tau)) \to C^*. \tag{4.42}$$

It remains to prove the instability of $(\tilde{u}, 0)$ for suitable μ when $\int_\Omega g\beta^2 < 0$. It suffices to show that the principal eigenvalue, denoted by λ_1, of the problem

$$\mu\Delta\varphi + (\beta - \tilde{u})\varphi = -\lambda\varphi \quad \text{in } \Omega, \qquad \frac{\partial\varphi}{\partial n}\Big|_{\partial\Omega} = 0 \tag{4.43}$$

is negative. Observe that $\varphi/\tilde{u}$ satisfies

$$\nabla\cdot\Big(\tilde{u}^2\nabla\frac{\varphi}{\tilde{u}}\Big) - \frac{\tau}{\mu}\, g\tilde{u}^2\frac{\varphi}{\tilde{u}} = -\frac{\lambda}{\mu}\,\tilde{u}^2\,\frac{\varphi}{\tilde{u}} \qquad \text{in } \ \Omega, \qquad \frac{\partial}{\partial n}\Big(\frac{\varphi}{\tilde{u}}\Big)\Big|_{\partial\Omega} = 0. \tag{4.44}$$

Hence λ_1 can be characterized as

$$\lambda_1 = \inf_{\{\psi\in H^1:\psi\neq 0\}} \frac{\mu\int_\Omega \tilde{u}^2|\nabla\psi|^2 + \tau\int_\Omega g\tilde{u}^2\psi^2}{\int_\Omega \tilde{u}^2\psi^2}\,. \tag{4.45}$$

By letting $\psi \equiv 1$ in (4.45), we have $\lambda_1 \le \tau\int_\Omega g\tilde{u}^2/\int_\Omega \tilde{u}^2$. Note that $\int_\Omega g\tilde{u}^2 \to \int_\Omega g(\beta+\tau g)_+^2$ as $\lim_{\mu\to 0+}\tilde{u} = (\beta+\tau g)_+$. For sufficiently small τ, $(\beta+\tau g)_+ = \beta+\tau g$ since $\beta > 0$, and $\int_\Omega g(\beta+\tau g)_+^2 < 0$ since $\int_\Omega g\beta^2 < 0$. Therefore there exist $\eta_3, \eta_4 > 0$ such that $\lambda_1 < 0$ for $0 < \mu \le \eta_3$ and $\tau \le \eta_4$. This proves the instability of $(\tilde{u}, 0)$ for $\mu \le \eta_3$; for $\mu \ge \eta_3$, choose $\eta = \eta_3$ in Theorem 1.2. From the proof of Theorem 1.2 we see that $(\tilde{u}, 0)$ does not change its stability for $\mu \in [\eta_3, \mu_1^*)$ if $\tau < \tau_0(\eta_3)$. This shows that $(\tilde{u}, 0)$ is unstable for $\mu \in (\mu_0, \mu_1^*)$ provided that $\tau < \min\{\eta_4, \tau_0(\eta_3)\}$. Thus, the theorem follows from (4.41).

Hence it suffices to establish (4.41). Since $\beta > 0$ in $\bar{\Omega}$, $\tilde{v} \to \beta$ uniformly when $\mu \to 0+$ (cf. (2.2a)), and thus (4.41a) follows by a standard argument.

For (4.41b), by the definition of $C(\mu)$, there exists $\Phi > 0$ such that

$$\int_\Omega \tilde{v}^2\nabla\Phi\cdot\nabla\Psi = C(\mu)\int_\Omega g\tilde{v}^2\Phi\Psi \tag{4.46}$$

for every $\Psi \in H^1(\Omega)$. We can normalize Φ so that $\|\Phi\|_{L^2(\Omega)} = 1$. Differentiate (4.46) with respect to μ, obtaining

$$2\int_\Omega \tilde{v}\tilde{v}_\mu\nabla\Phi\cdot\nabla\Psi + \int_\Omega \tilde{v}^2\nabla\Phi_\mu\nabla\Psi = C'\int_\Omega g\tilde{v}^2\Phi\Psi + 2C\int_\Omega g\tilde{v}\tilde{v}_\mu\Phi\Psi + C\int_\Omega g\tilde{v}^2\Phi_\mu\Psi \tag{4.47}$$

for every $\Psi \in H^1(\Omega)$, where $C' = dC/d\mu$. Letting $\Psi = \Phi$ in (4.47), we have

$$C' \int_\Omega g\tilde{v}^2\Phi^2 = 2\int_\Omega \tilde{v}\tilde{v}_\mu |\nabla\Phi|^2 - 2C\int_\Omega g\tilde{v}\tilde{v}_\mu\Phi^2, \tag{4.48}$$

where (4.46) has been used again.

We first observe that $\|\Phi\|_{H^1}$ is uniformly bounded for small μ: to see this, set $\Psi = \Phi$ in (4.46). As $\tilde{v} \to \beta > 0$ uniformly and $\|\Phi\|_{L^2} = 1$, we get $\|\nabla\Phi\|_{L^2} \le K_1$, where K_1 is independent of μ when $\mu \ll 1$. This proves that $\|\Phi\|_{H^1}$ is uniformly bounded.

Next we show that $\int_\Omega g\tilde{v}^2\Phi^2 \ge K_2 > 0$ for some K_2 independent of μ. We argue by contradiction: if not, suppose that $\int_\Omega g\tilde{v}^2\Phi^2 \to 0$ as $\mu \to 0$. By the Sobolev embedding theorem (see [A]), we may assume, passing to a sequence if necessary, that $\Phi \to \Phi_0$ weakly in H^1 and strongly in L^2. This implies that $\Phi_0 \ge 0$ a.e. in Ω, $\|\Phi_0\|_{L^2(\Omega)} = 1$, and by (4.46) $\int_\Omega \beta^2\nabla\Phi_0 \cdot \nabla\Psi = 0$ for every $\Psi \in H^1$. Therefore the only possibility is $\Phi_0 \equiv 1$. However, this is impossible since $\int_\Omega g\tilde{v}^2\Phi^2 \to \int_\Omega g\beta^2 < 0$, which contradicts $\int_\Omega g\tilde{v}^2\Phi^2 \to 0$ as $\mu \to 0$. Therefore

$$\|\Phi\|_{H^1} \le K_3, \qquad \int_\Omega g\tilde{v}^2\Phi^2 \ge K_2 > 0, \tag{4.49}$$

provided that $\mu \ll 1$.

By (4.48) and (4.49), to prove (4.41b) we need to show only that $\mu\tilde{v}_\mu \to 0$ uniformly as $\mu \to 0+$. To this end, differentiating (1.3) with respect to μ we have

$$\mu\Delta\tilde{v}_\mu + (\beta - 2\tilde{v})\tilde{v}_\mu + \Delta\tilde{v} = 0 \quad \text{in } \Omega, \qquad \frac{\partial\tilde{v}_\mu}{\partial n} = 0 \quad \text{on } \partial\Omega. \tag{4.50}$$

Choose $x_\mu \in \overline{\Omega}$ such that $\tilde{v}_\mu(x_\mu) = \max_{\overline{\Omega}} \tilde{v}_\mu$. We have

$$(\beta - 2\tilde{v})\tilde{v}_\mu + \Delta\tilde{v} = -\mu\Delta\tilde{v}_\mu \ge 0$$

at $x = x_\mu$. This is obvious if $x_\mu \in \Omega$; for $x_\mu \in \partial\Omega$ it follows from the boundary condition (cf. Proposition 2.2 of [LN]).

Note that $\beta - 2\tilde{v} \to -\beta < 0$ uniformly as $\mu \to 0$. Therefore for $0 < \mu \ll 1$, we have $\tilde{v}_\mu(x_\mu) \le K_4\|\Delta\tilde{v}\|_\infty$ for some positive constant K_4 which is independent of μ. That is, $\max_{\overline{\Omega}} \tilde{v}_\mu \le K_4\|\Delta\tilde{v}\|_\infty$. Similarly we can show that $\min_{\overline{\Omega}} \tilde{v}_\mu \ge -K_5\|\Delta\tilde{v}\|_\infty$. Hence we have $\|\tilde{v}_\mu\|_{L^\infty} \le K_6\|\Delta\tilde{v}\|_{L^\infty}$. Therefore

$$\mu\|\tilde{v}_\mu\|_{L^\infty} \le K_6\mu\|\Delta\tilde{v}\|_{L^\infty} = K_6\|\tilde{v}(\beta - \tilde{v})\|_{L^\infty} \le K_7\|\beta - \tilde{v}\|_{L^\infty} \to 0 \tag{4.51}$$

as $\mu \to 0+$. This implies that (4.41b) holds. □

Remark 4.6. We conjecture that for τ small, if there is a coexistence state of (1.8), it is always unique and is the global attractor of (1.8). The following weaker uniqueness result may be proved, but we omit the proof here: suppose that the assumptions in Theorems 1.2 and 1.4 hold. Then for any $\eta > 0$, there exists $\hat{\tau}(\eta) > 0$ such that if $0 < \tau \le \hat{\tau}$ and $\mu \ge \eta\tau$, (1.8) has at most one coexistence state; moreover, if a coexistence state exists, it must be the global attractor of (1.8).

4.3. Coexistence or fixation? We now turn to the question raised in the introduction as to whether there are mutants (which mathematically are represented by functions g) which invade but always (that is, for all small $\tau > 0$) yield coexistence. Recall from the remarks in the introduction that necessarily $G(\mu) = \int_\Omega g(x)\tilde{v}^2(x,\mu) =$

0, so the question is to discover whether there are functions g satisfying this requirement which lead to coexistence for all small τ.

To answer this question, first recall that near any μ with $G(\mu) = 0$ and $G'(\mu) \neq 0$, we have shown that there is a branch of positive coexistence states which connects both branches of semitrivial coexistence states $(\tilde{u}, 0)$ and $(0, \tilde{v})$. Moreover, for each μ, there exists at most one coexistence state; if it exists, it is stable and the global attractor of (1.8). Therefore the question is basically about the location of the two ends of this bifurcation branch. From Proposition 4.2, we know that this branch of coexistence states can be represented by a smooth curve as

(4.52)
$$(u, v, \mu) = \left(s\tilde{v} + \tau u_1(\cdot, s) + O(\tau^2), (1-s)\tilde{v} + \tau v_1(\cdot, s) + O(\tau^2), \mu + \tau\tilde{\mu}_1(s) + O(\tau^2)\right),$$

where $s \in (0, 1)$. By Proposition 4.4(ii), $\tilde{\mu}_1(s)$ is given as

$$\tilde{\mu}_1(s) = \frac{1}{G'(\mu)} \int_\Omega g\tilde{v} \left[2s(\mathcal{L} - \tilde{v})^{-1}(g\tilde{v}) + (1-2s)\mathcal{L}^{-1}(g\tilde{v})\right]. \tag{4.53}$$

It is now clear that the answer to the above question depends on whether $\tilde{\mu}_1(0)$ and $\tilde{\mu}_1(1)$ have opposite signs under the assumption $\int_\Omega g\tilde{v}^2 = 0$. It is easy to see that $\operatorname{sign}(\tilde{\mu}_1(0)) = \operatorname{sign}(-G'(\mu))$ since by (4.53),

$$\tilde{\mu}_1(0)G'(\mu) = \int_\Omega g\tilde{v}\mathcal{L}^{-1}(g\tilde{v}) \leq 0. \tag{4.54}$$

The inequality in (4.54) is in fact strict since $g\tilde{v} \in \Theta^\perp$ and the self-adjoint operator $\mathcal{L}^{-1}$ from $\Theta^\perp$ to $\Theta^\perp$ is negative.

Therefore it remains to find the sign of $\tilde{\mu}_1(1)$, i.e., the sign of

$$I(\mu, g) := \int_\Omega g\tilde{v} \left[2(\mathcal{L} - \tilde{v})^{-1} - \mathcal{L}^{-1}\right](g\tilde{v}). \tag{4.55}$$

Note that $I(\mu, -g) = I(\mu, g)$, thus the sign of $\int g$ is irrelevant in our computations. It is easy to see that Theorem 1.5 follows from the next result.

THEOREM 4.7. *There exists a (unique) $\mu^* > 0$, depending only on β and Ω, with the following properties:*

(i) *If $\mu > \mu^*$, then $I(\mu, g) < 0$ for any g satisfying $\int_\Omega g\tilde{v}^2(\cdot, \mu) = 0$.*

(ii) *If $\mu < \mu^*$, then there exists a nonempty open subset U of*

$$\left\{g \in C^1(\overline{\Omega}) : \int_\Omega g\tilde{v}^2(\cdot, \mu) = 0\right\}$$

such that $I(\mu, g) > 0$ for any $g \in U$.

Proof. Consider the following linear eigenvalue problem with weight function $\tilde{v}$:

$$-\mu\Delta\varphi - \beta\varphi = \lambda\tilde{v}\varphi \quad \text{in } \Omega, \qquad \frac{\partial\varphi}{\partial n}\Big|_{\partial\Omega} = 0. \tag{4.56}$$

Denote the eigenvalues of (4.56) by $\lambda_1 < \lambda_2 \leq \lambda_3 \cdots$ and the corresponding eigenfunctions by $\varphi_1, \varphi_2, \ldots$; from the definition of $\tilde{v}$ we see that $\lambda_1 \equiv -1$ for any μ, and φ_1 is a scalar multiple of $\tilde{v}$. The eigenfunctions can be chosen such that the following hold:

(a) $\int_\Omega \tilde{v}\varphi_i^2 = 1$, $i \geq 1$.

(b) $\int_\Omega \tilde{v}\varphi_i\varphi_j = 0$, $i \neq j$, $i, j \geq 1$. In particular, $\int_\Omega \tilde{v}^2\varphi_i = 0$ for any $i \geq 2$.

(c) $\{\varphi_i\}_{i=1}^{\infty}$ is a basis for the Hilbert space $L^2(\Omega,\tilde{v}) := \{\psi : \int_\Omega \tilde{v}\psi^2 < \infty\}$, with the inner product

$$\langle \phi, \psi \rangle = \int_\Omega \tilde{v}\phi\psi. \tag{4.57}$$

Now for any function φ satisfying $\int_\Omega \varphi\tilde{v} = 0$ (and in particular for $\varphi = g\tilde{v}$) $\varphi/\tilde{v}$ is orthogonal to $\tilde{v}$ in $L^2(\Omega,\tilde{v})$. Therefore

$$\frac{\varphi}{\tilde{v}} = \sum_{i=2}^{\infty} a_i\varphi_i, \tag{4.58}$$

where the convergence of (4.58) is in the $L^2(\Omega,\tilde{v})$ norm, which is equivalent to convergence in the L^2 norm. It is easy to check that the following hold:

$$\begin{aligned} (\mu\Delta + \beta - \tilde{v})^{-1}(\tilde{v}\varphi_i) &= \frac{\varphi_i}{-\lambda_i - 1} + c_i\tilde{v}, \qquad i \geq 2, \\ 2(\mu\Delta + \beta - 2\tilde{v})^{-1}(\tilde{v}\varphi_i) &= \frac{2\varphi_i}{-\lambda_i - 2}, \qquad i \geq 2, \end{aligned} \tag{4.59}$$

where c_i is some constant. Notice that $\lambda_i + 1 > 0$ for any $i \geq 2$ since $\lambda_1 = -1$. It follows from (4.59) that

$$(2(\mathcal{L} - \tilde{v})^{-1} - \mathcal{L}^{-1})(\tilde{v}\varphi_i) = \frac{-\lambda_i\varphi_i}{(\lambda_i + 2)(\lambda_i + 1)} - c_i\tilde{v}, \qquad i \geq 2. \tag{4.60}$$

Therefore from (4.58) and (4.60),

$$\int_\Omega \varphi(2(\mathcal{L} - \tilde{v})^{-1} - \mathcal{L}^{-1})\varphi = -\sum_{i=2}^{\infty} \frac{a_i^2\lambda_i}{(\lambda_i + 2)(\lambda_i + 1)}. \tag{4.61}$$

In the calculation of (4.61), the terms involving c_i vanish because $\int_\Omega \varphi\tilde{v} = 0$. Note that the λ_i depend on μ. Applying (4.61), (4.58) to $\varphi = g\tilde{v}$ and noting that $I(\mu, g)$ is continuous in g, we see that Theorem 4.7 follows from the following lemma. ∎

LEMMA 4.8. *There exists a unique $\mu^* > 0$ such that $\lambda_2(\mu) > 0$ if $\mu > \mu^*$, and $\lambda_2(\mu) < 0$ if $\mu < \mu^*$.*

Proof. Recall that φ_1 is a multiple of $\tilde{v}$, and thus by the variational characterization of eigenvalues we see that

$$\lambda_2(\mu) = \inf_{\varphi \in Q_\mu} \frac{\int_\Omega \left[\mu|\nabla\varphi|^2 - \beta\varphi^2\right]}{\int_\Omega \tilde{v}\varphi^2}, \tag{4.62}$$

where the set Q_μ is defined by

$$Q_\mu := \left\{ \varphi \in H^1(\Omega), \int_\Omega \varphi\tilde{v}^2 = 0 \right\}. \tag{4.63}$$

Notice that $\tilde{v} \to \beta_+$ uniformly as $\mu \to 0$. Therefore by choosing a suitable test function in (4.62) we see that $\lambda_2(\mu) < 0$ for sufficiently small μ; on the other hand, $\tilde{v} \to \int_\Omega \beta/|\Omega|$ uniformly as $\mu \to \infty$. Therefore standard arguments imply that $\lambda_2(\mu) > 0$ for large μ. In fact, $\lambda_2(\mu)/\mu \geq c > 0$ for some constant $c > 0$ when μ is large. This implies that $\lambda_2(\mu) = 0$ has at least one root. For the uniqueness, we need the following result.

Claim. If $\lambda_2(\mu_0) = 0$ for some $\mu_0 > 0$, then for $\mu \approx \mu_0$, $\lambda_2(\mu) < 0$ ($\mu < \mu_0$) and $\lambda_2(\mu) > 0$ ($\mu > \mu_0$).

To prove this assertion, consider the operator on $L^2(\Omega, \tilde{v})$ defined by

$$\tilde{\mathcal{L}}(\mu)\varphi = -\frac{1}{\tilde{v}}\,(\mu\Delta\varphi + \beta\varphi) \tag{4.64}$$

with domain $H^2_N(\Omega) := \{\psi \in H^2(\Omega), \frac{\partial\psi}{\partial n}|_{\partial\Omega} = 0\}$. Assume that for some μ_0 we have $\lambda_2(\mu_0) = 0$ for the second eigenvalue of $\tilde{\mathcal{L}}(\mu_0)$ and that this eigenvalue has multiplicity k. Let φ_i^0, $i = 1, \dots, k$, be an $L^2(\Omega, \tilde{v})$-orthonormal basis of the eigenspace of the eigenvalue 0. By standard perturbation theory (see [K2]), there exist functions $\varphi_i(\mu)$, $i = 1, \dots, k$, defined and smooth for μ in a neighborhood of μ_0 such that

$$\varphi_i(\mu^0) = \varphi_i^0, \qquad i = 1, \dots, k, \tag{4.65}$$

the space

$$X(\mu) := \operatorname{span}\{\varphi_i(\mu) : i = 1, \dots, k\} \tag{4.66}$$

is invariant under $\tilde{\mathcal{L}}(\mu)$, and the functions $\varphi_i(\mu)$ are orthonormal. (Note that $\varphi_i(\mu)$ may not be eigenfunctions of $\tilde{\mathcal{L}}(\mu)$.) Now, with respect to the basis $\varphi_i(\mu)$, the restriction of $\tilde{\mathcal{L}}(\mu)$ to $X(\mu)$ is represented by the matrix

$$M(\mu) = \left(\langle\tilde{\mathcal{L}}(\mu)\varphi_i, \varphi_j\rangle\right)_{1\le i,j\le k} = -\left(\int_\Omega (\mu\Delta + \beta)\varphi_i\varphi_j\right)_{1\le i,j\le k}. \tag{4.67}$$

We have $M(\mu_0) = 0$ and

$$M'(\mu_0) = \left(\int_\Omega \nabla\varphi_i^0\nabla\varphi_j^0\right)_{1\le i,j\le k}. \tag{4.68}$$

To prove (4.68), observe that

$$\begin{aligned}\frac{d}{d\mu}\int_\Omega (\mu\Delta + \beta)\varphi_i\varphi_j &= \int_\Omega \Delta\varphi_i\varphi_j + \int_\Omega (\mu\Delta + \beta)\,\frac{\partial\varphi_i}{\partial\mu}\,\varphi_j + \int_\Omega (\mu\Delta + \beta)\varphi_i\,\frac{\partial\varphi_j}{\partial\mu}\\ &= -\int_\Omega \nabla\varphi_i\nabla\varphi_j + \int_\Omega \frac{\partial\varphi_i}{\partial\mu}\,(\mu\Delta + \beta)\varphi_j + \int_\Omega (\mu\Delta + \beta)\varphi_i\,\frac{\partial\varphi_j}{\partial\mu}\\ &= I + II + III, \end{aligned} \tag{4.69}$$

where in the second equality the self-adjointness of the operator $\mu\Delta + \beta$ has been used. Obviously, $I \to -\int \nabla\varphi_i^0\nabla\varphi_j^0$ as $\mu \to \mu_0$. For II, as $\mu \to \mu_0$, $(\mu\Delta + \beta)\varphi_j \to (\mu_0\Delta + \beta)\varphi_j^0 \equiv 0$ since $\lambda_2(\mu_0) = 0$. This implies that $II \to 0$ as $\mu \to \mu_0$. A similar conclusion holds for III.

It remains to show that the matrix $M'(\mu_0)$ is positive definite. For any $\zeta = (\zeta_1, \dots, \zeta_k)$, set

$$\bar{\varphi} = \sum_{i=1}^k \zeta_i\varphi_i^0. \tag{4.70}$$

Then

$$\sum_{i,j=1}^k \int_\Omega \nabla\varphi_i^0\nabla\varphi_j^0\zeta_i\zeta_j = \int_\Omega |\nabla\bar{\varphi}|^2 \ge 0. \tag{4.71}$$

If the equality in (4.71) holds, then $\bar{\varphi}$ is equal to some constant, say c_1. By the equations for φ_i^0, we see that $\mu_0 \Delta \bar{\varphi} + \beta \bar{\varphi} = 0$, which implies that $\beta c_1 = 0$. That is, $c_1 = 0$. Hence $\bar{\varphi} = 0$. Since the φ_i^0 are orthogonal to each other in $L^2(\Omega, \tilde{v})$, we see that $\zeta_i = 0$ for all i. This together with (4.71) proves that the matrix $M'(\mu_0)$ is positive definite. Hence for μ close to μ_0, all eigenvalues of $M(\mu)$ are positive when $\mu > \mu_0$, and negative when $\mu < \mu_0$. □

5. The case $\tau \gg 1$. The goal of this section is to study (1.8) with $\tau \gg 1$, and Theorem 1.6 will be established. Throughout this section, we assume the hypotheses of the theorem to be satisfied.

LEMMA 5.1. *If τ is sufficiently large, $(0, \tilde{v})$ is unstable for any $\mu > 0$.*

Proof. It suffices to show that the principal eigenvalue, denoted by λ_1, of the linear eigenvalue problem (2.5) is negative. Note that λ_1 can be characterized by

$$\lambda_1 = \inf_{\{\varphi \in H^1 : \varphi \neq 0\}} \frac{\int_\Omega \left[\mu |\nabla \varphi|^2 - (\beta + \tau g - \tilde{v}) \varphi^2\right]}{\int_\Omega \varphi^2}. \tag{5.1}$$

With $\varphi = \tilde{v}$ in (5.1) we have

$$\lambda_1 \leq -\frac{\tau \int_\Omega g \tilde{v}^2}{\int_\Omega \tilde{v}^2} \tag{5.2}$$

for every $\tau > 0$ and $\mu > 0$. Since $\lim_{\mu \to +\infty} \tilde{v} = \int_\Omega \beta / |\Omega|$ uniformly (see (2.2b)) and $\int_\Omega g > 0$, there exists $\bar{\mu}$ such that if $\mu \geq \bar{\mu}$, $\int_\Omega g \tilde{v}^2 > 0$. Therefore for $\mu \geq \bar{\mu}$ and $\tau > 0$, $\lambda_1 < 0$.

Next we consider the case $0 < \mu \leq \bar{\mu}$: Choose a test function φ such that $\varphi \geq 0$, $\varphi \not\equiv 0$, and $\operatorname{supp} \varphi \sqsubset \Omega^+$. Then by (5.1) and the boundedness of $\tilde{v}(\cdot, \mu)$ (see (2.2a)), if $\tau \gg 1$ and $0 < \mu \leq \bar{\mu}$, we have

$$\lambda_1 \leq \frac{\int_\Omega \left[\bar{\mu} |\nabla \varphi|^2 + \tilde{v} \varphi^2\right] - \tau \int_\Omega g \varphi^2}{\int_\Omega \varphi^2} < 0. \tag{5.3}$$

This completes the proof of Lemma 5.1. □

The above lemma yields (i) of Theorem 1.6. We next prove (ii).

Proof of Theorem 1.6(ii). Recall that $\tilde{u} = \tilde{u}(x, \mu, \tau)$ is the unique positive solution of (1.3) with $\alpha = \beta + \tau g$, and the stability of $(\tilde{u}, 0)$ relative to (1.8) is determined by the sign of the principal eigenvalue of the linear eigenvalue problem (2.4b), (2.4d).

Claim. Let $\tilde{\Omega}$ be any compact subset of Ω^-. Then, as $\tau \to +\infty$ and $\tau/\mu \to +\infty$, $\tilde{u}/\mu \to 0$ uniformly with respect to $x \in \tilde{\Omega}$.

To establish this assertion, choose a domain $\hat{\Omega}$ such that $\tilde{\Omega} \sqsubset\sqsubset \hat{\Omega} \sqsubset\sqsubset \Omega^-$, where $\sqsubset\sqsubset$ means the inclusion of the closure of a given domain. By Proposition A.1 of [HLM] and the comparison arguments given there, for some $k_1 > 0$ one has $\tilde{u} \leq k_1 \tau^{2/3} \mu^{1/3}$ in Ω^- for $\tau \gg 1$ and $\tau/\mu \gg 1$. Here and below, $k_1, k_2, \ldots$ are constants independent of τ and μ, provided that both τ and τ/μ are sufficiently large.

Set $\hat{u} = \tilde{u} / \left(\tau^{2/3} \mu^{1/3}\right)$. Since $g < 0$ in $\hat{\Omega}$, we have

$$-\mu \Delta \hat{u} \leq -k_2 \tau \hat{u} \quad \text{in } \hat{\Omega}, \qquad \hat{u}\big|_{\partial \hat{\Omega}} \leq k_3 \tag{5.4}$$

for sufficiently large τ and τ/μ. Let w_τ be the unique solution of

$$-\mu \Delta w_\tau = -k_2 \tau w_\tau \quad \text{in } \hat{\Omega}, \qquad w_\tau\big|_{\partial \hat{\Omega}} = k_3. \tag{5.5}$$

It can be shown that

$$w_\tau(x) \le k_4 \exp\left[-k_5 \operatorname{dist}(x, \partial\hat{\Omega})(\tau/\mu)^{1/2}\right] \tag{5.6}$$

for every $x \in \hat{\Omega}$, where k_4 and k_5 are positive constants independent of x and large τ and τ/μ. Indeed, for any fixed $x \in \hat{\Omega}$, let B_x denote the ball centered at x with radius $\operatorname{dist}(x, \partial\hat{\Omega})$. Let z be the unique solution of

$$\mu\Delta z = k_2\tau z \quad \text{in } B_x, \qquad z|_{\partial B_x} = k_3. \tag{5.7}$$

By the maximum principle, $w_\tau \le k_3$ in $\hat{\Omega}$. Hence by the comparison principle, $w_\tau \le z$ in B_x. It is easy to see that z is radially symmetric, from which it can be shown that there are positive constants k_4, k_5, independent of τ, τ/μ, and x, such that $z(x) \le k_4 \exp[-k_5 \operatorname{dist}(x, \partial\hat{\Omega})(\tau/\mu)^{1/2}]$. This proves (5.6).

By the comparison principle, $\hat{u} \le w_\tau$ in $\hat{\Omega}$. Hence for $\tilde{u}$ we have, as $\tau/\mu \to +\infty$,

$$\tilde{u}(x)/\mu \le k_4(\tau/\mu)^{2/3} \exp\left[-k_5 \operatorname{dist}(x, \partial\hat{\Omega})(\tau/\mu)^{1/2}\right] \to 0 \tag{5.8}$$

for any $x \in \tilde{\Omega} \sqsubset\sqsubset \hat{\Omega}$. This proves the claim.

To continue with the proof of assertion (ii), let μ_0 be as in the theorem and fix $\epsilon > 0$. For $\mu \le \mu_0 - \epsilon$ and τ sufficiently large we want to choose a ψ such that

$$II := \int_\Omega \left[\mu|\nabla\psi|^2 - (\beta - \tilde{u})\psi^2\right] < 0, \tag{5.9}$$

which will yield the instability of $(\tilde{u}, 0)$. To this end, choose some $\tilde{\Omega} \sqsubset\sqsubset \Omega^-$ such that the number $\tilde{\mu}$, uniquely determined by the requirement that there exist a solution of the problem

$$\tilde{\mu}\Delta\tilde{\varphi} + \beta\tilde{\varphi} = 0 \quad \text{in} \quad \tilde{\Omega}, \qquad \tilde{\varphi} > 0 \quad \text{in } \tilde{\Omega}, \qquad \tilde{\varphi}|_{\partial\tilde{\Omega}} = 0, \tag{5.10}$$

satisfies $\tilde{\mu} \in \left(\mu_0 - \frac{\epsilon}{2}, \mu_0\right)$. We refer here to standard continuity and monotonicity properties of principal eigenvalues (see the definition of μ_0). Set

$$\psi = \begin{cases} \tilde{\varphi}(x), & x \in \tilde{\Omega}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.11}$$

Then it is easy to check that for $\mu \le \mu_0 - \epsilon$,

$$\begin{aligned} II &= \int_{\tilde{\Omega}} \left[\mu|\nabla\tilde{\varphi}|^2 - (\beta - \tilde{u})\tilde{\varphi}^2\right] \\ &= \int_{\tilde{\Omega}} \left[\left(\frac{\mu}{\tilde{\mu}} - 1\right)\beta + \tilde{u}\right]\tilde{\varphi}^2 \\ &\le \int_{\tilde{\Omega}} \left[\left(\frac{\mu_0 - \epsilon}{\mu_0 - \epsilon/2} - 1\right)\beta + \tilde{u}\right]\tilde{\varphi}^2 \\ &\le \left[\|\tilde{u}\|_{L^\infty(\tilde{\Omega})} - \frac{\epsilon}{2\mu_0}\min_{\bar{\Omega}}\beta\right]\int_{\tilde{\Omega}} \tilde{\varphi}^2 < 0, \end{aligned} \tag{5.12}$$

provided that $\tau \gg 1$ since $\tilde{u} \to 0$ uniformly in $\tilde{\Omega}$.

Next we consider the case $\mu \geq \mu_0 + \epsilon$ and show that $(\tilde{u}, 0)$ is the global attractor of (1.8) provided that $\tau \gg 1$. Since $(0, \tilde{v})$ is unstable in this case, the result will follow if the existence of a coexistence equilibrium is ruled out. We proceed by contradiction, assuming (u, v) is a coexistence state of (1.8) for a sequence of values $\mu \geq \mu_0 + \epsilon$ and $\tau \to +\infty$.

First consider the case when $\mu/\tau \to 0$. Since $v > 0$, we see that u/τ is a subsolution to the following problem:

$$\frac{\mu}{\tau}\Delta w_+ + w_+\left(\frac{\|\beta\|_\infty}{\tau} + g - w_+\right) = 0 \qquad \text{in } \Omega, \qquad \frac{\partial w_+}{\partial n}\Big|_{\partial\Omega} = 0. \tag{5.13}$$

As discussed in section 2, (5.13) has a unique positive solution w_+ which is globally attractive for the corresponding logistic parabolic equation. Therefore, by a standard super-sub solution method we conclude that $u/\tau \leq w_+$. Similarly, since $\beta \geq 0$ and $\|v\|_\infty \leq \|\beta\|_\infty$, u/τ is a supersolution of

$$\frac{\mu}{\tau}\Delta w_- + w_-\left(-\frac{\|\beta\|_\infty}{\tau} + g - w_-\right) = 0 \qquad \text{in } \Omega, \qquad \frac{\partial w_-}{\partial n}\Big|_{\partial\Omega} = 0. \tag{5.14}$$

Thus $u/\tau \geq w_-$ for the positive solution w_- of (5.14). It is not difficult to prove (cp. [HMP, proof of Lemma 3.4]) that both w_+ and w_- converge to g_+ uniformly as $\tau \to \infty$ and $\mu/\tau \to 0+$. This proves that $u/\tau \to g_+$ uniformly as $\mu/\tau \to 0$ and $\tau \to \infty$. Multiplying the equation for v by any $\varphi \in H^1(\Omega)$ and integrating it over Ω we have

$$\mu \int_\Omega \nabla v \cdot \nabla \varphi = \int_\Omega v\varphi(\beta - v - u). \tag{5.15}$$

It is easy to see that $\|v\|_{L^\infty}$ and $\|\nabla v\|_{L^2}$, and so $\|v\|_{H^1}$, are uniformly bounded. Therefore by the Sobolev embedding theorem, passing to a sequence if necessary, $v \to \bar{v}$ weakly in H^1, strongly in L^2 as $\mu/\tau \to 0$. Obviously, $\bar{v} \geq 0$ a.e. in Ω. Dividing (5.15) by τ and passing to the limit, we have

$$\int_\Omega g_+ \bar{v}\varphi = 0 \tag{5.16}$$

for every $\varphi \in H^1$. This implies that $\bar{v} = 0$ a.e. in Ω^+. Therefore $v \to 0$ weakly in $H^1(\Omega^+)$, and by the trace theorem (see [A]), $\bar{v}|_\Gamma = 0$.

We claim that $\bar{v} = 0$ a.e. in Ω^-. To prove this, choose $\varphi \in C_0^\infty(\Omega^-)$ in (5.15). For any coexistence state of (1.8), by the comparison principle we have $u \leq \tilde{u}$. Hence it follows from the claim above that $u/\mu \to 0$ uniformly in any compact subset of Ω^- as $\mu/\tau \to 0$ and $\tau \to \infty$. Dividing (5.15) by μ and passing to the limit we have either $\mu \to \bar{\mu}$ for some $\bar{\mu} \in (\mu_0, \infty)$ and

$$\int_{\Omega^-} \nabla \bar{v} \cdot \nabla \varphi = \bar{\mu}^{-1} \int_{\Omega^-} \varphi \bar{v}(\beta - \bar{v}), \qquad \varphi \in C_0^\infty(\Omega^-), \tag{5.17}$$

or $\bar{\mu} = \infty$. When $\bar{\mu} = \infty$, we see that $\bar{v}$ is a harmonic function, which together with $\bar{v}|_\Gamma = 0$ ensures that $\bar{v} = 0$ a.e. in Ω^-. When $\bar{\mu} < \infty$, since $\bar{v}|_\Gamma = 0$, by standard elliptic regularity we see that $\bar{v}$ is a classical solution of

$$\bar{\mu}\Delta \bar{v} + \bar{v}(\beta - \bar{v}) = 0 \quad \text{in } \Omega^-, \qquad \bar{v}|_\Gamma = 0, \qquad \bar{v} \geq 0 \quad \text{in } \Omega^-. \tag{5.18}$$

To show that $\bar{v} = 0$ a.e. in Ω^-, we argue by contradiction: If not, then by the strong maximum principle and the Hopf boundary lemma (see [PW, GNN]), $\bar{v} > 0$ in Ω^- and $\partial\bar{v}/\partial n < 0$, where n is the unit outward normal to $\partial\Omega$. Multiplying (5.18) by $\varphi_0 > 0$ (a solution of (1.14) corresponding to μ_0) and integrating in Ω^- we have

$$\left[-\frac{\bar{\mu}}{\mu_0} + 1\right] \int_{\Omega^-} \beta\varphi_0\bar{v} = \int_{\Omega^-} \varphi_0\bar{v}^2 > 0, \tag{5.19}$$

which is a contradiction since $\bar{\mu} > \mu_0$.

Therefore $\bar{v} = 0$ a.e. in Ω^-, and thus $\bar{v} = 0$ a.e. in Ω. Hence we see that $v \to 0$ weakly in $H^1(\Omega)$ and strongly in $L^2(\Omega)$. We shall use this to reach a contradiction. To this end, set $\hat{v} = v/\|v\|_{L^2}$. Then $\|\hat{v}\|_{L^2} = 1$ and $\hat{v}$ satisfies

$$\mu \int_\Omega \nabla\hat{v} \cdot \nabla\varphi = \int_\Omega \hat{v}\varphi(\beta - v - u) \tag{5.20}$$

for every $\varphi \in H^1(\Omega)$.

Since $\|\hat{v}\|_{L^2} = 1$, by letting $\varphi = \hat{v}$ in (5.20) we see that $\|\nabla\hat{v}\|_{L^2}$ is uniformly bounded, i.e., $\|\hat{v}\|_{H^1}$ is uniformly bounded. Hence we may assume that $\hat{v} \to v^*$ weakly in H^1 and strongly in L^2. In particular, this implies that $\|v^*\|_{L^2(\Omega)} = 1$, $v^* \geq 0$ a.e. in Ω. Similarly, as before we can show that $\int_\Omega v^*\varphi g_+ = 0$ for every $\varphi \in H^1(\Omega)$. Then $v^* = 0$ a.e. in Ω^+ and $v^*|_\Gamma = 0$. Again, by passing to a sequence if necessary, we may assume that $\mu \to \bar{\mu} \in (0, \infty]$. If $\bar{\mu} < \infty$, by similar argument as before we see that v^* solves

$$\bar{\mu}\Delta v^* + \beta v^* = 0 \quad \text{in } \Omega^-, \qquad v^*|_\Gamma = 0. \tag{5.21}$$

Note that $v^* \geq 0$, and $\|v^*\|_{L^2(\Omega^-)} = \|v^*\|_{L^2(\Omega)} = 1$ since $v^* = 0$ a.e. in Ω^+. By the strong maximum principle, $v^* > 0$ in Ω^-. This implies that $\bar{\mu} = \mu_0$, which is a contradiction! When $\bar{\mu} = \infty$, since $u/\mu \to 0$ uniformly in any compact subset of Ω^- we see that v^* is a harmonic function, which implies that $v^* = 0$ in Ω^- because $v^*|_\Gamma = 0$. This is again impossible since $\|v^*\|_{L^2(\Omega^-)} = 1$.

It remains to handle the case of $\mu/\tau \to \gamma$ for some $\gamma \in (0, \infty]$. For this case, it can be shown, by passing to a sequence if necessary, that u/τ converges to some positive function uniformly. (Arguments here are similar to those used in the case $\mu/\tau \to 0$ and are omitted.) This implies that $\beta(x) - u - v$ is strictly negative in Ω for large τ. However, by integrating the equation of v in Ω, we get $\int_\Omega v(\beta(x) - u - v) = 0$, which is a contradiction since v is positive in Ω. This completes the proof. □

Remark 5.2. Let $\mu^*(\tau)$ denote any value of μ where $(\tilde{u}, 0)$ changes stability. Theorem 1.6 simply says that $\lim_{\tau\to\infty} \mu^*(\tau) = \mu_0$. We suspect that such μ^* is unique, and any coexistence state of (1.8), if it exists, should also be unique and globally stable.

Appendix A. Proof of Proposition 4.4. The proof of Proposition 4.4 is given here after establishing some preliminary computational results. Throughout, (u, v) will be a coexistence state of (1.8), $\tilde{\mu}_1, u_1, v_1$ are given by (4.7), and (φ, ψ) is the solution of (2.3) corresponding to the eigenvalue $\lambda(\tau, s)$. Throughout the appendix we assume that $G(\tilde{\mu}) = 0$, where G is given by (1.9). We normalize (φ, ψ) such that $\int_\Omega \varphi^2 + \int_\Omega \psi^2 = 2\int_\Omega \tilde{v}^2$ and $\varphi > 0 > \psi$. In particular, for $\tau = 0$ $\varphi = \tilde{v}$, $\psi = -\tilde{v}$.

LEMMA A.1. *The following statements hold:*

(i)

$$\int_\Omega guv = 0. \tag{A.1}$$

(ii)

$$\lambda(\tau, s) = -\tau \int_\Omega g(\varphi v + \psi u) \Big/ \int (\varphi v - \psi u). \tag{A.2}$$

Note that by (4.7) and the normalization of (φ, ψ), we have for small $\tau > 0$

$$\varphi v - \psi u = \tilde{v}^2 + O(\tau)$$

and

$$\int_\Omega g(\varphi v + \psi u) = (1 - 2s) \int_\Omega g\tilde{v}^2 + O(\tau) = O(\tau).$$

In particular, the denominator in (A.2) is nonzero for small τ, and (A.2) implies

$$\lambda_\tau(0, s) = 0.$$

Proof of Lemma A.1. (i) Multiply (1.8a) by v, (1.8b) by u and subtract, obtaining

$$\mu(v\Delta u - u\Delta v) + \tau guv = 0. \tag{A.3}$$

The result follows on integrating (A.3) over Ω.

(ii) Multiply (2.3a) by v, (2.3b) by u and subtract, obtaining

$$-\lambda(\tau, s)(\varphi v - \psi u) = v\left[\mu\Delta\varphi + \varphi(\beta + \tau g - u - v)\right] - u\left[\mu\Delta\psi + \psi(\beta - u - v)\right]. \tag{A.4}$$

Integrating (A.4) over Ω and using (1.8) we deduce that

$$\begin{aligned} -\lambda(\tau, s) \int (\varphi v - \psi u) &= \int_\Omega \varphi\left[\mu\Delta v + v(\beta + \tau g - u - v)\right] - \int_\Omega \psi\left[\mu\Delta u + u(\beta - u - v)\right] \\ &= \tau \int_\Omega g(\varphi v + \psi u). \qquad \square \end{aligned}$$

Set

$$A = (\mathcal{L} - \tilde{v})^{-1}(\tilde{\mu}_1 \Delta\tilde{v}), \tag{A.5a}$$

$$B = (\mathcal{L} - \tilde{v})^{-1}(g\tilde{v}), \tag{A.5b}$$

$$C = \mathcal{L}^{-1}(g\tilde{v}), \tag{A.5c}$$

and expand the eigenfunctions φ, ψ in the form

$$\varphi = \tilde{v} + \tau\varphi_1(\cdot, s) + \tau^2\varphi_2(\cdot, \tau, s), \tag{A.6a}$$

$$\psi = -\tilde{v} + \tau\psi_1(\cdot, s) + \tau^2\psi_2(\cdot, \tau, s). \tag{A.6b}$$

Lemma A.2. *For some $\gamma_i \in \mathbb{R}$, we have*

(i)

$$u_1 = -s\left[A + sB + (1 - s)C\right] + \gamma_1\tilde{v}, \tag{A.7a}$$

$$v_1 = -(1 - s)\left[A + sB - sC\right] - \gamma_1\tilde{v}, \tag{A.7b}$$

(ii)

$$\varphi_1 = -A - 2sB + (2s-1)C + \gamma_2\tilde{v}, \tag{A.8a}$$

$$\psi_1 = A + (2s-1)B - (2s-1)C - \gamma_2\tilde{v}. \tag{A.8b}$$

Proof. (i) From direct calculation u_1, v_1 satisfy zero Neumann boundary conditions on $\partial\Omega$, and the following hold in Ω:

$$\tilde{\mu}\Delta u_1 + (\beta - \tilde{v})u_1 + s\tilde{v}(g - u_1 - v_1) + s\tilde{\mu}_1\Delta\tilde{v} = 0, \tag{A.9a}$$

$$\tilde{\mu}\Delta v_1 + (\beta - \tilde{v})v_1 + (1-s)\tilde{v}(-u_1 - v_1) + (1-s)\tilde{\mu}_1\Delta\tilde{v} = 0. \tag{A.9b}$$

Multiplying (A.9a), (A.9b) by $(1-s)$ and s, respectively, and subtracting, we find that

$$\mathcal{L}[(1-s)u_1 - sv_1] + s(1-s)g\tilde{v} = 0,$$

from which it follows on taking the inverse and using definition (A.5c) that

$$(1-s)u_1 - sv_1 = -s(1-s)C + \gamma_3\tilde{v}. \tag{A.10}$$

Adding (A.9a) and (A.9b) we have in a similar manner

$$u_1 + v_1 = -A - sB. \tag{A.11}$$

Then (A.7) follows from (A.10) and (A.11) by straightforward manipulation.

(ii) Since $\lambda_\tau(0,s) = 0$, it is easy to check that φ_1 and ψ_1 satisfy the following in Ω, together with zero Neumann boundary conditions:

$$\mathcal{L}\varphi_1 - s\tilde{v}(\varphi_1 + \psi_1) + \tilde{\mu}_1\Delta\tilde{v} + \tilde{v}(g - u_1 - v_1) = 0, \tag{A.12a}$$

$$\mathcal{L}\psi_1 - (1-s)\tilde{v}(\varphi_1 + \psi_1) - \tilde{\mu}_1\Delta\tilde{v} + \tilde{v}(u_1 + v_1) = 0. \tag{A.12b}$$

Adding (A.12a) and (A.12b), we have by an argument similar to that used in the previous lemma

$$\varphi_1 + \psi_1 = -B. \tag{A.13}$$

By (A.11), $u_1 + v_1 = -A - sB$. Substituting this and (A.13) in (A.12a), we obtain the equation which determines φ_1 up to an additive term $\gamma_4\tilde{v}$. Using definitions (A.5), it is easy to see that φ_1 given by (A.8a) satisfies that equation, which verifies (A.8a). This and (A.13) yield (A.8b). □

LEMMA A.3. *The following holds:*

$$\int_\Omega g\tilde{v}[2A + 2sB + (1-2s)C] = 0. \tag{A.14}$$

Proof. By (4.7) and (A.1),

$$0 = \int_\Omega guv = \tau\int_\Omega g\tilde{v}[sv_1 + (1-s)u_1] + O(\tau^2)$$

since $\int_\Omega g\tilde{v}^2 = 0$. Therefore

$$\int_\Omega g\tilde{v}[sv_1 + (1-s)u_1] = 0. \tag{A.15}$$

The result follows from (A.15) together with (A.7). □

Proof of Proposition 4.4. The result is a consequence of the following calculation of the denominator and numerator of (A.2). In the following $H_i(\tau, s)$ denote quantities that are uniformly bounded for $s \in [0,1]$ and small τ. From (A.6),

$$\int_\Omega (\varphi v - \psi u) = \int_\Omega \tilde{v}^2 + \tau H_1(\tau, s). \tag{A.16}$$

We now use (A.7) and (A.8) successively, obtaining

$$\begin{aligned}\int_\Omega g(\varphi v + \psi u) &= (1-2s)\int_\Omega g\tilde{v}^2 + \tau \int_\Omega g\tilde{v}[v_1 - u_1 + (1-s)\varphi_1 + s\psi_1] + \tau^2 H_2(\tau, s)\\ &= \tau \int_\Omega g\tilde{v}[(4s-2)A + (6s^2-4s)B + (-1+6s-6s^2)C] + \tau^2 H_3(\tau, s). \qquad \text{(A.17)}\end{aligned}$$

From (A.14) and (A.17),

$$\int_\Omega g(\varphi v + \psi u) = 2s(1-s)\tau \int_\Omega g\tilde{v}(C-B) + \tau^2 H_4(\tau, s). \tag{A.18}$$

As a consequence of (A.2), (A.16), and (A.18) we have

$$\lambda(\tau, s) = \tau^2 \left[\frac{2s(1-s)\int_\Omega g\tilde{v}(B-C)}{\int_\Omega \tilde{v}^2} + \tau H_5(\tau, s) \right]. \tag{A.19}$$

Since $\lambda(\tau, 0) = \lambda(\tau, 1) \equiv 0$ for all τ (see (4.23)), we have $H_5(0, \tau) = H_5(1, \tau) \equiv 0$. This implies that we can write H_5 as $H_5(\tau, s) = s(1-s)H_6(\tau, s)$. This proves part (i) of Proposition 4.4.

Part (ii) follows directly from Lemma A.3 and the relation

$$A = \tilde{\mu}_1 (\mathcal{L} - \tilde{v})^{-1} (\Delta \tilde{v}) = -\tilde{\mu}_1 \tilde{v}_\mu.$$

The latter equality is obtained by differentiating the equation for $\tilde{v}$ with respect to μ. □

Appendix B. Proof of Proposition 1.3. (i) Fix any β satisfying A1, and define the map

$$\Psi : (\mu, g) \mapsto \Psi(\mu, g) := \int_\Omega \tilde{v}^2(x, \mu) g(x)\, dx : (0, \infty) \times C^1(\overline{\Omega}) \to \mathbb{R}.$$

Then Ψ is smooth, it is linear in g, and, since $\tilde{v} > 0$, $\Psi(\mu, \cdot)$ is surjective. Thus 0 is a regular value of Ψ. Consequently, by the parametric transversality theorem [AR, Q, H2], for generic $g \in C^1(\overline{\Omega})$, 0 is a regular value of $G = \Psi(\cdot, g)$. This proves statement (i).

(ii) Fix any sequence $0 < \mu_1 < \mu_2 < \cdots$. We claim that for generic $\beta \in U = \{\beta \in C^1(\overline{\Omega}) : \int_\Omega \beta > 0\}$ the following condition holds for all $k = 1, 2, \ldots$:

$$\tilde{v}^2(\cdot, \mu_1), \ldots, \tilde{v}^2(\cdot, \mu_{k+1}) \text{ are linearly independent functions.} \tag{B.1}$$

Suppose for the moment that the claim is true. Observe that (B.1) allows us to choose a function $\tilde{g} \in C^1(\overline{\Omega})$ such that

$$\int_\Omega \tilde{g}(x)\tilde{v}^2(x, \mu_i)\, dx \int_\Omega \tilde{g}(x)\tilde{v}^2(x, \mu_{i+1})\, dx < 0 \quad (i = 1, \ldots, k) \tag{B.2}$$

(for example, we can choose $\tilde{g}$ as a linear combination of the functions in (B.1)). Then, for any g in a sufficiently small neighborhood of $\tilde{g}$, the function G has at least k zeros (at least one per each interval (μ_i, μ_{i+1})). By (i), we can choose g in this neighborhood such that G has only simple zeros, and, replacing $\tilde{g}$ by $-\tilde{g}$ if necessary, we can in addition take $\tilde{g} \in U$. Conclusion (ii) thus follows from our claim.

To prove the claim, it is clearly sufficient to show that for each fixed k, the set of all $\beta \in U$ for which (B.1) holds is open and dense in U. The openness is obvious, as $\tilde{v}$ is a continuous function of β. To prove the density, fix any $\tilde{\beta} \in U$. Arbitrarily close to $\tilde{\beta}$, we have to find $\beta \in U$ for which (B.1) holds. To this end let $0 = \lambda_1 < \lambda_2 \le \lambda_3 \cdots$ be the eigenvalues of $-\Delta$ on Ω under Neumann boundary condition. Let $\phi_1, \phi_2, \dots$ be an orthonormal basis of $L^2(\Omega)$ consisting of the corresponding eigenfunctions. Clearly, we can find $\bar{\beta} \in U$, as close to $\tilde{\beta}$ as we wish, such that

$$\beta_j := \int_\Omega (\bar{\beta}(x) - 1)\phi_j(x)\,dx = \int_\Omega \bar{\beta}(x)\phi_j(x)\,dx \neq 0 \quad (j = 2, 3, \dots); \tag{B.3}$$

that is, the Fourier coefficients of $\bar{\beta}$ with respect to the eigenfunctions are all nonzero (the first one is positive, as $\bar{\beta} \in U$). We further show that arbitrarily close to 1 there is a constant $\delta < 1$ such that (B.1) holds for

$$\beta = 1 - \delta + \delta\bar{\beta}.$$

This will complete the proof of density.

Denote the solution of (1.6) with $\beta = 1 - \delta + \delta\bar{\beta}$ by $\tilde{v}(\cdot, \mu, \delta)$. Observe that

$$\delta \mapsto \tilde{v}(\cdot, \mu, \delta) \in L^2(\Omega)$$

is an analytic function. Consider the Gram determinant

$$D(\delta) = \det\left(\int_\Omega \tilde{v}^2(x, \mu_i, \delta)\tilde{v}^2(x, \mu_j, \delta)\,dx\right)_{i,j=1}^{k+1}.$$

We have $D(\delta) \neq 0$ if and only if

$$\tilde{v}^2(\cdot, \mu_1, \delta), \dots, \tilde{v}^2(\cdot, \mu_{k+1}, \delta) \text{ are linearly independent functions.} \tag{B.4}$$

Since $\delta \mapsto D(\delta)$ is analytic, the desired property (i.e., $D(\delta) \neq 0$ for any $\delta < 1$ sufficiently close to 1) is established, provided $D \not\equiv 0$. To prove the latter we examine the limit $\delta \to 0$. At $\delta = 0$, we have the following equation for $\tilde{v}(\cdot, \mu, 0)$:

$$\mu\Delta v + (1 - v)v = 0 \quad \text{in } \Omega, \qquad \frac{\partial v}{\partial n} = 0 \quad \text{on } \partial\Omega.$$

Thus $\tilde{v}(\cdot, \mu, 0) \equiv 1$. As $\tilde{v}(\cdot, \mu, \delta)$ is a nondegenerate solution, we can expand

$$\tilde{v}(\cdot, \mu, \delta) = 1 + \delta z(\cdot, \mu) + O(\delta^2)$$

and find the equation for z to be

$$\mu\Delta z - z + \bar{\beta} - 1 = 0 \quad \text{in } \Omega, \qquad \frac{\partial z}{\partial n} = 0 \quad \text{on } \partial\Omega.$$

By an eigenfunction expansion,

$$z(\cdot, \mu) = \sum_{\ell=1}^{\infty} \frac{\beta_\ell}{\mu\lambda_\ell + 1}\phi_\ell,$$

where β_ℓ, $\ell = 1, 2, \ldots$, are the Fourier coefficients of $\bar{\beta} - 1$; cf. (B.3). Now,

$$\tilde{v}^2(\cdot, \mu_j, \delta) = 1 + 2\delta z(\cdot, \mu_j) + O(\delta^2), \quad j = 1, \ldots, k+1.$$

These functions are linearly independent if such are the functions

$$\delta Q z(\cdot, \mu_j) + O(\delta^2), \quad j = 1, \ldots, k+1, \tag{B.5}$$

where $Qu = u - \int u$ is the orthogonal projection with kernel $\operatorname{span}\{\phi_1\}$. Clearly, (B.5) are linearly independent for all small $\delta > 0$ if and only if the functions $Qz(\cdot, \mu_j)$ are linearly independent. This is equivalent to the independence of the $k+1$ infinite vectors

$$\left(\frac{\beta_\ell}{\mu_j \lambda_\ell + 1} \right)_{\ell=2}^{\infty}, \quad j = 1, \ldots, k+1,$$

and a sufficient condition for this is the independence of the $k+1$ vectors in $\mathbb{R}^{k+1}$

$$\left(\frac{\beta_{\ell_i}}{\mu_j \lambda_{\ell_i} + 1} \right)_{i=1}^{k+1}, \quad j = 1, \ldots, k+1, \tag{B.6}$$

for some choice of indices ℓ_i. We choose the ℓ_i such that $\tilde{\lambda}_i := \lambda_{\ell_i}$, $i = 1, \ldots, k+1$, are mutually distinct. To test for the independence, we compute the determinant of the matrix with rows (B.6). It is easy to check by listing the obvious roots of the determinant that

$$\begin{aligned} \det \left(\frac{\beta_{\ell_i}}{\mu_j \tilde{\lambda}_i + 1} \right)_{i,j=1}^{k+1} &= \prod_{m=1}^{k+1} \frac{\beta_{\ell_m}}{\lambda_m} \det \left(\frac{1}{\mu_j + \tilde{\lambda}_i^{-1}} \right)_{i,j=1}^{k+1} \\ &= \prod_{m=1}^{k+1} \frac{\beta_{\ell_m}}{\lambda_m} \prod_{1 \le i,j \le k+1} \frac{1}{\mu_j + \tilde{\lambda}_i^{-1}} \prod_{1 \le i < j \le k+1} (\mu_j - \mu_i) \prod_{1 \le i < j \le k+1} (\tilde{\lambda}_j^{-1} - \tilde{\lambda}_i^{-1}). \end{aligned}$$

Since $\beta_i \neq 0$ (see (B.3)), our choice of $\tilde{\lambda}_j$ implies that the determinant is nonzero. Therefore the vectors (B.6) and the functions (B.5) are linearly independent. This completes the proof. □

Acknowledgment. The authors express their gratitude to the anonymous referees of this paper for their careful reading and suggestions leading to an improvement of the exposition.

REFERENCES

[A] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[AR] J. Abraham and R. Robbin, *Transversal Mappings and Flows*, W. A. Benjamin, New York, 1967.

[CC] R. S. Cantrell and C. Cosner, *On the effects of spatial heterogeneity on the persistence of interacting species*, J. Math. Biol., 37 (1998), pp. 103–145.

[De] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, 1985.

[DHMP] J. Dockery, V. Hutson, K. Mischaikow, and M. Pernarowski, *The evolution of slow dispersal rates: A reaction-diffusion model*, J. Math. Biol., 37 (1998), pp. 61–83.

[F] W. H. Fleming, *A selection-migration model in population genetics*, J. Math. Biol., 2 (1975), pp. 219–233.

[G] P. Grant, *Ecology and Evolution of Darwin's Finches*, Princeton University Press, Princeton, NJ, 1999.

[GNN] B. Gidas, W.-M. Ni, and L. Nirenberg, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209–243.

[GT] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equation of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.

[H1] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Springer-Verlag, New York, 1981.

[H2] D. Henry, *Perturbation of the Boundary for Boundary Value Problems of Partial Differential Operators*, Cambridge University Press, Cambridge, to appear.

[He] P. Hess, *Periodic-Parabolic Boundary Value Problems and Positivity*, Longman Scientific & Technical, Harlow, UK, 1991.

[HSW] S. Hsu, H. Smith, and P. Waltman, *Competitive exclusion and coexistence for competitive systems on ordered Banach spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 4083–4094.

[HLM] V. Hutson, Y. Lou, and K. Mischaikow, *Spatial heterogeneity of resources versus Lotka-Volterra dynamics*, J. Differential Equations, 185 (2002), pp. 97–136.

[HLMV] V. Hutson, J. Lopez-Gomez, K. Mischaikow, and G. Vickers, *Limit behaviour for a competing species problem with diffusion*, in Dynamical Systems and Applications, World Sci. Ser. Appl. Anal. 4, World Scientific, River Edge, NJ, 1995, pp. 343–358.

[HMP] V. Hutson, K. Mischaikow, and P. Poláčik, *The evolution of dispersal rates in a heterogeneous time-periodic environment*, J. Math. Biol., 43 (2001), pp. 501–533.

[K1] T. Kato, *Superconvexity of the spectral radius, and convexity of the spectral bound and the type*, Math. Z., 180 (1982), pp. 265–273.

[K2] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[LN] Y. Lou and W. M. Ni, *Diffusion, self-diffusion and cross-diffusion*, J. Differential Equations, 72 (1996), pp. 79–131.

[L] A. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems*, Birkhäuser, Berlin, 1995.

[PW] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, 2nd ed., Springer-Verlag, Berlin, 1984.

[Q] F. Quinn, *Transversal approximation on Banach manifolds*, in Global Analysis, Proc. Sympos. Pure Math. 15, AMS, Providence, RI, 1970, pp. 213–222.

[SK] N. Shigesada and K. Kawasaki, *Biological Invasions: Theory and Practice*, Oxford Series in Ecology and Evolution, Oxford University Press, Oxford, New York, Tokyo, 1997.

SIAM J. MATH. ANAL.
Vol. 35, No. 2, pp. 492–519

VISCOUS APPROXIMATION OF STRONG SHOCKS OF SYSTEMS OF CONSERVATION LAWS*

FREDERIC ROUSSET†

Abstract. We consider a piecewise smooth solution of a one-dimensional hyperbolic system of conservation laws with a single Lax or overcompressive noncharacteristic shock. We show that it is a zero dissipation limit assuming that there exist linearly stable viscous profiles associated with the discontinuities. In particular, following the approach of [Grenier and Rousset, *Comm. Pure Appl. Math.*, 54 (2001), pp. 1343–1385], we replace the smallness condition obtained by energy methods in [Goodman and Xin, *Arch. Ration. Mech. Anal.*, 121 (1992), pp. 235–265] in the case of Lax shocks by a weaker spectral assumption which is sharp.

Key words. vanishing viscosity, stability of viscous shocks

AMS subject classifications. 35L65, 35L67

DOI. 10.1137/S0036141002403110

1. Introduction.

Consider a one-dimensional system of conservation laws

$$u_t + f(u)_x = 0 \tag{1}$$

with a smooth flux $f : \mathbb{R}^n \to \mathbb{R}^n$. We assume that (1) is hyperbolic: there exist smooth matrices $P(u)$, $D(u)$ such that $P(u)^{-1} f'(u) P(u) = D(u)$, where $D(u)$ is a diagonal matrix. The eigenvalues of D will be denoted by $\lambda_1(u), \dots, \lambda_n(u)$. We consider a piecewise smooth solution u which is a distributional solution of (1) in the domain $\mathbb{R} \times [0, T^*]$ with a single shock; that is to say that $u(x,t)$ is smooth at any point (x,t), $x \neq s(t)$, where $x = s(t)$ is a smooth curve in the (x,t) plane. Moreover the limits

$$\partial_x^k u^-(t) := \partial_x^k u(s(t) - 0, t) = \lim_{x \to s(t)^-} \partial_x^k u(x,t),$$

$$\partial_x^k u^+(t) := \partial_x^k u(s(t) + 0, t) = \lim_{x \to s(t)^+} \partial_x^k u(x,t)$$

exist. We also assume that the shock is a noncharacteristic Lax or overcompressive shock that is

$$\lambda_1(u^-(t)), \dots, \lambda_{n-i^-}(u^-(t)) < s'(t) < \lambda_{n-i^-+1}(u^-(t)) < \cdots < \lambda_n(u^-(t)), \tag{2}$$

$$\lambda_1(u^+(t)), \dots, \lambda_{i^+}(u^+(t)) < s'(t) < \lambda_{i^++1}(u^+(t)) < \cdots < \lambda_n(u^+(t)), \tag{3}$$

$i^- + i^+ = n + q + 1$. We have a Lax shock when $q = 0$ and an overcompressive shock when $q > 0$. A general conjecture is that the admissible solutions of (1) can be obtained as limits of solutions of

$$u_t^\varepsilon + f(u^\varepsilon)_x = \varepsilon u_{xx}^\varepsilon \tag{4}$$

*Received by the editors February 25, 2002; accepted for publication (in revised form) September 14, 2002; published electronically August 6, 2003.
http://www.siam.org/journals/sima/35-2/40311.html
†UMPA, ENS-Lyon, 46 Allée d' Italie, 69364 Lyon Cedex 07, France (frousset@umpa.ens-lyon.fr).

when ε tends to zero. This conjecture has been proved for scalar conservation laws by using the maximum principle [19] and for some special 2×2 systems (for which a weak maximum principle holds) by the method of compensated compactness [4]. This conjecture motivated the work of Goodman and Xin [8], who have shown that a solution of (1) with a Lax shock is the limit of a solution of (4) under a smallness assumption on the amplitude of the shock

$$\sup_{t\in[0,T]} |u(s(t)+0,t) - u(s(t)-0,t)| \leq \eta_0, \tag{5}$$

η_0 being sufficiently small.[1] Under this smallness assumption and in the case of Lax shocks, they have built an approximate solution u^{app} of (4) thanks to the method of matched asymptotic expansions. More precisely, away from the shock there is an outer expansion

$$O(x,t) = u(x,t) + \varepsilon u_1(x,t) + \varepsilon^2 u_2(x,t), \tag{6}$$

where u is the piecewise smooth solution of (1) that we considered, and u_i, $i \geq 2$, are solutions of some linear hyperbolic systems. Similarly, near the shock there is an inner expansion

$$I(x,t) = V(\xi,t) + \varepsilon V_1(\xi,t) + \varepsilon^2 V_2(\xi,t), \tag{7}$$

where $\xi = \frac{x-s(t)}{\varepsilon} + \delta(t)$ is the stretched variable. We also have an expansion of $\delta(t)$ which is a perturbation of the shock position:

$$\delta(t) = \delta_0(t) + \varepsilon\delta_1(t).$$

Note that it is actually possible to get expansions at every order by assuming more regularity on u.

The viscous shock profile $V(\xi,t)$ is a solution of

$$V_\xi = f(V) - f(u^-(t)) - s'(t)(V - u^-(t)) \tag{8}$$

such that

$$\lim_{\xi\to\pm\infty} V(\xi,t) = u^\pm(t). \tag{9}$$

The higher order terms are solutions of some linear ordinary differential equations. Taking a smooth function m such that $m(x) = 1$ when $|x| \leq 1$ and $m(x) = 0$, when $|x| \geq 2$, one finally gets an approximate solution of (4),

$$u^{app}(x,t) = m\left(\frac{x-s(t)}{\varepsilon^\gamma}\right) I(x,t) + \left(1 - m\left(\frac{x-s(t)}{\varepsilon^\gamma}\right)\right) O(x,t) + d(x,t), \tag{10}$$

where d is a higher order correction term that allows us to put the error term in conservative form (see section 2.2 for more details), and $\gamma \in (\frac{2}{3}, 1)$. The approximate solution u^{app} then solves

$$u^{app}_t + f(u^{app})_x - \varepsilon u^{app}_{xx} = \left(f(u^{app}) - f(u^{app} - d)\right)_x.$$

[1] More recently the convergence of u^ε was proved in [1] if the initial data of (4) has sufficiently small total variation.

It is then shown that a solution u^ε of (4) with initial condition $u^{app}(x,0)$ tends to u^{app} when ε goes to zero. As explained in [7], it is better to deal with an integrated equation when we study the stability of shocks. It relies on the fact that in case $\lambda_x < 0$ the scalar equation

$$v_t + \lambda(x)v_x = 0$$

has a better behavior than the equation

$$v_t + (\lambda(x)v)_x = 0$$

when we make energy estimates. Actually, the first one gives

$$\frac{d}{dt}\int_{\mathbb{R}} v^2(t,x)\,dx - \int_{\mathbb{R}} \lambda_x v^2\,dx = 0,$$

and hence $\frac{d}{dt}\int_{\mathbb{R}} v^2(t,x)\,dx \le 0$, as the second one gives

$$\frac{d}{dt}\int_{\mathbb{R}} v^2(t,x)\,dx + \int_{\mathbb{R}} \lambda_x v^2\,dx = 0,$$

which does not have the good sign. Consequently, as in [8], we set $w_x = u^\varepsilon - u^{app}$, which gives

$$w_t + f'(u^{app})w_x - \varepsilon w_{xx} = Q(u^{app}, w_x) + (f(u^{app}) - f(u^{app} - d)), \tag{11}$$

$$w(x,0) = 0.$$

The error term $Q(u^{app}, w_x)$ is $\mathcal{O}(|w_x|^2)$. The convergence of w_x to zero is shown in [8] by means of energy estimates. This method leads to a smallness assumption on the amplitude of the shock. Consequently, the convergence is shown only for weak shocks satisfying (5). The same kind of result with a more precise description of the convergence and the study of the evolution of an initial layer into a shock layer is shown in [20] using the method of approximate Green's functions in [14], but there is still the restriction (5). These results can be seen as a kind of generalization of the results of [7], [12], [18] about the asymptotic stability of weak viscous shock profiles. Nevertheless, the smallness assumption is not sharp even when we study the linear stability of shock profiles. There are strong shocks for which a viscous shock profile $V(\xi,\tau)$ associated is linearly stable for zero-mass perturbation; that is to say that the solutions of

$$\partial_t u + \partial_\xi(f'(V(\xi,\tau))u) - \partial_{\xi\xi}u = 0,$$

$$u(\xi,0) = \partial_\xi u_0(\xi)$$

tend to zero when $t \to +\infty$. However, when the shock does not satisfy a smallness assumption, the classical energy estimates are not sufficient to prove the stability. The main difficulty is due to the fact that the energy tends to zero when $t \to +\infty$ but not in a monotonous way. To conclude, refined methods are needed as in [10], [21]. Note that these methods do not apply in our time-dependent case since they both rely on the Laplace transform. The aim of this paper is to show the convergence of u^ε towards u^{app} when the viscous profiles are linearly stable using the method of [9]. This implies the result of [8] in the case of weak Lax shocks.

We now present our hypotheses more precisely. First we assume the following:

(H0) $\forall t \in [0, T^*]$, there exists a viscous profile which is a solution of (8), (9). Moreover in the case of overcompressive shocks ($q \geq 1$), we make the generic assumption that there exists a q-dimensional manifold of profiles $\varphi(\xi, t, h)$, $h \in U$, U being a vicinity of zero in $\mathbb{R}^q$.

Note that thanks to (2) and (3), $u^+(t)$ and $u^-(t)$ are hyperbolic rest points for the ordinary differential equation (8); consequently, we have for any α, β,

$$|\partial_t^\alpha \varphi(\xi, t, h)) - \partial_t^\alpha u^+(t)| \leq e^{-\omega\xi} \ \forall \xi \geq 0, \quad |\partial_t^\alpha \varphi(\xi, t, h)) - \partial_t^\alpha u^-(t)| \leq e^{\omega\xi} \ \forall \xi \leq 0,$$

$$|\partial_\xi^\alpha \varphi(\xi, t, h)| \leq e^{-\omega|\xi|}, \ |\partial_h^\beta \varphi(\xi, t, h)| \leq e^{-\omega|\xi|} \ \forall \xi$$

for some $\omega > 0$.

Let us formalize the notion of linear stability. Consider for each $\tau \in [0, T^*]$ the operators

$$\mathcal{L}_\tau v = v_{\xi\xi} - \Big(f'(\varphi(\xi, \tau, 0)) - s'(\tau)\Big) v_\xi$$

and

$$\tilde{L}_\tau v = v_{\xi\xi} - \Big(f'(\varphi(\xi, \tau, 0) - s'(\tau)) v\Big)_\xi$$

in L^p with domain $W^{2,p}$ for $p < +\infty$. For each time τ we want the profile $\varphi(\xi, \tau, 0)$ to be linearly stable. As stated in [21], it is equivalent to the Evans function criterion

(H) $\forall \tau \in [0, T^*]$, $\tilde{L}_\tau$ is such that $\tilde{D}_\tau(\lambda) \neq 0 \, \forall \lambda, \Re\, \lambda \geq 0$, $\lambda \neq 0$, and $\tilde{D}_\tau^{(q+1)}(0) \neq 0$,

where $\tilde{D}_\tau$ is the Evans function of $\tilde{L}_\tau$. We refer to [6] and [21] for the definition of the Evans function. This hypothesis means that $\tilde{L}_\tau$ does not have eigenvalues in the closed right half plane and that zero is semisimple in the effective spectrum. (We again refer to [21] for the definition of the effective spectrum.) This hypothesis is necessary for the linear stability of Lax shocks and overcompressive shocks as stated in [21]. Note that (H) can be checked by energy methods in the case of weak shocks. We point out that this hypothesis has an equivalent form (see [21]):

(H′) $\forall \tau \in [0, T^*]$, the "integrated operator" $\mathcal{L}_\tau$ is such that $\mathcal{D}_\tau(\lambda) \neq 0 \, \forall \lambda, \Re\lambda \geq 0$,

where $\mathcal{D}_\tau$ is the Evans function associated with $\mathcal{L}_\tau$. This second form is more useful for our work since we deal with an integrated equation (11). Note that we can prove by a direct computation that $\mathcal{D}_\tau(0) = \tilde{D}_\tau^{(q+1)}(0)$. Note also that this is another justification of the better behavior of the "integrated" equation explained in [7]. We finally point out that the assumption $\tilde{D}_\tau^{(q+1)}(0) \neq 0$ implies that the eigenspace of $\tilde{L}_\tau$ associated with the eigenvalue zero is generated by $\varphi_\xi, \varphi_{h^1}, \ldots, \varphi_{h^q}$.

The main theorems of this paper are as follows.

THEOREM 1. *Assume* (H0) *and* $\tilde{D}'_\tau(0) \neq 0$ *for any* $\tau \in [0, T^*]$ *in the case of Lax shocks and* $\tilde{D}_0^{(q+1)}(0) \neq 0$ *in the case of overcompressive shocks; then there exists an approximate solution of* (4) *defined on* $[0, T]$ *for some* $T > 0$. *Moreover, in the case of Lax shocks, we have* $T = T^*$.

THEOREM 2. *Assume* (H0) *and* (H)*; then we have*

$$||u^\varepsilon - u^{app}||_{L^\infty([0,T], L^1(\mathbb{R}))} \to 0$$

and

$$||u^\varepsilon - u^{app}||_{L^\infty([0,T]\times\mathbb{R})} \to 0$$

when ε tends to zero. Consequently, we have

$$||u^\varepsilon - u||_{L^\infty([0,T],L^1(\mathbb{R}))} \to 0$$

and, for any $\eta \in (0,1)$,

$$\sup_{0\le t\le T,\ |x-s(t)|\ge\varepsilon^\eta} |u^\varepsilon(x,t) - u(x,t)| \to 0$$

when ε tends to zero. Moreover, in the case of Lax shocks, we have $T^* = T$.

We can also get a convergence in $L^\infty([0,T], W^{m,1}(\mathbb{R}))$, where m depends only on the regularity of u. Note that this theorem is sharp since (H) is necessary for linear stability of each profile and since we can expect that linear instability implies nonlinear instability as in [3]. Moreover this theorem implies the result of [8] up to a change of L^2-type to L^1-type Sobolev space.

In the first part of this paper, we prove Theorem 1, and we show that the error term of the approximate solution satisfies the same estimates as in [8]. The fact that even if we assume that $\tilde{D}_\tau^{(q+1)} \neq 0$ for any $\tau \in [0, T^*]$, we may have $T < T^*$ in the case of overcompressive shocks will appear in this construction. Actually, since we also look for a perturbation $h(t)$ in U of the viscous profile, the construction can be led only for small t since in general $U \neq \mathbb{R}^n$. A motivating example is given in [5]. Hence the convergence will be shown only for T sufficiently small. It seems to be related to the "instability" of overcompressive shocks that is pointed out in [13].

Next, we set $z = x - s(t) + \varepsilon\delta(t)$; hence (11) becomes

$$\begin{aligned} &\tilde{w}_t + \Big(f'(\tilde{u}^{app}(z,t)) - s'(t) + \varepsilon\delta'(t)\Big)\tilde{w}_z - \varepsilon\tilde{w}_{zz} \\ &= Q(\tilde{u}^{app}, \tilde{w}_z) + (f(\tilde{u}^{app}) - f(\tilde{u}^{app} - \tilde{d})), \end{aligned} \tag{12}$$

where $\tilde{u}^{app}(z,t) = u^{app}(z + s(t) - \varepsilon\delta(t), t)$, $\tilde{d}(z,t) = d(z + s(t) - \varepsilon\delta(t))$, and $\tilde{w}(z,t) = w(z + s(t) - \varepsilon\delta(t), t)$. We note that now the viscous shock is located at $z = 0$. In the second part of this paper, we use the iterative construction of Green's functions described in [9] to get a Green's function for the operator

$$\mathcal{L}^\varepsilon w = w_t + (f'(\tilde{u}^{app}) - s'(t) + \varepsilon\delta'(t))w_z - \varepsilon w_{zz}$$

and to derive uniform estimates in ε on the L^1 norm of this Green's function. To use this method, we construct an approximate Green's function. This function is obtained by combining approximate Green's functions away from the shock and the exact Green's function for the shock problem with frozen time in the viscous profiles built in [21]. Finally we show the nonlinear convergence using a standard argument for parabolic equations as in [10].

2. Construction and estimates on the error term of the approximate solution. In this section we explain how the hypothesis (H) allows us to make the same construction as in [8] with the same type of estimates on the error term but without using the smallness assumption (5).

2.1. Proof of Theorem 1. We look for an outer expansion under the form (6) where u_1, u_2 are smooth up to $x = s(t)$, but since we also want to deal with overcompressive shocks, we look for an inner expansion under the more general form

$$I(x,t) = \varphi(\xi, t, h_0(t)) + \varepsilon\Big(\partial_h\varphi(\xi, t, h_0(t))h_1(t) + V_1(\xi,t)\Big) + \varepsilon^2 V_2(\xi,t),$$

where ξ is still under the form $\xi = \frac{x-s(t)}{\varepsilon} + \delta(t)$, φ is a solution of

$$\varphi_\xi = f(\varphi) - f(u^-(t)) - s'(t)(\varphi - u^-(t)),$$

$\varphi(-\infty, t, h_0(t)) = u^-(t)$, and $\varphi(+\infty, t, h_0(t)) = u^+(t)$. In the case of Lax shocks, we have $h_0(t) = 0$ and $h_1(t) = 0$. In the case of overcompressive shocks, $h_0(t)$ is a parametrization of the viscous profiles. The method of matched asymptotic expansion leads to

$$u_{1t} + (f'(u)u_1)_x = u_{xx}, \tag{13}$$

$$u_{2t} + (f'(u)u_2)_x = u_{1xx} + (f''(u)(u_1,u_1))_x \tag{14}$$

for the outer expansion and to

$$V_{1\xi\xi} = \Big((f'(\varphi) - s'(t))V_1\Big)_\xi + \delta_0'(t)\varphi_\xi + \partial_h\varphi h_0'(t) + \varphi_t, \tag{15}$$

$$\begin{aligned} V_{2\xi\xi} = {} & \Big((f'(\varphi) - s'(t))V_2\Big)_\xi + \delta_1'(t)\varphi_\xi + \partial_h\varphi h_1'(t) \\ & + (\partial_t\partial_h\varphi + \delta_0'\partial_\xi\partial_h\varphi + \partial_h^2\varphi\cdot h_0')\cdot h_1 \\ & + \Big(f''(\varphi)(\partial_h\varphi h_1 + V_1, \partial_h\varphi h_1 + V_1)\Big)_\xi + V_{1t} + \delta_0' V_{1\xi} \end{aligned} \tag{16}$$

for the inner expansion. Here we have used

$$(\partial_h\varphi)_{\xi\xi} = \Big((f'(\varphi) - s'(t))\partial_h\varphi\Big)_\xi.$$

Moreover we want the two solutions to be valid in an intermediate zone. This gives the matching conditions

$$V_1(\xi,t) = u_1^\pm(t) + (\xi - \delta_0)\partial_x u^\pm(t) + o(1), \tag{17}$$

$$\begin{aligned} V_2(\xi,t) = {} & u_2^\pm(t) + (\xi-\delta_0)\partial_x u_1^\pm(t) - \delta_1(t)\partial_x u^\pm(t) \\ & + \frac{1}{2}(\xi-\delta_0)^2\partial_x^2 u^\pm(t) + o(1) \end{aligned} \tag{18}$$

when $\xi \to \pm\infty$.

We give the construction of u_1, V_1, δ_0, and h_0, which are built simultaneously. The construction of u_2, V_2, δ_1, and h_1 will be similar. As in [8], it is convenient to deal with bounded solutions; hence we first lift the dominant term in the asymptotic expansion (17) by choosing D_1 smooth such that

$$D_1(\xi,t) = \begin{cases} \xi\partial_x u(s(t)-0,t), & \xi < -1, \\ \xi\partial_x u(s(t)+0,t), & \xi > 1. \end{cases}$$

Consequently, $U_1 = V_1 - D_1$ solves

$$U_{1\xi\xi} - \Big((f'(\varphi) - s'(t))U_1\Big)_\xi = \delta_0'(t)\varphi_\xi + \partial_h\varphi h_0'(t) + g(\xi,t,h_0(t)),$$

where

$$g(\xi,t,h_0(t)) = \Big((f'(\varphi)-s')D_1\Big)_{\xi} + \partial_t\varphi(\xi,t,h_0(t)). \tag{19}$$

Note that

$$g(\xi,t,h_0(t)) = \Big(f'(u^{\pm})-s'\Big)\partial_x u^{\pm} + \frac{d}{dt}u^{\pm}(t) + \mathcal{O}(e^{-\alpha|\xi|}), \qquad \alpha>0.$$

Hence

$$|g(\xi,t,h_0(t))| \le Ce^{-\alpha|\xi|},$$

since thanks to (1)

$$\partial_t u^{\pm} + f'(u^{\pm})\partial_x u^{\pm} = 0.$$

Moreover (17) becomes

$$U_1(\xi,t) = u_1^{\pm}(t) - \delta_0\partial_x u_0^{\pm}(t) + o(1), \qquad \xi\to\pm\infty. \tag{20}$$

Setting

$$G(\xi,t,h_0(t)) = \int_0^{\xi} g(\eta,t,h_0(t))\,d\eta, \tag{21}$$

we get

$$\begin{aligned} U_{1\xi} - \Big(f'(\varphi)-s'(t)\Big)U_1 = \delta_0'(t)\varphi + \int_0^{\xi}\partial_h\varphi(\zeta,t,h(t))h_0'(t)\,d\zeta \\ + G(\xi,t,h_0(t)) + c(t), \end{aligned} \tag{22}$$

where $c(t)$ is an integration constant.

We have to solve the coupled systems (13), (22), (20). The first step of the proof is to show that with t, δ_0, h_0 fixed we can find a solution of (22) having limits at $\pm\infty$ which depend on t, δ_0, h_0. Then, using these limits and the matching conditions (20), we rewrite (13) as a hyperbolic initial boundary value problem where δ_0 and h_0 are also unknown. We use the techniques of [11], [16] to solve it. Finally using the δ_0 and h_0 obtained, we take the corresponding solutions of (22) which have limits which by construction verify the matching condition (20).

Let us consider the homogeneous ordinary differential system

$$v_{\xi} = \Big(f'(\varphi)-s'(t)\Big)v. \tag{23}$$

Since

$$\begin{aligned} |f'(\varphi)-f'(u^+(t))| &\le e^{-\alpha\xi}, \qquad \xi\ge 0,\\ |f'(\varphi)-f'(u^-(t))| &\le e^{\alpha\xi}, \qquad \xi\le 0, \end{aligned}$$

and since $f'(u^+(t))-s'(t)$, $f'(u^-(t))-s'(t)$ do not have eigenvalues on the imaginary axis thanks to (2), (3), Proposition 1 in [2, Chap. 2] concerning the roughness of exponential dichotomy states that the differential system (23) has an exponential dichotomy on both half lines. Hence we can use Lemma 4.2 of [17]. Let us denote

by $C_b(\mathbb{R})$ the space of continuous bounded functions on $\mathbb{R}$ and by $C_b^1(\mathbb{R})$ the space of C^1 functions bounded together with their derivative. We consider the linear operator A_{t,h_0} defined from C_b^1 to C_b^0 by

$$A_{t,h_0}v = v' - \Big(f'(\varphi) - s'(t)\Big)v.$$

Lemma 4.2 of [17] states that A_{t,h_0} is a Fredholm operator. The index of A_{t,h_0} is $\dim E + \dim F - n$, where E and F are the stable and unstable subspaces of (23); that is to say

$$E = \{x \in \mathbb{R}^n,\ \exists v \text{ solution of (23)}, v(0) = x,\ v(+\infty) = 0\},$$
$$F = \{x \in \mathbb{R}^n,\ \exists v \text{ solution of (23)}, v(0) = x,\ v(-\infty) = 0\}.$$

Using [2] again, E has the same dimension as the stable subspace of

$$v' = \Big(f'(u^+(t)) - s'(t)\Big)v,$$

i.e., i^+, thanks to (3). Similarly F has the same dimension as the unstable subspace of

$$v' = \Big(f'(u^-(t)) - s'(t)\Big)v,$$

i.e., i^-, thanks to (2). Consequently, the index of the Fredholm operator (A_{t,h_0}) is $q+1$. Moreover, in the case of Lax shocks, our assumption $\tilde{D}_t'(0) \neq 0$ for any $t \in [0, T^*]$ implies that $\ker\ A_t = \mathrm{Span}(\varphi_\xi)$ (in this case, there is no dependence on h); hence $\dim \ker A_t = 1$ and A_t is onto. In the case of overcompressive shocks, by continuity we have $\tilde{D}_{t,h_0}^{(q+1)}(0) \neq 0$ for any $t \in [0, T]$, for some positive T, and for any h_0 in a vicinity of zero in $\mathbb{R}^q$, where $\tilde{D}_{t,h_0}$ is the Evans function of the operator

$$\tilde{L}_{t,h_0}v = v_{\xi\xi} - \Big((f'(\varphi(\xi,t,h_0)) - s'(t))v\Big)_\xi.$$

Consequently $\ker\ A_{t,h_0} = \mathrm{Span}(\varphi_\xi, \varphi_{h^1}, \ldots, \varphi_{h^q})$; hence $\dim \ker A_{t,h_0} = q+1$ and A_{t,h_0} is onto. Finally, there exists a bounded solution U_1 of (22) in a closed supplementary of the kernel for any $t \in [0, T^*]$, $\delta_0 \in \mathbb{R}$ in the case of Lax shocks and for any $t \in [0, T]$, $\delta_0 \in \mathbb{R}$ and h_0 in a vicinity of zero in the case of overcompressive shocks.

We now show that every bounded solution U_1 of (22) actually has limits at both $\pm\infty$ which do not depend on its choice, and we compute them. For $\xi \geq 0$, we have

$$U_{1\xi} = \Big(f'(u^+(t)) - s'(t)\Big)U_1 + F, \tag{24}$$

where

$$F = \Big(f'(\varphi) - f'(u^+(t))\Big)U_1 + \delta_0'\varphi + \int_0^\xi \partial_h\varphi \cdot h_0'(t) + G + c$$

is bounded and such that

$$\lim_{\xi\to+\infty} F = \delta_0' u^+ + \int_0^{+\infty} \partial_h\varphi \cdot h_0' + G^+ + c,$$

where we denote $\lim_{\xi\to+\infty} G$ by G^+. Denoting by $r_i^+(t)$ and $a_i^+(t)$ eigenvectors and eigenvalues of $f'(u^+(t)) - s'(t)$, we can write

$$U_1(\xi,t) = \sum_{i=1}^{n} \alpha_i(\xi,t) r_i^+(t),$$
$$F(\xi,t) = \sum_{i=1}^{n} F_i(\xi,t) r_i^+(t).$$

Hence, we get

$$\alpha_{i\xi} = a_i^+(t)\alpha_i + F_i.$$

Consequently, we get the expression of α_i:

$$\alpha_i(\xi,t) = -\int_x^{+\infty} F_i(z,t) e^{a_i^+(t)(\xi-z)}\, dz \qquad \text{if } a_i^+(t) > 0,$$
$$\alpha_i(\xi,t) = Ce^{a_i^+(t)x} + \int_0^x F_i(z,t) e^{a_i^+(t)(\xi-z)}\, dz \qquad \text{if } a_i^+(t) < 0.$$

Hence, in both cases we find that $\alpha_i(\xi,t) = -\frac{1}{a_i}\lim_{\xi\to+\infty} F_i + \mathcal{O}(e^{-\omega\xi})$, $\omega > 0$, when $\xi \to +\infty$. Consequently, we have shown

(25)
$$\lim_{\xi\to+\infty} U_1(\xi,t) = -\Big(f'(u^+(t)) - s'(t)\Big)^{-1}\left(\delta_0'(t)u^+(t) + \int_0^{+\infty} \partial_h\varphi \cdot h_0' + G_+(t,h_0(t)) + c(t)\right).$$

Similarly, we can show

(26)
$$\lim_{\xi\to-\infty} U_1(\xi,t) = -\Big(f'(u^-(t)) - s'(t)\Big)^{-1}\left(\delta_0'(t)u^-(t) + \int_0^{-\infty} \partial_h\varphi \cdot h_0' + G^-(t,h_0(t)) + c(t)\right).$$

Going back to (20), we must also have

$$\lim_{\xi\to+\infty} U_1 = u_1^+(t) - \delta_0(t)\partial_x u^+(t), \tag{27}$$
$$\lim_{\xi\to-\infty} U_1 = u_1^-(t) - \delta_0(t)\partial_x u^-(t). \tag{28}$$

We now set $A^+(t) = f'(u^+(t)) - s'(t)$, $A^-(t) = f'(u^-(t)) - s'(t)$. By combining (25), (27), (26), (28) and eliminating $c(t)$, we get

$$\begin{aligned} A^+(t)u_1^+(t) - A^-(t)u_1^-(t) = {} & \delta_0(A^+(t)\partial_x u^+(t) - A^-(t)\partial_x u^-(t)) \\ & - \delta_0'(u^+(t) - u^-(t)) - \partial_h m \cdot h_0' - (G^+ - G^-), \end{aligned} \tag{29}$$

where

$$m(t,h_0) = \int_{-\infty}^{+\infty} (\varphi(\xi,t,h_0) - \varphi(\xi,t,0))\, d\xi.$$

Using (1) and (21), we get

$$A^+\partial_x u^+ - A^-\partial_x u^- = -\frac{d}{dt}(u^+ - u^-)$$

and

$$G^+ - G^- = -\frac{d}{dt}(u^+ - u^-) + \int_{-\infty}^{+\infty} \partial_t \varphi(\xi, t, h_0) d\xi.$$

Hence we can rewrite (29) as

$$A^+ u_1^+ - A^- u_1^- + \frac{d}{dt}\Big(\delta_0(u^+ - u^-) + m(t, h_0)\Big) = \frac{d}{dt}(u^+ - u^-). \tag{30}$$

We want to use (30) to solve (13). We first rewrite (13) as a hyperbolic initial boundary value problem. Let

$$W(z,t) = \begin{pmatrix} u_1(z + s(t), t) \\ u_1(s(t) - z, t) \end{pmatrix}$$

be the new unknown. We get that W solves

$$W_t + \big(\mathbb{A}(z,t)W\big)_z = S, \qquad z > 0, \tag{31}$$

where

$$\mathbb{A}(z,t) = \begin{pmatrix} f'(u_0(z + s(t), t)) - s'(t) & 0 \\ 0 & -(f'(u(s(t) - z, t)) - s'(t)) \end{pmatrix}$$

and

$$S = \begin{pmatrix} u_{zz}(s(t) + z, t) \\ u_{zz}(s(t) - z, t) \end{pmatrix}.$$

Note that

$$\mathbb{A}(0,t) = \begin{pmatrix} A^+(t) & 0 \\ 0 & -A^-(t) \end{pmatrix}.$$

Hence thanks to (2), (3) the boundary $\{z = 0\}$ is noncharacteristic for (31).

We first give the end of the proof in the case of Lax shocks. In this case $q = 0$, and we can forget the dependence on h_0. Consequently, (31), (30) is a linear system. We can solve it with initial conditions $W(z,0) = W_0(z)$, $\delta_0(0) = 0$, where W_0 is a smooth function which satisfies suitable compatibility conditions at 0 with the source term S to get a solution W sufficiently smooth to build the next terms of the asymptotic expansion. Using [16], it suffices to check the well-posedness for every $s \in [0, T^*]$ for the system

$$W_t + \mathbb{A}(0,s)W_z = 0, \qquad z > 0, \tag{32}$$

$$A^+(s)u_1^+(t) - A^-(s)u_1^-(t) + \delta_0'(t)(u^+(s) - u^-(s)) = 0, \tag{33}$$

where

$$W(0,t) = \begin{pmatrix} u_1^+(t) \\ u_1^-(t) \end{pmatrix}.$$

This system is similar to the one which arises in the study of linear stability of inviscid shocks [16]. The system (32), (33) is well posed if there is no nonzero solution under the form

$$W(z,t) = e^{\tau t} X(z), \qquad X(+\infty) = 0, \qquad \Re\tau > 0, \tag{34}$$

$$\delta_0(t) = e^{\tau t} d_0. \tag{35}$$

In our one-dimensional setting, this condition is equivalent to the Majda–Liu condition

$$\Delta(s) = \det\Big(r^+_{i^++1}(s), \dots, r^+_n(s), r^-_1(s), \dots, r^-_{n-i^-}(s), u^+(s) - u^-(s)\Big) \neq 0 \ \forall s \in [0, T^*].$$

Using [22, Proposition 5.3], we have

$$\tilde{D}'_s(0) = \Gamma\Delta(s).$$

Hence our hypothesis $\tilde{D}'_s(0) \neq 0$ implies the Majda–Liu condition. Consequently, using [11], [15], [16], there exists W and δ_0 which are solutions of (32), (33) defined on $[0, T^*]$. We now choose a bounded solution U_1 of (22), W and δ_0 being known. This solution satisfies the matching condition thanks to the choice of W and δ_0. This ends the proof in the case of Lax shocks.

In the case of overcompressive shocks, (30) is a nonlinear equation in h_0. Using the techniques of [16], [15], there exist W, δ_0, and h_0 solutions of (31), (30) with initial conditions $W(0,z) = W_0(z)$, $\delta_0(0) = 0$, $h_0(0) = 0$ defined on $[0,T]$ for some positive T if we have well-posedness for the linearized system

$$W_t + \mathbb{A}(0,0)W_z = 0,$$

$$A^+(0)u_1^+(t) - A^-(0)u_1^-(0) + \delta_0'(t)(u^+(0) - u^-(0)) + m_h(0,0)h_0'(t) = 0.$$

Thanks to [11], [16], this linear system is well posed if there is no nonzero solution (W, δ_0, h_0) under the form (34), (35) and

$$h_0(t) = e^{\tau t} H, \qquad h \in \mathbb{R}^q.$$

This is equivalent to

$$\Delta = \det\Big(r^+_{i^++1}(0), \dots, r^+_n(0), r^-_1(0), \dots, r^-_{n-i^-}(0), u^+(0) - u^-(0), \\ \partial_{h^1} m(0,0), \dots, \partial_{h^q} m(0,0)\Big) \neq 0.$$

Using [22, Proposition 6.2], we have

$$\tilde{D}_0^{(q+1)}(0) = \Gamma\Delta.$$

Hence $\Delta \neq 0$ since $\tilde{D}_0^{(q+1)}(0) \neq 0$. The end of the proof is similar to the case of Lax shocks.

2.2. Bounds on the error terms of the approximate solution. In the following we will denote $\varphi(\xi,t,h_0(t))$ by $V(\xi,t)$. Since we want to integrate the equation, we need to choose an approximate solution of (4) with an error term in a conservative form. If we choose an approximate solution under the basic form

$$u^{app}(x,t)=m\left(\frac{x-s(t)}{\varepsilon^\gamma}\right)I(x,t)+\left(1-m\left(\frac{x-s(t)}{\varepsilon^\gamma}\right)\right)O(x,t),$$

we find that

$$u_t^{app}+f(u^{app})_x-\varepsilon u_{xx}^{app}=q,$$

where $q(x,t)=\sum_{i=1}^3 q_i(x,t)$,

$$q_1(x,t)=(1-m)\Bigg\{\left(f(O)-f(u)-\varepsilon f'(u)u_1-\varepsilon^2 f'(u)u_2-\frac{1}{2}f''(u)(u_1,u_1)\right)_x - \varepsilon^3 u_{2xx}\Bigg\},$$

$$q_2(x,t)=m\Bigg\{\left(f(I)-f(V)-\varepsilon f'(V)V_1-\varepsilon^2 f'(V)V_2-\frac{1}{2}\varepsilon^2 f''(V)(V_1,V_1)\right)_x + \varepsilon^2\left(\delta_1'(t)\partial_\xi\partial_h\varphi+\delta_1'V_{1\xi}+\partial_t\partial_h\varphi+V_{2t}+\delta' V_{2\xi}\right)\Bigg\},$$

$$\begin{aligned}q_3(x,t)=\;&m_t\left(I-O\right)-\varepsilon m_{xx}(I-O)-2\varepsilon m_x\left(I-O\right)_x+m_x\left(f(I)-f(O)\right)\\&+\left(f(mI+(1-m)O)_x-\left(mf(I)+(1-m)f(O)\right)\right)_x.\end{aligned}$$

Consequently, the error is not in a conservative form. Hence, following the idea of [8], we choose an approximate solution in the form (10),

$$u^{app}(x,t)=m\left(\frac{x-s(t)}{\varepsilon^\gamma}\right)I(x,t)+\left(1-m\left(\frac{x-s(t)}{\varepsilon^\gamma}\right)\right)O(x,t)+d(x,t),$$

where d is such that

$$d_t-\varepsilon d_{xx}=-q(x,t),$$

$$d(0,x)=d_0(x).$$

Thanks to this choice, u^{app} now solves

$$u_t^{app}+(f'(u^{app}))_x-\varepsilon u_{xx}^{app}=(f(u^{app}-d)-f(u^{app}))_x,$$

and we will be able to integrate the equation. Let us set $R^\varepsilon(z,t)=f(\tilde{u}^{app}-\tilde{d})-f(\tilde{u}^{app})$, which is the error term of the approximate solution in (12). Recall that we have set $z=x-s(t)+\varepsilon\delta(t)$. We show the following.

PROPOSITION 3. *There exists a positive constant C independent of ε such that $\forall$ $\gamma\in(\frac{2}{3},1)$, $t\in[0,T]$,*

$$||R^\varepsilon(\cdot,t)||_{L^1}\le C\varepsilon^{3\gamma},\quad ||R_t^\varepsilon(\cdot,t)||_{L^1}\le C\varepsilon^{3\gamma-\frac{1}{2}},\quad ||R_{tt}^\varepsilon(\cdot,t)||_{L^1}\le C\varepsilon^{3\gamma-1},$$

$$||R^\varepsilon_z(\cdot,t)||_{L^1} \le C\varepsilon^{3\gamma-1}, \quad ||R^\varepsilon_{zz}(\cdot,t)||_{L^1} \le C\varepsilon^{2\gamma-\frac{1}{2}}.$$

Proof. We have a result similar to that of [8] thanks to the matching conditions

$$||\tilde q(\cdot,t)||_{L^1} + ||\tilde q(\cdot,t)_t||_{L^1} + ||\tilde q_{tt}(\cdot,t)||_{L^1} \le C\varepsilon^{3\gamma}$$

and

$$||\tilde q||_{L^\infty} + ||\tilde q_z(\cdot,t)||_{L^1} \le C\varepsilon^{2\gamma}.$$

Next since $\tilde d(z,t) = d(x+s(t)-\varepsilon\delta(t),t)$, $\tilde d$ is a solution of

$$\tilde d_t - (s'(t)-\varepsilon\delta(t))\tilde d_z - \varepsilon\tilde d_{zz} = -\tilde q(z,t), \tag{36}$$

with the initial condition $\tilde d(z,0)=0$. Consequently, we can write

$$\tilde d(z,t) = \int_0^t \int_{-\infty}^{+\infty} k^\varepsilon(z-y,t,\tau)\tilde q(y,\tau)\,dyd\tau,$$

where

$$k^\varepsilon(z,t,\tau) = \frac{1}{\sqrt{4\pi\varepsilon(t-\tau)}} \exp\left(-\frac{\left(z+\int_\tau^t s'(\mu)-\varepsilon\delta'(\mu)\,d\mu\right)^2}{4\varepsilon(t-\tau)}\right).$$

Since $||k(\cdot,t,\tau)||_{L^1} + \sqrt{\varepsilon}||k_z(\cdot,t,\tau)||_{L^1}$ is independent of ε, we get

$$||\tilde d(\cdot,t)||_{L^1} \le C\varepsilon^{3\gamma}, \qquad ||\tilde d_z(\cdot,t)||_{L^1} \le C\varepsilon^{3\gamma-\frac{1}{2}},$$

$$||\tilde d(\cdot,t)||_{L^\infty} \le C\varepsilon^{2\gamma}, \qquad ||\tilde d_z(\cdot,t)||_{L^\infty} \le C\varepsilon^{2\gamma-\frac{1}{2}}.$$

To bound $||\tilde d_t(\cdot,t)||_{L_1}$ and $||\tilde d_{tt}(\cdot,t)||_{L_1}$, we take the time derivative of (36) and get

$$(\tilde d_t)_t - (s'(t)-\varepsilon\delta(t))(\tilde d_t)_z - \varepsilon(\tilde d_t)_{zz} = \tilde q_t + (s'(t)-\varepsilon\delta(t))\tilde d_z,$$

with the initial condition $\tilde d_t(z,0) = \tilde q(z,0)$. Hence we can write

$$\begin{aligned}\tilde d_t(z,t) = &\int_{-\infty}^{+\infty} k^\varepsilon(z-y,t,0)\tilde q(y,0)\,dy \\ &+ \int_0^t\int_{-\infty}^{+\infty} k^\varepsilon(z-y,t,\tau)(\tilde q_t + (s'(\tau)-\varepsilon\delta(\tau))\tilde d_z)(y,\tau)\,dyd\tau.\end{aligned}$$

This leads to

$$||\tilde d_t(\cdot,t)||_{L^1} \le C\varepsilon^{3\gamma-\frac{1}{2}}, \qquad ||\tilde d_{tz}(\cdot,t)||_{L^1} \le C\varepsilon^{3\gamma-1}.$$

The bound for $||\tilde d_{tt}||_{L^1}$ is obtained by taking again the time derivative of the equation and using the bound on $||\tilde d_{tz}(\cdot,t)||_{L^1}$. To get the bound on $||\tilde d_{zz}||_{L^1}$, we take the derivative of (36) with respect to z. To end the proof of the proposition, we use the Taylor formula, the previous bounds, and

$$||\tilde u^{app}||_{L^\infty} + ||\tilde u^{app}_t||_{L^\infty} + ||\tilde u^{app}_{tt}||_{L^\infty} + \varepsilon||\tilde u^{app}_z||_{L^\infty} + \varepsilon^2||\tilde u^{app}_{zz}||_{L^\infty} \le C,$$

where C is independent of ε. ∎

3. The Green's function for a pure viscous profile problem. Here we recall part of the results of [21] about the behavior of the Green's function $G_\tau^S(t,x,y)$ of the operator

$$L_\tau^\varepsilon u = \partial_t u - \left(f'\left(V\left(\frac{x}{\varepsilon},\tau\right)\right) - s'(\tau)\right) u_x - \varepsilon u_{xx}.$$

The Green's function is a solution of $L_\tau^\varepsilon G_\tau^S = 0$ for $t>0$ such that

$$\lim_{t\to 0^+} G_\tau^S(t,x,y) = \delta_y(x) I_n$$

and

$$G_\tau^S(t,x,y) = 0, \qquad t<0.$$

Note that

$$G_\tau^S(t,x,y) = \frac{1}{\varepsilon} G_\tau^{HZ}\left(\frac{t}{\varepsilon},\frac{x}{\varepsilon},\frac{y}{\varepsilon}\right), \tag{37}$$

where $G_\tau^{HZ}(t,x,y)$ is the Green's function of the operator

$$L_\tau u = \partial_t u - \left(f'(V(x,\tau)) - s'(\tau)\right) u_x - u_{xx}$$

studied very precisely in [21]. We will denote by $a_j^+(\tau)$, $r_j^+(\tau)$ the eigenvalues and the associated eigenvectors of $f'(u^+(\tau)) - s'(\tau)$ and by $a_j^-(\tau)$, $r_j^-(\tau)$ the eigenvalues and the associated eigenvectors of $f'(u^-(\tau)) - s'(\tau)$. Using (37) and [21], we get the following.

THEOREM 4 (Zumbrun–Howard [21]). *Under hypothesis* (H), *we have*

$$\begin{aligned} G_\tau^S(t,x,y) = &\sum_{j,a_j^+(\tau)>0} \mathcal{O}\left(\frac{\exp\left(-\frac{\left(x-a_j^+(\tau)t\right)^2}{M\varepsilon t}\right)}{\sqrt{\varepsilon t}}\right) r_j^+(\tau)\chi_{x\geq 0} \\ &+ \sum_{j,a_j^-(\tau)<0} \mathcal{O}\left(\frac{\exp\left(-\frac{\left(x-a_j^-(\tau)t\right)^2}{M\varepsilon t}\right)}{\sqrt{\varepsilon t}}\right) r_j^-(\tau)\chi_{x\leq 0} + \mathcal{O}\left(\frac{e^{-\frac{(x-y)^2}{M\varepsilon t}}}{\sqrt{\varepsilon t}} e^{-\sigma\frac{t}{\varepsilon}}\right), \end{aligned} \tag{38}$$

$$\begin{aligned} \partial_x G_\tau^S(t,x,y) = &\sum_{j,a_j^+(\tau)>0} \mathcal{O}\left(\frac{\exp\left(-\frac{\left(x-a_j^+(\tau)t\right)^2}{M\varepsilon t}\right)}{\varepsilon t}\right) r_j^+(\tau)\chi_{x\geq 0} \\ &+ \sum_{j,a_j^-(\tau)<0} \mathcal{O}\left(\frac{\exp\left(-\frac{\left(x-a_j^-(\tau)t\right)^2}{M\varepsilon t}\right)}{\varepsilon t}\right) r_j^-(\tau)\chi_{x\leq 0} + \mathcal{O}\left(\frac{e^{-\frac{(x-y)^2}{M\varepsilon t}}}{\varepsilon t} e^{-\sigma\frac{t}{\varepsilon}}\right), \end{aligned} \tag{39}$$

where M and σ are positive constants, and

$$\chi_{z\leq 0} = \begin{cases} 1 & \text{if } z\leq 0, \\ 0 & \text{if } z>0. \end{cases}$$

All the $\mathcal{O}$'s stand for linear forms which are uniformly bounded in (t,x) and locally bounded in y.

Note that we do not use the whole expansion of the Green's function obtained in [21]. We use simplified (and weaker) bounds since we only need estimates for bounded y. Indeed the far field behavior will be handled by another Green's function in our iterative construction. Moreover, there are no excited terms thanks to the hypothesis $\mathcal{D}_\tau(0) \neq 0$ since we deal with an integrated equation. Similar weaker estimates were obtained in [9] for boundary layers.

4. The Green's function for the general operator $\mathcal{L}_\varepsilon$. The aim of this section is to prove the following theorem.

THEOREM 5. *There exists a Green's function $G(t,\tau,x,y)$ of $\mathcal{L}_\varepsilon$ defined for $0 \leq \tau,\ t \leq T,\ z,\ y \in \mathbb{R}$ such that $G(t,\tau,z,y) = 0$ if $t < \tau$ and*

$$\sup_{0\leq\tau\leq T,y} \int_0^T \int_{\mathbb{R}} |G(t,\tau,z,y)|\,dzdt + \sqrt{\varepsilon} \sup_{0\leq\tau\leq T,y} \int_0^T \int_{\mathbb{R}} |\partial_z G(t,\tau,z,y)|\,dzdt \tag{40}$$

is bounded uniformly in ε.

Proof. We use the iterative construction of the Green's functions given in [9]. We choose an approximate Green's function $G^{app}(t,\tau,z,y)$ under the form

$$G^{app}(t,\tau,z,y) = \sum_{k=1}^{N} S_k(t,\tau,z,y)\Pi_k(\tau,y),$$

where $S_k(t,\tau,z,y)$ are Green's kernels and $\Pi_k \in \mathcal{C}^\infty([0,T]\times\mathbb{R},\mathcal{L}(\mathbb{R}^n))$ are such that

$$||\Pi_k(t,x)v|| \leq C||v|| \qquad \forall x \geq 0,\ t \in [0,T],\ v \in \mathbb{R}^n$$

and

$$\sum_k \Pi_k = \mathrm{Id}.$$

For each S_k, we define the error $R_k(t,\tau,z,y) = \mathcal{L}_\varepsilon S_k$. We then define the matrix of errors $\mathcal{M}(T_1,T_2) = (\sigma_{kl}(T_1,T_2))_{1\leq k,l\leq N}$, where

$$\sigma_{kl}(T_1,T_2) = \sup_{T_1\leq\tau\leq T_2,\ y\in \mathrm{supp}\ \Pi_l} \int_{T_1}^{T_2} \int_{\mathbb{R}} |\Pi_k(t,z)R_l(t,\tau,z,y)|\,dzdt.$$

This matrix describes how each part of the approximate Green's function is handled at the next step of the iterative method. Theorem 5 of [9] states that to prove (40), it suffices to check that there exists η such that $T_2 - T_1 \leq \eta$ implies

$$\lim_{p\to+\infty} \mathcal{M}^p(T_1,T_2) = 0.$$

Let us now introduce some definitions that are necessary for the construction of our approximate Green's function. We use two smooth cut-off functions χ^+ and χ^- such that

$$\chi^+(z) = \begin{cases} 0 & \text{if } z \leq 1, \\ 1 & \text{if } z \geq 2 \end{cases} \qquad \text{and} \qquad \chi^-(z) = \begin{cases} 0 & \text{if } z \geq -1, \\ 1 & \text{if } z \leq -2. \end{cases}$$

We also assume that the cut-off function m already used was under the form $(1-\chi^+)(1-\chi^-)$. We denote by $P^\pm(t,z)$, $D^\pm(t,z)$ matrices such that

$$f'(u(z+s(t),t))-s'(t)=P^+D^+(P^+)^{-1}\qquad\forall z>0,$$

$$f'(u(z+s(t),t))-s'(t)=P^-D^-(P^-)^{-1}\qquad\forall z<0,$$

and $D^\pm(\pm z,t)=\operatorname{diag}(\lambda_1(u(s(t)\pm z,t))-s'(t),\dots,\lambda_n(u(s(t)\pm z,t))-s'(t))$ if $z>0$. Setting

$$\lambda_i^+(z,t)=\begin{cases}\lambda_i(u(z+s(t),t))-s'(t) & \text{if } z>0,\\ \lambda_i(u(s(t)+0,t))-s'(t) & \text{if } z\le 0\end{cases}$$

and

$$\lambda_i^-(z,t)=\begin{cases}\lambda_i(u(z+s(t),t))-s'(t) & \text{if } z<0,\\ \lambda_i(u(s(t)-0,t))-s'(t) & \text{if } z\ge 0,\end{cases}$$

we define the characteristic curves $X_i^\pm(t,\tau,y)$ by

$$\partial_t X_i^\pm(t,\tau,y)=\lambda_i^\pm(X_i^\pm(t,\tau,y),t),\qquad t\ge\tau,$$

with initial data $X_i^\pm(\tau,\tau,y)=y$. As in [9], we first make a stronger hypothesis than (2), (3) and assume that

$$|\lambda_k(z,t)|\ge C\qquad\forall z,t. \tag{41}$$

This hypothesis will be removed in the last part of the proof of the theorem by a localization argument.

We also define the projections

$$\begin{aligned}\mathcal{P}^+_{out}(t,z)&=P^+(t,z)D^+_{out}(t,z)(P^+)^{-1}(t,z),\\ \mathcal{P}^+_{in}(t,z)&=P^+(t,z)D^+_{in}(t,z)(P^+)^{-1}(t,z),\\ \mathcal{P}^-_{out}(t,z)&=P^-(t,z)D^-_{out}(t,z)(P^-)^{-1}(t,z),\\ \mathcal{P}^-_{in}(t,z)&=P^-(t,z)D^-_{in}(t,z)(P^-)^{-1}(t,z),\end{aligned}$$

where

$$\begin{aligned}D^+_{out}&=\operatorname{diag}(0,\dots,0,1,\dots 1),\ \text{with } p+1 \text{ null coefficients},\\ D^+_{in}&=\operatorname{diag}(1,\dots,1,0,\dots,0),\ \text{with } p \text{ unit coefficients},\\ D^-_{out}&=\operatorname{diag}(1,\dots,1,0,\dots,0),\ \text{with } p-1 \text{ unit coefficients},\\ D^-_{in}&=\operatorname{diag}(0,\dots,0,1,\dots,1),\ \text{with } p \text{ unit coefficients}.\end{aligned}$$

Let

$$G_T^\pm(t,\tau,z,y)=\operatorname{diag}\left(\frac{e^{-\frac{(z-X_i^\pm(t,\tau,y))^2}{4\varepsilon(t-\tau)}}}{\sqrt{4\pi\varepsilon(t-\tau)}}\right).$$

We define the Green's functions for the incoming and outgoing waves as

$$G^+_{out}(t,\tau,z,y) = \chi^+\left(\frac{z}{M_1\varepsilon}\right) P^+(t,z)D^+_{out}(t,z)G_T(t,\tau,z,y)(P^+(\tau,y))^{-1} = \chi^+\tilde{G}^+_{out},$$
$$G^+_{in}(t,\tau,z,y) = \chi^+\left(\frac{z}{M_1\varepsilon}\right) P^+(t,z)D^+_{in}(t,z)G_T(t,\tau,z,y)(P^+(\tau,y))^{-1} = \chi^+\tilde{G}^+_{in},$$
$$G^-_{out}(t,\tau,z,y) = \chi^-\left(\frac{z}{M_1\varepsilon}\right) P^-(t,z)D^-_{out}(t,z)G_T(t,\tau,z,y)(P^-(\tau,y))^{-1} = \chi^-\tilde{G}^-_{out},$$
$$G^-_{in}(t,\tau,z,y) = \chi^-\left(\frac{z}{M_1\varepsilon}\right) P^-(t,z)D^-_{in}(t,z)G_T(t,\tau,z,y)(P^-(\tau,y))^{-1} = \chi^-\tilde{G}^-_{in},$$

where $M_1 > 0$ is to be chosen. We also define

$$G^{shock}(t,\tau,z,y) = m\left(\frac{z}{M_3\varepsilon}\right) G^S_\tau(t-\tau,z,y),$$

where G^S_τ was defined in section 3. The kernels of the theorem of [9] will be $S_1 = S_2 = G^-_{out}$, $S_3 = G^-_{in}$, $S_4 = G^{shock}$, $S_5 = G^+_{in}$, $S_6 = S_7 = G^+_{out}$. The truncation functions will be

$$\Pi_1(\tau,y) = \chi^-\left(\frac{y}{M_2\varepsilon}\right)\left(1-\chi^-\left(\frac{2y}{M_3\varepsilon}\right)\right)\mathcal{P}^-_{out}(\tau,y),$$
$$\Pi_2(\tau,y) = \chi^-\left(\frac{2y}{M_3\varepsilon}\right)\mathcal{P}^-_{out}(\tau,y),$$
$$\Pi_3(\tau,y) = \chi^-\left(\frac{y}{M_2\varepsilon}\right)\mathcal{P}^-_{in}(\tau,y),$$
$$\Pi_4(\tau,y) = m\left(\frac{y}{M_2\varepsilon}\right),$$
$$\Pi_5(\tau,y) = \chi^+\left(\frac{y}{M_2\varepsilon}\right)\mathcal{P}^+_{in}(\tau,y),$$
$$\Pi_6(\tau,y) = \chi^+\left(\frac{2y}{M_3\varepsilon}\right)\mathcal{P}^+_{out}(\tau,y),$$
$$\Pi_7(\tau,y) = \chi^+\left(\frac{y}{M_2\varepsilon}\right)\left(1-\chi^+\left(\frac{2y}{M_3\varepsilon}\right)\right)\mathcal{P}^+_{out}(\tau,y).$$

The constants M_1, M_2, and M_3 are such that $M_1 \le 4M_2 \le 16M_3$ and will be carefully chosen at the end of the proof. G_1 and G_5 describe the creation and propagation of outgoing waves in a vicinity of the shock layer, G_2 and G_6 describe the creation and propagation of outgoing waves away from the shock layer, G_3 and G_7 describe the creation and propagation of ingoing waves, and G_4 describes the dynamics of the shock layer. Note that with this choice we get a relevant approximate Green's function since $G^{app}(\tau,\tau,z,y) = \delta_y(z)I_n$. Moreover G^{app} satisfies the estimate (40); hence we can use Theorem 5 of [9]. We have to compute $\mathcal{M}$, which is the aim of the following lemmas. For any error term $E(t,\tau,z,y)$, we use the notation

$$||E|| = \sup_{0\le\tau\le T, y\in\mathbb{R}} \int_\tau^T \int_{-\infty}^{+\infty} |E(t,\tau,z,y)|\, dzdt.$$

The following lemma of [9] is crucial for the estimations of the error terms. Consequently, we recall it for the sake of completeness.

4.1. A technical lemma. Let us define for some trajectory $X(t)$ such that $X(0)=0$

$$I(y)=\int_0^{+\infty}\int_0^{+\infty}\exp\Big(-\frac{(x-y-X(t))^2}{t}\Big)\exp(-\sigma x)\frac{dxdt}{\sqrt{t}},$$

$$J(y)=\int_0^{+\infty}\int_0^{+\infty}\exp\Big(-\frac{(x-y-X(t))^2}{t}\Big)\exp(-\sigma x)\frac{dxdt}{t},$$

$$K(y)=\frac{1}{M}\int_0^{+\infty}\int_M^{2M}\exp\Big(-\frac{(x-y-X(t))^2}{t}\Big)\frac{dxdt}{\sqrt{t}},$$

$$L(y)=\frac{1}{M}\int_0^{+\infty}\int_M^{2M}\exp\Big(-\frac{(x-y-X(t))^2}{t}\Big)\frac{dxdt}{t},$$

LEMMA 6 (see [9]).

(i) *If $\gamma\geq X'(t)\geq\delta>0$, then $I(y)$ and $J(y)$ are bounded uniformly in $y\geq 0$ and go to 0 as $y\to+\infty$.*

(ii) *If $-\gamma\leq X'(t)\leq-\delta<0$, then $I(y)$ and $J(y)$ are bounded uniformly in $y\geq 0$.*

(iii) *If $\gamma\geq X'(t)\geq\delta>0$, then $K(y)$ and $L(y)$ are bounded uniformly in $y\geq 0$ and M and go to 0 as $y-2M$ goes to $+\infty$.*

(iv) *If $-\gamma\leq X'(t)\leq-\delta<0$, then $K(y)$ and $L(y)$ are bounded uniformly in $y\geq 0$ and M.*

Note that as $xe^{-x^2}\leq Ce^{-x^2/2}$ the same result is true if we replace $J(y)$ by

$$\int_0^{+\infty}\int_0^{+\infty}\frac{|x-y-X(t)|}{t}\exp\Big(-\frac{(x-y-X(t))^2}{t}\Big)\exp(-\sigma x)\frac{dxdt}{\sqrt{t}}$$

and $L(y)$ by

$$\frac{1}{M}\int_0^{+\infty}\int_M^{2M}\frac{|x-y-X(t)|}{t}\exp\Big(-\frac{(x-y-X(t))^2}{t}\Big)\frac{dxdt}{\sqrt{t}}.$$

Moreover if $M\geq 0$,

$$\int_0^{+\infty}\int_M^{+\infty}\exp\Big(-\frac{(x-y-X(t))^2}{t}\Big)\exp(-\sigma x)\frac{dxdt}{\sqrt{t}}\leq e^{-\sigma M}I(y-M)$$

and therefore goes to 0 as $M\to+\infty$, provided that $y\geq M$, and similarly for J.

Note also that we have opposite results according to the direction of the transport if we study the integrals for x varying in $(-\infty,0)$ or in $(-2M,-M)$ when y goes to $-\infty$.

4.2. Bounds on the error terms of the shock Green's function. Call $E^S(t,\tau,z,y)$ this error term; we have

$$E^S(t,\tau,z,y)=\mathcal{L}_\varepsilon\left(m\left(\frac{z}{M_3\varepsilon}\right)G^S_\tau(t-\tau,z,y)\right).$$

As in [9], we split it into an evolution error

$$E_1^S = m\left(\frac{z}{M_3\varepsilon}\right)\mathcal{L}_\varepsilon G_\tau^S$$

and a truncation error

$$E_2^S = \left(\frac{1}{M_3\varepsilon}m'\left(\frac{z}{M_3\varepsilon}\right)(f'(u^{app}) - s'(t) + \varepsilon\delta'(t)) - \frac{1}{M_3^2\varepsilon}m''\left(\frac{z}{M_3\varepsilon}\right)\right)G_\tau^S - \frac{2}{M_3}m'\left(\frac{z}{M_3\varepsilon}\right)\partial_x G_\tau^S. \tag{42}$$

We then split E_2^S into four parts: the ingoing parts at the right- and left-hand sides of the shock and the outgoing part at the right- and left-hand sides of the shock

$$E_{2in}^{S+} = \chi_{z\geq 0}\mathcal{P}_{in}^+ E_2^S, \qquad E_{2out}^{S+} = \chi_{z\geq 0}\mathcal{P}_{out}^+ E_2^S,$$

$$E_{2in}^{S-} = \chi_{z\leq 0}\mathcal{P}_{in}^- E_2^S, \qquad E_{2out}^{S-} = \chi_{z\leq 0}\mathcal{P}_{out}^- E_2^S.$$

Estimates on these terms are given by the following lemma.

LEMMA 7. *We have*

(i) $||1_{|y|\leq 2M_2\varepsilon}E_1^S(t,\tau,z,y)||_{L^\infty_{\tau,y},L^1_{t,z}} \leq C_1(T+\varepsilon^{2\gamma-1})$,

(ii) $||1_{|y|\leq 2M_2\varepsilon}E_{2out}^{S+}(t,\tau,z,y)||_{L^\infty_{\tau,y},L^1_{t,z}} + ||1_{|y|\leq 2M_2\varepsilon}E_{2out}^{S-}(t,\tau,z,y)||_{L^\infty_{\tau,y},L^1_{t,z}} \leq C_2$,

(iii) $||1_{|y|\leq 2M_2\varepsilon}E_{2in}^{S+}(t,\tau,z,y)||_{L^\infty_{\tau,y},L^1_{t,z}} + ||1_{|y|\leq 2M_2\varepsilon}E_{2in}^{S-}(t,\tau,z,y)||_{L^\infty_{\tau,y},L^1_{t,z}} \leq C_3 + C_4 T$,

where C_1 is locally bounded in M_2 and M_3, C_2 and C_4 are locally bounded in M_2 (uniformly in M_3), and C_3, which depends on M_2 and M_3, goes to 0 as $M_3 \to +\infty$.

Note that $2\gamma - 1 > 0$.

Proof. We give only the proof for the "+" terms, that is, estimates for $z \geq 0$, since the proof for the other side is completely symmetric. Let us begin with the evolution error. We have

$$E_1^S = m\left(\frac{z}{M_3\varepsilon}\right)\left(f'\left((1-m)\left(\frac{z-\varepsilon\delta(t)}{\varepsilon^\gamma}\right)O(z+s(t)-\varepsilon\delta(t),t) + m\left(\frac{z-\varepsilon\delta(t)}{\varepsilon^\gamma}\right)I\left(\frac{z}{\varepsilon},t\right) + \tilde{d}(z,t)\right) - f'\left(V\left(\frac{z}{\varepsilon},\tau\right)\right) + (s'(t)-s'(\tau)) + \varepsilon\delta(t)\right)\partial_z G_\tau^S.$$

Hence

$$|E_1^S| \leq m\left(\frac{z}{M_3\varepsilon}\right)(|t-\tau| + \varepsilon^\gamma)|\partial_z G_\tau^S|$$

since for ε sufficiently small, we have $m(\frac{z}{M_3\varepsilon})(1-m)(\frac{z-\varepsilon\delta(t)}{\varepsilon^\gamma}) = 0\ \forall t \in [0,T]$. It leads

to estimate

$$\alpha_{1,j} = \int_\tau^T \int_0^{2M_3\varepsilon} \frac{1}{\varepsilon} e^{-\frac{(z-a_j^+(\tau)(t-\tau))^2}{\varepsilon(t-\tau)}}\, dzdt,$$
$$\alpha_{1,r} = \int_\tau^T \int_0^{2M_3\varepsilon} e^{-\sigma\frac{t-\tau}{\varepsilon}} e^{-\frac{(z-y)^2}{\varepsilon(t-\tau)}}\, dzdt,$$
$$\tilde{\alpha}_{1,j} = \varepsilon^\gamma \int_\tau^T \int_0^{2M_3} \frac{e^{-\frac{(z-a_j^+(\tau)(t-\tau))^2}{\varepsilon(t-\tau)}}}{\varepsilon(t-\tau)}\, dzdt,$$
$$\tilde{\alpha}_{1,r} = \varepsilon^\gamma \int_\tau^T \int_0^{2M_3\varepsilon} \frac{e^{-\sigma\frac{t-\tau}{\varepsilon}} e^{-\frac{(z-y)^2}{\varepsilon(t-\tau)}}}{\varepsilon(t-\tau)}\, dzdt.$$

The integrals $\alpha_{1,j}$ and $\alpha_{1,r}$ are obviously bounded by $2M_3T$. To bound $\tilde{\alpha}_{1,j}$, we set $\tilde{z} = \frac{z-\alpha_j^+(t-\tau)}{\sqrt{\varepsilon(t-\tau)}}$, giving an $\mathcal{O}(\varepsilon^{\gamma-\frac{1}{2}}\sqrt{T})$ bound. We easily get a similar bound for $\tilde{\alpha}_{1,r}$. Hence (i) is proved.

We now prove (ii). Using (42) and estimates on G_τ^S and $\partial_x G_\tau^S$, given in Theorem 5, we have to bound

$$\beta_{1,j} = \frac{1}{M_3\varepsilon} \int_\tau^T \int_{M_3\varepsilon}^{2M_3\varepsilon} \frac{e^{-\frac{(z-a_j^+(\tau)(t-\tau))^2}{\varepsilon(t-\tau)}}}{\sqrt{\varepsilon(t-\tau)}}\, dzdt,$$
$$\beta_{2,j} = \frac{1}{M_3\varepsilon} \int_\tau^T \int_{M_3\varepsilon}^{2M_3\varepsilon} \frac{e^{-\frac{(z-a_j^+(\tau)(t-\tau))^2}{\varepsilon(t-\tau)}}}{\varepsilon(t-\tau)}\, dzdt,$$
$$\beta_{1,r} = \frac{1}{M_3} \int_\tau^T \int_{M_3\varepsilon}^{2M_3\varepsilon} \frac{e^{-\frac{(z-y)^2}{\varepsilon(t-\tau)}}}{\sqrt{\varepsilon(t-\tau)}} e^{-\frac{\sigma(t-\tau)}{\varepsilon}}\, dzdt,$$
$$\beta_{2,r} = \frac{1}{M_3} \int_\tau^T \int_{M_3\varepsilon}^{2M_3\varepsilon} \frac{e^{-\frac{(z-y)^2}{\varepsilon(t-\tau)}}}{\varepsilon(t-\tau)} e^{-\frac{\sigma(t-\tau)}{\varepsilon}}\, dzdt,$$

To bound $\beta_{1,j}$ and $\beta_{2,j}$, we set $\tilde{z} = \frac{z}{\varepsilon}$ and $s = \frac{t-\tau}{\varepsilon}$ and use Lemma 6. To bound $\beta_{1,r}$ and $\beta_{2,r}$, we set $\tilde{z} = \frac{z-y}{\sqrt{\varepsilon(t-\tau)}}$ and then set $s = \frac{t-\tau}{\varepsilon}$. We get $\mathcal{O}(\frac{1}{M_3})$ bounds.

It remains to prove (iii). Thanks to Theorem 4, we can write for $z \geq 0$

$$G_\tau^S(t-\tau, z, y) = \mathcal{P}_{out}^+(\tau, 0)\tilde{G}(t, \tau, z, y) + R(t, \tau, z, y),$$

where $\tilde{G}$ is bounded by Gaussians which travel at speeds $a_j^+(\tau, 0)$ and R is the residual term bounded by $\frac{1}{\varepsilon(t-\tau)} e^{-\frac{(z-y)^2}{M\varepsilon(t-\tau)}} e^{-\sigma\frac{t-\tau}{\varepsilon}}$. Consequently, we get

$$|\mathcal{P}_{in}^+(t, z) G_\tau^S| \leq C\left((|t-\tau| + z)\,|\tilde{G}| + |R|\right).$$

Similarly, since

$$m'\left(\frac{z}{M_3\varepsilon}\right) f'(u^{app}) = m'\left(\frac{z}{M_3\varepsilon}\right) f'\left(V\left(\frac{z}{\varepsilon}, t\right) + \mathcal{O}(\varepsilon^\gamma)\right)$$
$$= m'\left(\frac{z}{M_3\varepsilon}\right) f'(u(z + s(t), t)) + \mathcal{O}\left(e^{-\alpha\frac{z}{\varepsilon}} + \varepsilon^\gamma\right),$$

we get

$$\left| m'\left(\frac{z}{M_3\varepsilon}\right)\mathcal{P}^+_{in}(t,z)f'(u^{app})G^S_\tau \right| \leq \left(|t-\tau|+|z|+e^{-\alpha\frac{z}{\varepsilon}}+\varepsilon^\gamma\right)|\tilde{G}|+|R|.$$

Using also the polarization of $\partial_z G^S_\tau$ given in Theorem 2, we get

$$|\mathcal{P}^+_{in}(t,z)\partial_z G^S_\tau| \leq C((|t-\tau|+z+\varepsilon^\gamma)|\tilde{G}|+|R|).$$

Hence E^{S+}_{2in} is bounded by terms like $\beta_{1,r}$, $\beta_{2,r}$, $\alpha_{1,j}$, $\alpha_{1,r}$ which have already been bounded and by

$$\gamma_{1,j} = \frac{1}{M_3\varepsilon}\int_\tau^T\int_{M_3\varepsilon}^{2M_3\varepsilon} x\frac{e^{-\frac{(z-a^+_j(\tau)(t-\tau))^2}{\varepsilon(t-\tau)}}}{\sqrt{\varepsilon(t-\tau)}}\,dzdt,$$

$$\gamma_{2,j} = \frac{1}{M_3}\int_\tau^T\int_{M_3\varepsilon}^{2M_3\varepsilon} x\frac{e^{-\frac{(z-a^+_j(\tau)(t-\tau))^2}{\varepsilon(t-\tau)}}}{\varepsilon(t-\tau)}\,dzdt,$$

$$\gamma_{3,j} = \frac{1}{M_3\varepsilon}\int_\tau^T\int_{M_3\varepsilon}^{2M_3\varepsilon} e^{-\alpha\frac{z}{\varepsilon}}\frac{e^{-\frac{(z-\lambda_j(t-\tau))^2}{\varepsilon(t-\tau)}}}{\sqrt{\varepsilon(t-\tau)}}\,dzdt.$$

We have

$$\gamma_{1,j} \leq 2\int_\tau^T\int_0^{+\infty}\frac{e^{-\frac{(z-a^+_j(\tau)(t-\tau))^2}{\varepsilon(t-\tau)}}}{\sqrt{\varepsilon(t-\tau)}}\,dzdt \leq 2\sqrt{\pi}T.$$

Similarly, we have the estimate $\gamma_{2,j} \leq 2\sqrt{\pi}\sqrt{\varepsilon T}$. Finally $\gamma_{3,j} \leq e^{-\alpha M_3}\beta_{1,j}$, which ends the proof since $\beta_{1,j}$ is uniformly bounded in M_3. □

4.3. Bounds on the error terms of $G^\pm_{out}$. In this section, we compute and estimate the error terms of G^+_{out} and G^-_{out}. These error terms are respectively denoted by R^+_{out} and R^-_{out}. As usual we write $R^\pm_{out} = E^\pm_{1out} + E^\pm_{2out}$, where $E^\pm_{1out} = \chi_\pm\left(\frac{z}{M\varepsilon}\right)\mathcal{L}^\varepsilon G^\pm_{out}$ is the evolution error and $E^\pm_{2out}$ is the truncation error.

LEMMA 8. *Let $M \geq M_2$. We have*

$$||1_{y\geq M\varepsilon}E^+_{1out}(t,\tau,z,y)|| + ||1_{y\leq -M\varepsilon}E^-_{1out}(t,\tau,z,y)|| \leq C_5(T+\varepsilon^{2\gamma-1})+C_6,$$

where C_5 is independent of M_1, M_2, and M_3, and where C_6, which depends only on M, goes to 0 as $M \to +\infty$. Moreover,

$$||1_{y\geq M\varepsilon}E^+_{2out}(t,\tau,z,y)|| + ||1_{y\leq -M\varepsilon}E^-_{2out}(t,\tau,z,y)|| \leq C_7(M)$$

with

$$\lim_{M\to+\infty} C_7(M) = 0.$$

Note that since $\gamma > \frac{2}{3}$, we have a good bound in ε for $||E^\pm_{1out}||$.

Proof. We give only the outlines of the proof since it is almost the same as in [9, section 5.3]. Note that here *in* and *out* stand for the direction of the transport according to the shock, whereas in [9] they stand for the direction of the transport according to the domain. Let us consider, for example, the "+" term. The proof

for the "−" term is symmetric. Compared to [9], the only terms that are different are those involving $f'(u^{app})$. More precisely, as $f'(u^{app})$ is uniformly bounded in ε as $A^{int}(t,x) + A^b\left(t, \frac{x}{\varepsilon}\right)$ was in [9], the truncation error is already studied in [9]. The only new term which has to be studied is

$$\chi^+\left(\frac{z}{M_1\varepsilon}\right)(f'(u^{app}) - f'(u(z+s(t),t)))P^+D^+_{out}\partial_z G_T(P^+)^{-1}.$$

Note that for $z \geq M_1\varepsilon$,

$$\begin{aligned}
|f'(u^{app}) - f'(u(z+s(t),t)| &\leq Cm\Big(\frac{z-\varepsilon\delta(t)}{\varepsilon^\gamma}\Big)\Big(\Big|V\Big(\frac{z}{\varepsilon},t\Big) - u(z+s(t),t)\Big| + \mathcal{O}(\varepsilon^\gamma)\Big)\\
&\leq C\Big(e^{-\alpha\frac{z}{\varepsilon}} + \varepsilon^\gamma\\
&\quad + m\Big(\frac{z-\varepsilon\delta(t)}{\varepsilon^\gamma}\Big)|u(z+s(t),t) - u(s(t)+0,t)|\Big)\\
&\leq C\left(e^{-\alpha\frac{z}{\varepsilon}} + \varepsilon^\gamma\right)
\end{aligned}$$

since $m(\frac{z-\varepsilon\delta(t)}{\varepsilon^\gamma})$ vanishes for $z \geq 2\varepsilon^\gamma + \varepsilon\delta(t)$. The term involving $e^{-\alpha\frac{z}{\varepsilon}}$ is already bounded in [9] thanks to Lemma 6. Consequently, it remains only to bound

$$\delta_j = \varepsilon^\gamma \int_\tau^T \int_{M_1\varepsilon}^{+\infty} \frac{|z - X_j^+|}{(\varepsilon(t-\tau))^{\frac{3}{2}}} e^{-\frac{(z-X_j^+)^2}{\varepsilon(t-\tau)}}\, dzdt.$$

Setting $\tilde{z} = \frac{x - X_j^+}{\sqrt{\varepsilon(t-\tau)}}$, we get

$$\delta_j \leq C\varepsilon^{\gamma-\frac{1}{2}}\sqrt{T} \leq C(T + \varepsilon^{2\gamma-1}). \qquad \square$$

4.4. Bounds on the error terms of G^{in}. In this section, we estimate the error terms of G^+_{in}, and G^-_{in}. These terms are respectively denoted by R^+_{in} and R^-_{in}. As usual we write $R^\pm_{in} = E^\pm_{1in} + E^\pm_{2in}$, where $E^\pm_{1in} = \chi_\pm\left(\frac{z}{M\varepsilon}\right)\mathcal{L}^\varepsilon\tilde{G}^\pm_{in}$ is the evolution error and $E^\pm_{2in}$ is the truncation error.

LEMMA 9. *We have*

$$||1_{y\geq M_2\varepsilon}E^+_{1in}|| + ||1_{y\leq -M_2\varepsilon}E^-_{1in}|| \leq C_8(T + \varepsilon^{2\gamma-1}) + C_9,$$

where C_8 is locally bounded in M_1 and $C_9 \to 0$ when $M_1 \to +\infty$. Moreover

$$||1_{y\geq M_2\varepsilon}E^+_{2in}|| + ||1_{y\leq -M_2\varepsilon}E^-_{2in}|| \leq C_{10},$$

where C_{10} is bounded uniformly in M_1.

We do not give the proof of this lemma since it is very similar to the proof of the corresponding lemma in [9]. The minor changes that we have to bring in the proof of [9] are already explained in the previous section. We point out only that the estimate on the truncation error is worse than in the previous section. It relies on the different estimates on $K(y)$ and $L(y)$ according to the direction of the transport in Lemma 6.

4.5. End of the proof with (41). Combining all the previous estimates, we can write the matrix of errors $\mathcal{M}(0,T)$ as

$$M = \begin{pmatrix} \eta(M_2) & \eta(M_3) & \eta(M_1) & \eta(M_3) & 0 & 0 & 0 \\ \eta(M_2) & \eta(M_3) & \eta(M_1) & C & 0 & 0 & 0 \\ \eta(M_2) & \eta(M_3) & \eta(M_1) & \eta(M_3) & 0 & 0 & 0 \\ \eta(M_2) & \eta(M_3) & C & \eta(M_3) & C & \eta(M_3) & \eta(M_2) \\ 0 & 0 & 0 & \eta(M_3) & \eta(M_1) & \eta(M_3) & \eta(M_2) \\ 0 & 0 & 0 & C & \eta(M_1) & \eta(M_3) & \eta(M_2) \\ 0 & 0 & 0 & \eta(M_3) & \eta(M_1) & \eta(M_3) & \eta(M_2) \end{pmatrix},$$

where "$\eta(M_1)$" stands for the coefficients which go to 0 as $M_1 \to +\infty$ (independently of M_2 and M_3), $T \to 0$ and $\varepsilon \to 0$, "$\eta(M_2)$" stands for coefficients which go to 0 as $M_2 \to +\infty$ (M_1 being fixed, but independently of M_3), $T \to 0$ and $\varepsilon \to 0$, and "$\eta(M_3)$" stands for the coefficients which go to 0 as $M_3 \to +\infty$ (M_1 and M_2 being fixed), $T \to 0$ and $\varepsilon \to 0$, and where C is a constant depending on M_1, M_2 (but independent of M_3). Next we conclude as in [9]. Let $\alpha < \frac{1}{1000}$, we first fix M_1 and T_1, ε_1 such that for $T \le T_1$, $\varepsilon \le \varepsilon_1$ all the $\eta(M_1)$ are smaller than α. Next we fix M_2, and we can reduce T_1 and ε_1 such that all the $\eta(M_2)$ are smaller than α. Finally, by taking M_3 sufficiently large, and by reducing T_1 and ε_1 we can make $\eta(M_3)$ arbitrarily small. After the choice of M_1, M_2, T_1, and ε_1 the constants are fixed. Using a perturbation argument, we have to consider only the powers of

$$\tilde{\mathcal{M}} = \begin{pmatrix} \alpha & 0 & \alpha & 0 & 0 & 0 & 0 \\ \alpha & 0 & \alpha & C & 0 & 0 & 0 \\ \alpha & 0 & \alpha & 0 & 0 & 0 & 0 \\ \alpha & 0 & C & 0 & C & 0 & \alpha \\ 0 & 0 & 0 & 0 & \alpha & 0 & \alpha \\ 0 & 0 & 0 & C & \alpha & 0 & \alpha \\ 0 & 0 & 0 & 0 & \alpha & 0 & \alpha \end{pmatrix}.$$

Since the eigenvalues of $\mathcal{M}$ are 0 and 2α, the theorem with assumption (41) is proved.

4.6. Without (41). In this section we only assume (2), (3). By continuity, inequalities (2), (3) are still true for $-4\eta \le z \le 4\eta$. We again use Theorem 5 of [9] with

$$\Pi_1 = \chi^+\left(\frac{y}{\eta}\right), \quad \Pi_2 = \chi^-\left(\frac{y}{\eta}\right), \quad \Pi_3 = m\left(\frac{y}{\eta}\right),$$

and

$$S_1 = \chi^+\left(\frac{4z}{\eta}\right) G_T^+, \qquad S_2 = \chi^-\left(\frac{4z}{\eta}\right) G_T^-,$$

$$S_3 = m\left(\frac{z}{2\eta}\right) G,$$

where G is the Green's function that was constructed in the previous section. Since the errors corresponding to each S_i are as small as we want if we choose T and ε small (T being independent of ε), Theorem 5 of [9] allows us to conclude.

5. Convergence of $\tilde{w}$. In this section we prove Theorem 2. To prove the convergence of $\tilde{w}$ to zero, we use a standard argument for parabolic equations as in [10], [9]. Local existence of a smooth solution for (12) with initial condition $\tilde{w}(z,0)=0$ is classical; hence we define

$$T^\varepsilon = \sup\left\{T_1 \in [0,T],\ \exists \tilde{w} \text{ solution on } \mathbb{R}\times[0,T_1),\ E(T_1)\le 1\right\}, \tag{43}$$

where

$$E(T_1)=\int_0^{T_1}\int_{-\infty}^{+\infty} \frac{|\tilde{w}|}{\varepsilon^{3\gamma-\alpha}}+\frac{|\tilde{w}_z|}{\varepsilon^{3\gamma-\alpha-\frac{1}{2}}}+\frac{|\tilde{w}_t|}{\varepsilon^{3\gamma-2\alpha-\frac{1}{2}}}+\frac{|\tilde{w}_{tz}|}{\varepsilon^{3\gamma-2\alpha-1}} + \frac{|\tilde{w}_{zz}|}{\varepsilon^{3\gamma-3\alpha-\frac{3}{2}}}+\frac{|\tilde{w}_{tt}|}{\varepsilon^{3\gamma-3\alpha-1}}+\frac{|\tilde{w}_{ttz}|}{\varepsilon^{3\gamma-3\alpha-\frac{3}{2}}}+\frac{|\tilde{w}_{tzz}|}{\varepsilon^{3\gamma-4\alpha-2}}\,dzdt;$$

the small positive constant α will be carefully chosen in the following. Note that we have

$$||\tilde{w}_z||_{L^\infty(\mathbb{R}\times[0,T_1])} \le ||\tilde{w}_{tzz}||_{L^1(\mathbb{R}\times[0,T_1])} \le \varepsilon^{3\gamma-4\alpha-2} \le 1 \tag{44}$$

as soon as $\varepsilon \le 1$ if we choose $\gamma \in (\frac{2}{3},1)$, and $\alpha>0$ so small that $3\gamma-4\alpha-2>0$. There are two possibilities:

(i) $T^\varepsilon = T$,

(ii) $T^\varepsilon < T$ and $E(T^\varepsilon)=1$.

Let us assume that we are in the second case. From now on, we will denote by C a generic number which may depend on T but which is independent of ε. Moreover, until the end of the proof, we set

$$||\tilde{w}|| = ||\tilde{w}||_{L^1(\mathbb{R}\times[0,T^\varepsilon])}, \qquad ||\tilde{w}||_\infty = ||\tilde{w}||_{L^\infty(\mathbb{R}\times[0,T^\varepsilon])}.$$

Since $\tilde{w}$ is a solution of (12), we have for all $t\in[0,T^\varepsilon)$,

$$\tilde{w}(z,t)=\int_0^t\int_{-\infty}^{+\infty} G(t,\tau,z,y)(R^\varepsilon(z,\tau)+Q(\tilde{u}^{app},\tilde{w}_z))\,dyd\tau, \tag{45}$$

where G is the Green's function built in the previous section. Consequently, using (40) and Proposition 3, we get

$$||\tilde{w}|| \le C(\varepsilon^{3\gamma}+||Q(\tilde{u}^{app},\tilde{w}_z)||).$$

Since

$$Q(\tilde{u}^{app},\tilde{w}_z)=\int_0^1 (1-\mu)f''(\tilde{u}^{app}+\mu\tilde{w}_z)\,d\mu\cdot(\tilde{w}_z,\tilde{w}_z), \tag{46}$$

we get, thanks to (44),

$$||Q(\tilde{u}^{app},\tilde{w})|| \le C||\tilde{w}_z||_\infty||\tilde{w}_z|| \le C\varepsilon^{6\gamma-5\alpha-\frac{5}{2}}.$$

Hence

$$\frac{||\tilde{w}||}{\varepsilon^{3\gamma-\alpha}} \le C(\varepsilon^\alpha+\varepsilon^{3\gamma-4\alpha-\frac{5}{2}}). \tag{47}$$

Note that we can choose $\gamma \in (\frac{2}{3}, 1)$ and $\alpha > 0$ such that $3\gamma - 4\alpha - \frac{5}{2} > 0$. Going back to (45), we also have

$$\tilde{w}_z(z,t) = \int_0^t \int_{-\infty}^{+\infty} \partial_z G(t,\tau,z,y)(R^\varepsilon(z,\tau) + Q(\tilde{u}^{app}, \tilde{w}_z))\, dyd\tau.$$

Hence, again using (40), we get

$$\frac{||\tilde{w}_z||}{\varepsilon^{3\gamma-\alpha-\frac{1}{2}}} \leq C(\varepsilon^\alpha + \varepsilon^{3\gamma-4\alpha-\frac{5}{2}}). \tag{48}$$

Next we take the time derivative of (12) to obtain

$$\begin{aligned}(\tilde{w}_t)_t + (f'(\tilde{u}^{app}) - s'(t) + \varepsilon\delta'(t))(\tilde{w}_t)_z - \varepsilon(\tilde{w}_t)_{zz} \\ = R^\varepsilon_t + \partial_t Q(\tilde{u}^{app}, w_z) - \partial_t(f'(\tilde{u}^{app}) - s'(t) + \varepsilon\delta'(t))\tilde{w}_z,\end{aligned} \tag{49}$$

with the initial condition

$$\tilde{w}_t(z,0) = R^\varepsilon(z,0) \tag{50}$$

given by (12). Hence, again using the Green's function, we can write

$$\begin{aligned}\tilde{w}_t(z,t) = \int_{-\infty}^{+\infty} G(t,0,z,y)R^\varepsilon(z,0)\, dy \\ + \int_0^t \int_{-\infty}^{+\infty} G(t,\tau,z,y)\Big(R^\varepsilon_t + \partial_t Q(\tilde{u}^{app}, w_z) - \partial_t(f'(\tilde{u}^{app}) - s'(t) + \varepsilon\delta'(t))\tilde{w}_z\Big).\end{aligned} \tag{51}$$

Consequently, thanks to (40) and Proposition 3, we get

$$||\tilde{w}_t|| \leq C(\varepsilon^{3\gamma-\frac{1}{2}} + \varepsilon^{3\gamma-\alpha-\frac{1}{2}} + ||\partial_t Q(\tilde{u}^{app}, w_z)||).$$

Taking the time derivative of (46), we get

$$||\partial_t Q(\tilde{u}^{app}, w_z)|| \leq C(||\tilde{w}_{zt}||||\tilde{w}_{tzz}|| + ||\tilde{w}_z||||\tilde{w}_{tzz}||) \leq C\varepsilon^{6\gamma-6\alpha-3}. \tag{52}$$

This gives the estimate

$$\frac{||\tilde{w}_t||}{\varepsilon^{3\gamma-2\alpha-\frac{1}{2}}} \leq C(\varepsilon^\alpha + \varepsilon^{3\gamma-4\alpha-\frac{5}{2}}). \tag{53}$$

Taking the derivative with respect to z of (51) and again using (40), we also get

$$\frac{||\tilde{w}_{tz}||}{\varepsilon^{3\gamma-2\alpha-1}} \leq C(\varepsilon^\alpha + \varepsilon^{3\gamma-3\alpha-\frac{5}{2}}). \tag{54}$$

Next we use (12) to express $||\tilde{w}_{zz}||$. We get

$$||\tilde{w}_{zz}|| \leq \frac{C}{\varepsilon}(||\tilde{w}_t|| + ||R^\varepsilon|| + ||Q(\tilde{u}^{app}, \tilde{w}_z)||.$$

Hence

$$\frac{||\tilde{w}_{zz}||}{\varepsilon^{3\gamma-3\alpha-\frac{3}{2}}} \leq C(\varepsilon^\alpha + \varepsilon^{3\gamma-\alpha-2}). \tag{55}$$

The next step is to take the time derivative of (49), which gives

$$\begin{aligned}&(\tilde{w}_{tt})_t + (f'(\tilde{u}^{app}) - s'(t) + \varepsilon\delta'(t))(\tilde{w}_{tt})_z - \varepsilon(\tilde{w}_{tt})_{zz}\\ &= R^{\varepsilon}_{tt} + \partial_{tt}Q(\tilde{u}^{app}, \tilde{w}_z) - 2\partial_t(f'(\tilde{u}^{app}) - s'(t) + \varepsilon\delta'(t))\tilde{w}_{zt}\\ &\quad - \partial_{tt}(f'(\tilde{u}^{app}) - s'(t) + \varepsilon\delta'(t))\tilde{w}_z.\end{aligned}$$

Using (50) now, we get

$$\tilde{w}_{tz}(z,0) = R^{\varepsilon}_z(z,0), \qquad \tilde{w}_{tzz} = R^{\varepsilon}_{zz}(z,0). \tag{56}$$

Hence, also using (49) and Proposition 3, we get

$$||\tilde{w}_{tt}(\cdot,0)|| \leq C(\varepsilon^{3\gamma-1} + \varepsilon^{2\gamma+\frac{1}{2}}) \leq C\varepsilon^{3\gamma-1},$$

since $\gamma < \frac{3}{2}$.

Consequently, again using the Green's function, we get

$$||\tilde{w}_{tt}|| \leq C(\varepsilon^{3\gamma-1} + \varepsilon^{3\gamma-2\alpha-1} + \varepsilon^{3\gamma-\alpha-\frac{1}{2}} + ||\partial_{tt}Q(\tilde{u}^{app}, \tilde{w}_t)||). \tag{57}$$

It remains to estimate $||\partial_{tt}Q(\tilde{u}^{app}, \tilde{w}_t)||$; we have

$$\begin{aligned}||\partial_{tt}Q(\tilde{u}^{app}, \tilde{w}_z)|| &\leq C(||\tilde{w}_{ztt}||||\tilde{w}_z|| + ||\tilde{w}_{ztt}||||\tilde{w}_{zt}|| + ||\ |\tilde{w}_{zt}|^2||)\\ &\leq C(\varepsilon^{6\gamma-7\alpha-\frac{7}{2}} + ||\ |\tilde{w}_{zt}|^2||).\end{aligned}$$

Since

$$||\ |\tilde{w}_{zt}|^2|| \leq \int_0^T \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{z} |\tilde{w}_{zzt}(t,y)|dy\right)\left(\int_0^t |\tilde{w}_{ztt}(s,z)|ds + |\tilde{w}_{zt}(0,z)|\right) dzdt,$$

we get, thanks to (56),

$$||\ |\tilde{w}_{zt}|^2|| \leq C\varepsilon^{6\gamma-7\alpha-\frac{7}{2}}.$$

Consequently, going back to (57), we have shown

$$\frac{||\tilde{w}_{tt}||}{\varepsilon^{3\gamma-3\alpha-1}} \leq C(\varepsilon^{\alpha} + \varepsilon^{3\gamma-4\alpha-\frac{5}{2}}). \tag{58}$$

As previously shown, we also have, thanks to the estimate on the derivative of the Green's function,

$$\frac{||\tilde{w}_{ttz}||}{\varepsilon^{3\gamma-3\alpha-\frac{3}{2}}} \leq (\varepsilon^{\alpha} + \varepsilon^{3\gamma-4\alpha-\frac{5}{2}}). \tag{59}$$

Finally, thanks to (49), (52), and Proposition 3, we find

$$\frac{||\tilde{w}_{tzz}||}{\varepsilon^{3\gamma-4\alpha-2}} \leq C(\varepsilon^{\alpha} + \varepsilon^{3\gamma-\alpha-2}). \tag{60}$$

To conclude, we choose $\gamma \in (\frac{2}{3}, 1)$ and $\alpha > 0$ such that $3\gamma - 4\alpha - \frac{5}{2} > 0$ and collect (47), (48), (53), (54), (55), (58), (59), (60), which gives us

$$E(T^{\varepsilon}) \leq C\varepsilon^{\beta}$$

for some $\beta > 0$. Consequently the equality is impossible in (ii), hence $T^\varepsilon = T$ and

$$E(T) \leq 1.$$

Moreover since

$$||u^\varepsilon - u^{app}||_{L^\infty([0,T],L^1(\mathbb{R}))} = ||\tilde{w}_z||_{L^\infty([0,T],L^1(\mathbb{R}))} \leq ||\tilde{w}_{tz}|| \leq \varepsilon^{3\gamma-2\alpha-1}$$

and

$$||u^\varepsilon - u^{app}||_{L^\infty([0,T]\times\mathbb{R})} = ||\tilde{w}_z||_{L^\infty(\mathbb{R}\times[0,T])} \leq ||\tilde{w}_{tzz}||_{L^1(\mathbb{R}\times[0,T])} \leq \varepsilon^{3\gamma-4\alpha-2}$$

the theorem is proved.

Acknowledgments. I thank Emmanuel Grenier and Denis Serre for many fruitful discussions.

REFERENCES

[1] S. Bianchini and A. Bressan, *Vanishing Viscosity Solutions of Nonlinear Hyperbolic Systems,* preprint, 2001.
[2] W. A. Coppel, *Dichotomies in Stability Theory*, Lecture Notes in Math. 629, Springer-Verlag, Berlin, 1978.
[3] B. Desjardins and E. Grenier, *Linear instability implies nonlinear instability for various boundary layers,* Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 87–106.
[4] R. J. DiPerna, *Convergence of the viscosity method for isentropic gas dynamics,* Comm. Math. Phys., 91 (1983), pp. 1–30.
[5] H. Freistühler and T.-P. Liu, *Nonlinear stability of overcompressive shock waves in a rotationally invariant system of viscous conservation laws,* Comm. Math. Phys., 153 (1993), pp. 147–158.
[6] R. A. Gardner and K. Zumbrun, *The gap lemma and geometric criteria for instability of viscous shock profiles,* Comm. Pure Appl. Math., 51 (1998), pp. 797–855.
[7] J. Goodman, *Nonlinear asymptotic stability of viscous shock profiles for conservation laws,* Arch. Ration. Mech. Anal., 95 (1986), pp. 325–344.
[8] J. Goodman and Z. P. Xin, *Viscous limits for piecewise smooth solutions to systems of conservation laws,* Arch. Ration. Mech. Anal., 121 (1992), pp. 235–265.
[9] E. Grenier and F. Rousset, *Stability of one-dimensional boundary layers by using Green's functions,* Comm. Pure Appl. Math., 54 (2001), pp. 1343–1385.
[10] G. Kreiss and H.-O. Kreiss, *Stability of systems of viscous conservation laws,* Comm. Pure Appl. Math., 51 (1998), pp. 1397–1424.
[11] H.-O. Kreiss, *Initial boundary value problems for hyperbolic systems,* Comm. Pure Appl. Math., 1970.
[12] T.-P. Liu, *Nonlinear Stability of Shock Waves for Viscous Conservation Laws,* Mem. Amer. Math. Soc., 56 (328), AMS, Providence, RI, 1985.
[13] T.-P. Liu, *Nonlinear stability and instability of overcompressive shock waves*, in Shock Induced Transitions and Phase Structures in General Media, Springer-Verlag, New York, 1993, pp. 159–167.
[14] T.-P. Liu, *Pointwise convergence to shock waves for viscous conservation laws,* Comm. Pure Appl. Math., 50 (1997), pp. 1113–1182.
[15] A. Majda, *The Existence of Multidimensional Shock Fronts,* Mem. Amer. Math. Soc., 43 (281), AMS, Providence, RI, 1985.
[16] A. Majda, *The Stability of Multidimensional Shock Fronts,* Mem. Amer. Math. Soc., 41 (275), AMS, Providence, RI, 1985.
[17] K. J. Palmer, *Exponential dichotomies and transversal homoclinic points,* J. Differential Equations, 55 (1984), pp. 225–256.
[18] A. Szepessy and Z. P. Xin, *Nonlinear stability of viscous shock waves,* Arch. Ration. Mech. Anal., 122 (1993), pp. 53–103.
[19] A. I. Volpert, *Spaces BV and quasilinear equations,* Mat. Sb. (N.S.), 73 (1967), pp. 255–302.

[20] S. H. Yu, *Zero-dissipation limit of solutions with shocks for systems of hyperbolic conservation laws,* Arch. Ration. Mech. Anal., 146 (1999), pp. 275–370.
[21] K. Zumbrun and P. Howard, *Pointwise semigroup methods and stability of viscous shock waves,* Indiana Univ. Math. J., 47 (1998), pp. 741–871.
[22] K. Zumbrun and D. Serre, *Viscous and inviscid stability of multidimensional planar shock fronts,* Indiana Univ. Math. J., 48 (1999), pp. 937–992.

SIAM J. MATH. ANAL.
Vol. 35, No. 2, pp. 520–546

TRAVELING WAVES OF BISTABLE DYNAMICS ON A LATTICE*

PETER W. BATES†, XINFU CHEN‡, AND ADAM J. J. CHMAJ§

Abstract. We prove the existence of stationary or traveling waves in a lattice dynamical system arising in the theory of binary phase transitions. The system allows infinite-range couplings with positive and negative weights. The allowance for negative coupling coefficients precludes the possibility of a maximum principle. Instead, a weakened type of ellipticity is stipulated that is used with spectral theory in a perturbative fixed point argument to construct a traveling wave when the nonlinearity is unbalanced and the coupling is sufficiently strong. When the nonlinearity is balanced, a variational technique is used to obtain stationary waves, which are then analyzed in more detail for strong couplings. From a physical perspective these models are important since long-range and indefinite interactions occur in nature and can lead to pattern formation. Our results provide conditions under which patterned states tend to be swept away by traveling waves even when the interaction is of excitatory-inhibitory type. The results also have implications for the numerical analysis of spatially discretized reaction-diffusion equations, where it is important to know whether solutions to the discretized equations converge to solutions to the continuum equation as the mesh size tends to zero.

Key words. lattice dynamical system, long-range indefinite coupling, spectral theory, global minimizer

AMS subject classifications. Primary, 82B26; Secondary, 34K20, 34K30, 34K60

DOI. 10.1137/S0036141000374002

1. Introduction.

We consider the heteroclinic traveling wave problem for the lattice dynamical system

$$\dot{u}_n = \frac{1}{\varepsilon^2} \sum_{k=-\infty}^{\infty} \alpha_k u_{n-k} - f(u_n), \qquad n \in \mathbb{Z}, \tag{1.1}$$

where $0 < \varepsilon$, $\sum_k \alpha_k = 0$, $\alpha_0 < 0$, $\alpha_{-k} = \alpha_k$, and f is a smooth bistable function with nondegenerate zeros at ± 1 and an intermediate zero at $q \in (-1, 1)$. Thus, we seek a solution having the form $u_n(t) = u(\varepsilon n + ct)$ for some constant c and with the profile satisfying $u(\pm\infty) = \pm 1$. We shall not assume positivity of the α_k's for $k \neq 0$; there may be some which are negative, but we assume that $\sum_{k>0} \alpha_k k^2 > 0$ (without loss of generality we take this sum to be 1). We will assume another (spectral) positivity condition, but the foregoing describes the basic problem under consideration. The symmetry, $\alpha_{-k} = \alpha_k$, is suggested by the application to material science as indicated below, but it also assures that the convolution operator on the right in (1.1) is self-adjoint in ℓ^2.

*Received by the editors June 16, 2000; accepted for publication (in revised form) February 7, 2003; published electronically August 6, 2003.

http://www.siam.org/journals/sima/35-2/37400.html

†Department of Mathematics, Michigan State University, East Lansing, MI 48824, and Department of Mathematics, Brigham Young University, Provo, UT 84602 (bates@math.msu.edu). This author was partially supported by National Science Foundation grant DMS-9974340.

‡Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (xinfu@pitt.edu, http://www.pitt.edu/~xinfu). This author was partially supported by National Science Foundation grant DMS-9971043.

§Department of Mathematics, Heriot-Watt University, Edinburgh EH14 4AS, UK, and Department of Mathematics, Brigham Young University, Provo, UT 84602 (A.J.Chmaj@ma.hw.ac.uk, chmaj@math.byu.edu). This author was partially supported by National Science Foundation grant DMS-0096182 and by a Marie Curie Fellowship of the European Community IHP program under contract HPMFCT-2000-00465.

Setting $u_n(t) = u(\varepsilon n + ct)$ and using the properties of the α_k's, we may write the traveling wave equation for (1.1) in variable $x = \varepsilon n + ct$ as

$$cu' - \sum_{k>0} \alpha_k k^2 \frac{u(x+k\varepsilon) + u(x-k\varepsilon) - 2u(x)}{(k\varepsilon)^2} + f(u) = 0. \tag{1.2}$$

We seek a solution u such that $u(\pm\infty) = \pm 1$.

Formally, as $\varepsilon \to 0$, we obtain the traveling wave equation for the bistable reaction-diffusion equation or Allen–Cahn equation:

$$cu' - u'' + f(u) = 0 \ \text{ in } \ \mathbb{R}, \qquad u(\pm\infty) = \pm 1. \tag{1.3}$$

It is well known (e.g., [13]) that (1.3) has a unique (up to translation) traveling wave profile, u_0, and a unique wave speed, c_0. Furthermore, $u_0' > 0$, and the operator obtained by setting $c = c_0$ and linearizing the left-hand side of (1.3) at u_0 has 0 as a simple isolated eigenvalue, the remaining spectrum being in the open right half-plane [16, section 5.4]. With this nondegeneracy of the wave u_0, it is therefore natural to hope that for $\varepsilon > 0$ and sufficiently small, (1.2) also has a unique traveling wave $(u_\varepsilon, c_\varepsilon)$ close to (u_0, c_0). This is a singular perturbation problem, however, and intuition is not always correct in such problems. In the case that $c_0 \neq 0$, which is equivalent to $\int_{-1}^{1} f(u)du \neq 0$ (we say that f is unbalanced), and under certain positivity conditions on the set of coefficients $\{\alpha_k\}$, we prove that there is a locally unique traveling wave for $\varepsilon > 0$ and sufficiently small. In the case that $c_0 = 0$, under slightly stronger assumptions on the α_k's, we prove that there is a stationary wave for *all* $\varepsilon > 0$, but when ε is small the solution is not unique. The proofs in the two cases are completely different, the first being obtained through a perturbation argument and the second through a variational argument.

When $c_0 \neq 0$, we use the Fredholm alternative and a contraction mapping argument to obtain a locally unique solution near (u_0, c_0). However, there are hurdles to overcome due to the facts that the linearized operator is not self-adjoint, and it is a bounded, nonlocal operator which is in some sense approximating an unbounded, local operator.

In the case that $c_0 = 0$ (f is balanced), the linearized operator is now self-adjoint, but, in addition to the approximation difficulties mentioned above, there is a loss of regularity since (1.2) is not even a differential equation, unlike the case when $c_0 \neq 0$. To compensate, we slightly strengthen the spectral positivity assumption on $\{\alpha_k\}$ but still allow some terms to be negative. A variational argument employing Fourier analysis is used to prove the existence of a stationary wave profile for any $\varepsilon > 0$. Thus, the result is global in ε in contrast to the perturbative nature of the result for $c_0 \neq 0$. The argument involves splitting the energy functional into a part which we call *kinetic energy* and a part which we call *potential energy*. By the assumptions on $\{\alpha_k\}$, the kinetic energy is nonnegative and the potential energy is coercive. This assures that any minimizing sequence of transition states has a convergent subsequence. However, because of the long-range and indefinite interaction terms, one cannot conclude that transitions are monotone, and in fact it is difficult to localize the transitions of the terms in the minimizing sequence. It can be shown that transitions have a minimal "cost" and that a pair of well-separated transitions costs more than a single transition. These ideas are used to obtain convergence of a subsequence of suitably translated members of a minimizing sequence to an energy minimizing state which has a single transition. This minimizer is the desired stationary wave. Uniqueness is not assured.

In fact, for $\varepsilon > 0$ sufficiently small, using a minimax argument we show the existence of a second wave, in sharp contrast to the continuum case.

Concerning the origins of (1.1), consider the free energy of an Ising-like spin system on a lattice, Λ, with spins taking values (probabilistically) in an interval, I, and which interact in pairs according to the spin values and their separation. Assume that the "temperature" is below critical so that two stable spin-phases exist.

Consistent with the second law of thermodynamics, we postulate that the spin field $\{u_r\}_{r\in\Lambda}$ evolves according to the negative gradient of the free energy. The resulting dynamical system is (1.1) in the case that $\Lambda = \mathbb{Z}$, the integers. This was described in some detail in [1], where it was also shown that, in general, the coefficients $\{\alpha_k\}$ may not all be nonnegative and that f may not be balanced, even without an external field. Because the interaction coefficients represent an aggregate of different forces, α_k may change sign with k, as in the Lennard–Jones potential (see [18]). The subscript k refers to the separation between a pair of interacting lattice sites, and so it is natural that we take $\alpha_{-k} = \alpha_k$.

In taking ε small, we are considering the case where the pairwise interaction is very large, which may be the case when the lattice sites are very closely packed.

Another motivation for studying (1.1) is that one may want to establish a rigorous connection between continuum theory and numerical approximations. In that case, ε is obviously proportional to the mesh size. Noticeably, in [5], only approximate solutions of (1.2) were constructed, although an exact solution would be of more use.

The nearest neighbor version of (1.1), using the finite difference approximation of the second derivative, was studied in [17] and [21]. We note that even in that special case our result improves upon the result in [21] to some extent. There the author considered the usual discretization of the Nagumo equation and constructed traveling waves for mesh size ε small enough, of the form $u_n(t) = u(n - ct)$ with $c \neq 0$. In this paper, we show not only that those waves have nonzero speed but also that the scaled profile converges to u_0 as $\varepsilon \searrow 0$. Furthermore, we find stationary waves for all $\varepsilon > 0$ when f is balanced.

There have been many recent works on other versions of (1.1) besides those mentioned above (see, for instance, [3], [8], [9], [1], [19], [14], [15], and [20], the latter of which has a review and other references). A discrete time version was studied in [10]. Continuum but nonlocal versions were studied in [2], [4], [11], [6], [7], and [12], for example. However, for most of these results the authors assume that the interaction is nonnegative. In that case a comparison principle holds, providing some type of compactness and allowing the construction of unique monotone traveling waves.

An interesting feature of (1.1), first discussed in [17], is that for some unbalanced nonlinearities, waves are "pinned"; that is, $\int_{-1}^{1} f(u)du \neq 0$, but the wave has zero velocity. Some recent results concerning pinned states may also be found in [19] and [1], where coupling with positive and negative weights is also allowed. Of course, that is only for ε sufficiently large, as our results show. That the usual propagation criterion ($\int_{-1}^{1} f(u)du \neq 0$) holds for ε sufficiently small is also established in [1] for the case $a_k \geq 0$ when $k \neq 0$.

For brevity and to suggest what is to follow, we introduce the notation

$$\Delta_\varepsilon u \equiv \frac{1}{\varepsilon^2} \sum_{k>0} \alpha_k (u(x + \varepsilon k) + u(x - \varepsilon k) - 2u(x)). \tag{1.4}$$

Thus, we study

$$c_\varepsilon u_\varepsilon{}' - \Delta_\varepsilon u_\varepsilon + f(u_\varepsilon) = 0 \ \text{ on } \ \mathbb{R}, \quad u_\varepsilon(\pm\infty) = \pm 1. \tag{1.5}$$

We now give explicitly the assumptions on f and the coefficients $\{\alpha_k\}$ and state our results.

Even though (1.5) contains only α_k's with positive indices, in (A2) below we mention the symmetry and mean value conditions on the coefficients to remind the reader of the origins of (1.5). Assume the following:

(A1) $f \in C^2(\mathbb{R})$ has exactly three zeros, -1, $q \in (-1,1)$, and 1, with $f_u(\pm 1) > 0$.

(A2) $\sum_{k\in\mathbb{Z}} \alpha_k = 0$, $\alpha_{-k} = \alpha_k$, $\sum_{k>0} \alpha_k k^2 > 0$ (without loss of generality we take this sum to be 1), $\sum_{k>0} |\alpha_k| k^2 < \infty$, and $A(z) \equiv \sum_{k>0} \alpha_k (1 - \cos(kz)) \geq 0$ for all $z \in [0, 2\pi]$.

Note that $A(z) \geq 0$ for all $z \in [0, 2\pi]$ implies $\sum_{k>0} \alpha_k k^2 \geq 0$ (since $A''(0) \geq 0$).

THEOREM 1. *Suppose that* $\int_{-1}^{1} f(u)du \neq 0$. *Assume that* f *satisfies* (A1) *and* $\{\alpha_k\}$ *satisfy* (A2). *Then there exists a positive constant* ε^* *such that for every* $\varepsilon \in (0, \varepsilon^*)$, *problem* (1.5) *admits at least one solution,* $(c_\varepsilon, u_\varepsilon)$, *which is locally unique in* $H^1(\mathbb{R})$ *up to translation and which has the property that*

$$\lim_{\varepsilon \searrow 0} (c_\varepsilon, u_\varepsilon) = (c_0, u_0) \quad \text{in } \mathbb{R} \times H^1(\mathbb{R}).$$

In the case that $c_0 = 0$ (i.e., $\int_{-1}^{1} f(u)du = 0$) the stationary wave equation derived from (1.1) becomes a functional equation rather than a functional *differential* equation, and it does not necessarily produce a solution that is defined at points other than the integers. Given a stationary solution $\{u_n^\varepsilon\}$ to (1.1) with $\lim_{n\to\pm\infty} = \pm 1$, one may construct a solution u_n^ε to (1.5) on $\mathbb{R}$ with $c_\varepsilon = 0$ by

$$u_\varepsilon(x) = \sum u_n^\varepsilon \chi_n^\varepsilon(x), \tag{1.6}$$

where χ_n^ε is the characteristic function of the interval $(\varepsilon(n - 1/2), \varepsilon(n + 1/2)]$.

Clearly, translates of u_ε are also stationary solutions to (1.5). While u_ε constructed in this way is not continuous, we may consider the continuous linear interpolant $\tilde{u}_\varepsilon$ of u_n^ε defined by

$$\tilde{u}_\varepsilon(x) = \sum [u_n^\varepsilon + (x/\varepsilon - n)(u_{n+1}^\varepsilon - u_n^\varepsilon)] \chi_n^\varepsilon(x - \varepsilon/2) \tag{1.7}$$

to compare with u_0.

As far as existence is concerned, the fact that from (1.1) we have only a functional equation suggests that conditions stronger than (A2) are needed. Also, since (1.1) involves indices from all $\mathbb{Z}$, unlike (1.5), which has only $k > 0$, we recall the initial symmetry and mean value conditions on the coefficients α_k. We will require the following:

(A3) $\sum_{k\in\mathbb{Z}} \alpha_k = 0$, $\alpha_{-k} = \alpha_k$, $\sum_{k>0} \alpha_k k^2 = 1$, $\sum_{k>0} |\alpha_k| k^2 < \infty$, and for $z \in (0, 2\pi)$, $A(z) \equiv \sum_{k>0} \alpha_k (1 - \cos(kz)) > 0$.

The main results in this case are summarized in the following theorem.

THEOREM 2. *Suppose that* $\int_{-1}^{1} f(u)du = 0$. *Assume that* f *satisfies* (A1) *and* $\{\alpha_k\}$ *satisfy* (A3). *Then for any* $\varepsilon > 0$ *there exists a stationary solution,* $\{u_n^\varepsilon\}$, *to* (1.1) *with* $\lim_{n\to\pm\infty} = \pm 1$. *Furthermore, with* u_ε *and* $\tilde{u}_\varepsilon$ *defined by* (1.6) *and* (1.7), *respectively,*

$$\lim_{\varepsilon \searrow 0} u_\varepsilon = u_0 \quad \text{in } L^\infty(\mathbb{R}) \quad \text{and} \quad \lim_{\varepsilon \searrow 0} \tilde{u}_\varepsilon = u_0 \quad \text{in } H^1(\mathbb{R}).$$

Finally, for $\varepsilon > 0$ *and sufficiently small, there exists a second stationary solution* ${u_\varepsilon}^2$ *to* (1.5), *not a translate of* u_ε, *with* $\lim_{\varepsilon \searrow 0} u_\varepsilon = u_0$ *in* $L^\infty(\mathbb{R})$.

Remark 1. The bistable nonlinearity f can have more zeros. What we require is that for some $c = c_0$ there exists a traveling wave solution, u_0, to (1.3) which approaches ± 1 exponentially fast as x approaches $\pm\infty$.

Remark 2. Conditions regarding $A(z)$ in (A2) and (A3) are weak ellipticity conditions or positivity conditions on the operator in the spectral sense. The decay condition on the interaction coefficients translates to a regularity condition in transform space. In particular, (A2) says that $\{\alpha_k\}$ are the Fourier coefficients of a smooth, nonpositive, even, 2π-periodic function which is strictly quadratic at 0. It follows that, even though many of these coefficients may be negative, $\sum_{k>0} \alpha_k > 0$ and $\sum_{k>0} k\alpha_k > 0$. We will not need these last two observations in our proofs, but we believe that they do shed more light on the restrictions imposed by (A2). To prove the first, one need only note that $A(z)$ is nonconstant, nonnegative, with mean $\sum_{k>0} \alpha_k$. The second may be seen by observing that $v(x,y) \equiv \sum_{k>0} \alpha_k e^{-yk} \cos(kx)$ is harmonic on $(-2\pi, 2\pi) \times (0,\infty)$ having a strict maximum at $(0,0)$, and so $0 > v_y(0,0) = -\sum_{k>0} k\alpha_k$.

Remark 3. It is natural to ask if the above conditions are vacuous or difficult to meet with coefficients that change sign. The simplest example may be had for only nearest and second-nearest neighbor interactions. Then (A3) is equivalent to $\alpha_1 > 0$ and $\alpha_2 \geq -\alpha_1/4$ (allowing α_2 to be negative). As another example we take the case of only three nonzero weights. Then (A3) is satisfied if, e.g., $\alpha_2 = \alpha_3 > 0$ and $\alpha_1 > -M\alpha_2$, where $0 < M \equiv \min_{z\in[0,2\pi]}(2 - \cos(2z) - \cos(3z))/(1 - \cos z)$. Note that in the second example α_1 need not be positive. Finally, (A3) holds for small (but global in range) perturbations of the usual finite difference approximation of the Laplacian.

Remark 4. It is also natural to ask whether or not our conditions (A2) or (A3) are sharp. The condition that $\sum_{k>0} \alpha_k k^2 = 1$ (or just the positivity of that sum) is clearly what is needed to even formally approximate the Laplacian. We have no proof that the positivity of $A(z)$ is needed, though we suspect that it is. To support our beliefs consider the case where $\alpha_0 = -1/8, \alpha_1 = -1/4, \alpha_2 = 5/16$, $\alpha_{-k} = \alpha_k$, and all other coefficients are zero. Then (A2) is violated only by $A(z)$ taking on negative values on the interval $(z_0, 2\pi - z_0) \subset [0, 2\pi]$ where $\cos(z_0) = -3/5$.

Now take $\varepsilon = 1$ and $f(u) = u \pm 1$ for $\pm u < 0$ with $f(0) = 0$. One can check that there are no solutions $\{u_n\}$ with $u_n \to \pm 1$ as $n \to \pm\infty$ and with $u_n > 0$ for $\pm n > 0$.

Even without requiring the extra condition that u_n change sign only at $n = 0$ and assuming only that f is bistable and balanced with $f'(\pm 1) = 1$, there is strong evidence against the existence of a connecting orbit between $u = 1$ and $u = -1$: If the recurrence relation involving the u_n's is written as a four-dimensional discrete dynamical system, then $u = -1$ has only a one-dimensional unstable manifold and $u = 1$ has only a one-dimensional stable manifold. Thus, generically, these one-dimensional curves will not meet in the four-dimensional phase space.

Returning to the above example with piecewise linear f but now taking $\varepsilon = 1/4$, one can easily show that there is an odd ($u_n = -u_{1-n}$ for $n > 0$) stationary solution that approaches stationary solutions approximately given by $\{\mp 1 \pm 0.1 \cos n\theta\}$ for some $\theta \neq 0$ as $n \to \mp\infty$ and another odd stationary solution that approaches stationary solutions approximately given by $\{\mp 1 \pm 0.05 \sin n\theta\}$ as $n \to \mp\infty$. We believe that these two connections to oscillatory states form a barrier to the existence of connections between -1 and $+1$. It is worth noting that neither of the connections to oscillatory states takes on the value 0 and therefore f can be modified in a neighborhood of zero so that it is smooth.

Part I. Nonstationary waves. Here we assume that $\int_{-1}^{1} f(s)ds \neq 0$.

2. Reformulation of the problem. Let $(c_0, u_0(x))$ be the unique solution to

$$c_0 u_0' - u_0'' + f(u_0) = 0 \ \text{ in } \ \mathbb{R}, \qquad u_0(\pm\infty) = \pm 1, \quad u(0) = 0. \tag{2.1}$$

Then $c_0 = \int_1^{-1} f(s)ds / \int_{\mathbb{R}} (u_0'(x))^2 dx \neq 0$. We write

$$u_\varepsilon = u_0 + \phi_\varepsilon, \qquad \phi_\varepsilon \in H^1(\mathbb{R}).$$

Then the traveling wave problem (1.5) is equivalent to finding $(c_\varepsilon, \phi_\varepsilon) \in \mathbb{R} \times H^1(\mathbb{R})$ such that

$$\mathcal{L}^+_{\varepsilon,\delta}\phi_\varepsilon = \mathcal{R}(c_\varepsilon, \phi_\varepsilon), \tag{2.2}$$

where

$$\mathcal{L}^\pm_{\varepsilon,\delta}\phi = \{\pm c_0 \tfrac{d}{dx} - \Delta_\varepsilon + f_u(u_0(x)) + \delta\}\phi \qquad \text{for all } \ \phi \in H^1(\mathbb{R}), \tag{2.3}$$

$$\mathcal{R}(c, \phi) = (c_0 - c)(u_0' + \phi') + (\Delta_\varepsilon - \tfrac{d^2}{dx^2})u_0 + \delta\,\phi - \mathbf{N}(u_0, \phi), \tag{2.4}$$

$$\mathbf{N}(u_0, \phi) = f(u_0 + \phi) - f(u_0) - f_u(u_0)\phi, \tag{2.5}$$

and $\delta > 0$ is a small positive constant chosen at our convenience. The operator $\mathcal{L}^-_{\varepsilon,\delta}$ is introduced since it is the adjoint of $\mathcal{L}^+_{\varepsilon,\delta}$, and we find it more efficient to study them together.

We shall show in the next section that $\mathcal{L}^+_{\varepsilon,\delta}$ has a bounded inverse $(\mathcal{L}^+_{\varepsilon,\delta})^{-1}$ from $L^2(\mathbb{R})$ to $H^1(\mathbb{R})$, and that when restricted to the orthogonal complement of $u_0' e^{-c_0 x}$, $(\mathcal{L}^+_{\varepsilon,\delta})^{-1}$ is bounded independent of ε and δ (for sufficiently small positive ε). Hence, for every small $\phi \in H^1(\mathbb{R})$, we choose $c_\varepsilon = c_\varepsilon(\phi)$ such that $\mathcal{R}(c_\varepsilon(\phi), \phi)$ is orthogonal to $u_0' e^{-c_0 x}$. Then we define $\tilde{\phi} = (\mathcal{L}^+_{\varepsilon,\delta})^{-1}\mathcal{R}(c_\varepsilon(\phi), \phi)$. In section 4, we shall show that the mapping $\phi \to \tilde{\phi}$ is a contraction and thus possesses a fixed point, in some small ball in $H^1(\mathbb{R})$, thereby establishing the existence of a solution to (1.5).

The introduction of a small positive δ serves dual purposes: (i) It makes the operator $\mathcal{L}^+_{\varepsilon,\delta}$ invertible, so that one does not have to worry about the kernel; (ii) the translation invariance of (1.5) is removed by requiring $\mathcal{R}$ to be orthogonal to $u_0' e^{-c_0 x}$. With positive δ, $(\mathcal{L}^+_{\varepsilon,\delta})^{-1}\mathcal{R}$ automatically selects the required solution; in this way one avoids the arbitrary addition of a multiple of u_0' which is in the kernel of the limiting operator with $\delta = 0$. Thus, it makes the proof more efficient.

In what follows, $\|\cdot\|_{L^2}$, $\|\cdot\|_{L^\infty}$, and $\|\cdot\|_{H^i}$ $(i = 1, 2)$ denote the norms of the spaces $L^2(\mathbb{R})$, $L^\infty(\mathbb{R})$, and $H^i(\mathbb{R})$, respectively. Also,

$$(\phi, \psi) \equiv \int_{\mathbb{R}} \phi\psi\, dx; \qquad \phi \perp \psi \quad \Longleftrightarrow \quad (\phi, \psi) = 0.$$

We end this section with a few properties of Δ_ε which may be useful in other contexts.

LEMMA 3. *Let Δ_ε be defined as in* (1.4)*, where $\{\alpha_k\}$ satisfy* (A2)*. Then*

(1) *for any $\phi \in L^\infty(\mathbb{R})$ with $\phi'' \in L^2(\mathbb{R})$, $\|\Delta_\varepsilon\phi - \phi''\|_{L^2} \to 0$ as $\varepsilon \searrow 0$;*

(2) *for any $\phi \in H^1(\mathbb{R})$, $(\Delta_\varepsilon\phi, \phi') = 0$;*

(3) *for any $\phi, \psi \in L^2(\mathbb{R})$, $(\Delta_\varepsilon\phi, \psi) = (\phi, \Delta_\varepsilon\psi)$ and $(\Delta_\varepsilon\phi, \phi) \leq 0$.*

The proof follows from a straightforward calculation and is omitted. We point out only that, by a Fourier transform and Parseval's identity,

$$(\Delta_\varepsilon \phi, \phi) = -\frac{1}{\varepsilon^2 \pi} \int_{\mathbb{R}} \sum_{k>0} \alpha_k (1 - \cos(\varepsilon k \xi)) \Big| \mathcal{F}[\phi] \Big|^2 \, d\xi, \tag{2.6}$$

where $\mathcal{F}[\phi](\xi) = \int_{\mathbb{R}} e^{\mathbf{i}\, x\xi} \phi(x)\, dx$.

Remark 5. As one can see from our proof, properties (1)–(3) are all that are required of Δ_ε for the assertion of Theorem 1 to hold.

3. The invertibility of $\mathcal{L}^{\mp}_{\varepsilon,\delta}$. In this section, we prove the following theorem.

THEOREM 4. *There exist a positive constant C_0 and a positive function $\varepsilon_0(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$ such that for every $\delta > 0$ and every $\varepsilon \in (0, \varepsilon_0(\delta))$, $\mathcal{L}^{\pm}_{\varepsilon,\delta}$ is a homeomorphism from $H^1(\mathbb{R})$ to $L^2(\mathbb{R})$ and*

$$\Big\| (\mathcal{L}^{\pm}_{\varepsilon,\delta})^{-1} \psi \Big\|_{H^1} \le C_0 \Big\{ \|\psi\|_{L^2} + \tfrac{1}{\delta} |(\psi, \phi_0^{\mp})| \Big\}, \tag{3.1}$$

where $\phi_0^{\pm}$ is as in (3.3) *below. Consequently, for all $\delta > 0$ and $\varepsilon \in (0, \varepsilon_0(\delta))$,*

$$\| (\mathcal{L}^{\pm}_{\varepsilon,\delta})^{-1} \psi \|_{H^1} \le C_0 \, \|\psi\|_{L^2} \qquad \text{for all } \psi \in L^2 \ \text{ with } \ \psi \perp \phi_0^{\pm}. \tag{3.2}$$

To prove this theorem, we need a few preparations.

3.1. The limiting operator $\mathcal{L}_0^{\pm}$. To study the operator $\mathcal{L}^{\pm}_{\varepsilon,\delta}$, we begin by studying the $\varepsilon \to 0$ limit case, where Δ_ε becomes $\frac{d^2}{dx^2}$. Hence, we introduce operators $\mathcal{L}_0^{\pm}$ and functions $\phi_0^{\pm}$ by

$$\begin{gathered} \mathcal{L}_0^{\pm} \phi \equiv \pm c_0 \phi' - \phi'' + f_u(u_0) \phi \,, \\ \phi_0^{+} = u_0' \,/\, \|u_0'\|_{L^2}, \quad \phi_0^{-} = u_0' e^{-c_0 x} \,/\, \|u_0' e^{-c_0 x}\|_{L^2} \,. \end{gathered} \tag{3.3}$$

LEMMA 5. *Let $\mathcal{L}_0^{\pm}$ and $\phi_0^{\pm}$ be as in* (3.3). *The following hold:*
(1) *$\phi_0^{\pm} \in H^2(\mathbb{R})$ and $\mathcal{L}_0^{\pm} \phi_0^{\pm} = 0$.*
(2) *For every $\psi \in L^2(\mathbb{R})$, the problem*

$$\mathcal{L}_0^{\pm} \phi = \psi, \qquad \phi \in H^2 \ \text{ with } \ \phi \perp \phi_0^{\pm}$$

has a unique solution ϕ if and only if $\psi \perp \phi_0^{\mp}$. In addition, there exists a positive constant C_1, which depends only on f, such that

$$\|\phi\|_{H^2} \le C_1 \| \mathcal{L}_0^{\pm} \phi \|_{L^2} \qquad \text{for all } \ \phi \in H^2(\mathbb{R}) \ \text{ satisfying } \ \phi \perp \phi_0^{\pm}.$$

(3) *There exists a positive constant C_2, depending only on f, such that for every $\delta > 0$,*

$$\|\phi\|_{H^2} \le C_2 \Big\{ \|\psi\|_{L^2} + \frac{1}{\delta} |(\psi, \phi_0^{\mp})| \Big\} \qquad \text{for all } \ \phi \in H^2(\mathbb{R}), \quad \text{where } \ \psi = \mathcal{L}_0^{\pm} \phi + \delta \phi. \tag{3.4}$$

Parts (1) and (2), we believe, can be found in the literature. Here, for completeness, we provide an elementary proof. In fact, $\mathcal{L}_0^{\pm} + \delta$ is invertible, as can be seen by the same argument used in subsection 3.3.

Proof. (1) The assertion $\mathcal{L}_0^{\pm}\phi_0^{\pm} = 0$ follows by differentiating $c_0u_0' - u_0'' + f(u_0) = 0$ and a direct evaluation. We note that for some positive constants $a_{\pm}$, depending only on f,

$$u_0' \sim a_{\pm} \exp\left\{\left(c_0/2 \mp \sqrt{c_0^2/4 + f_u(\pm 1)}\right) x\right\} \quad \text{as} \quad x \to \pm\infty,$$

so that $\phi_0^{\pm} \in H^2(\mathbb{R})$. For later use, we remark that for some positive constant C,

$$\phi_0^{\pm}(x) \int_0^x \frac{dy}{\phi_0^{\pm}(y)} \le C, \quad \frac{1}{[\phi_0^{\pm}(x)]^2} \int_x^{\infty} [\phi_0^{\pm}(y)]^2\, dy \le C \qquad \text{for all } x > 0. \tag{3.5}$$

Similar estimates also hold for $x < 0$.

(2) Since for every $\phi \in H^2(\mathbb{R})$, $(\mathcal{L}_0^{\pm}\phi, \phi_0^{\mp}) = (\phi, \mathcal{L}_0^{\mp}\phi_0^{\mp}) = 0$, a necessary condition for $\mathcal{L}_0^{\pm}\phi = \psi$ to have a solution is $\psi \perp \phi_0^{\mp}$. We now show that this is also sufficient.

By using the method of variation of constants, one finds a special solution ϕ_{sp} to $\mathcal{L}_0^{\pm}\phi = \psi$ is given by

$$\phi_{sp}(x) = \phi_0^{\pm}(x) \int_0^x \frac{1}{\phi_0^{+}(y)\phi_0^{-}(y)}\, dy \int_y^{\infty} \phi_0^{\mp}(z)\psi(z)dz.$$

Using (3.5), we have, for all $y > 0$,

$$\left| \int_y^{\infty} \phi_0^{\mp}\psi \right| \le \sqrt{\textstyle\int_y^{\infty} \psi^2 \int_y^{\infty} (\phi_0^{\mp})^2} \le C\phi_0^{\mp} \sqrt{\textstyle\int_y^{\infty} \psi_2}\,.$$

Thus, for $x > 0$,

$$|\phi_{sp}(x)| \le C\phi_0^{\pm}(x) \int_0^x (\phi_0^{\pm})^{-1} \sqrt{\textstyle\int_y^{\infty} \psi^2}.$$

Hence, by (3.5), $\|\phi_{sp}\|_{L^{\infty}((0,\infty))} \le C\|\psi\|_{L^2}$. Also, by l'Hôpital's rule, $\lim_{x\to\infty} \phi_{sp}(x) = 0$.

Now if $\psi \perp \phi_0^{\mp}$, then $\int_y^{\infty} \phi_0^{\mp}\psi = \int_y^{-\infty} \phi_0^{\mp}\psi$. In a manner similar to the case where x is positive, one can show that $\|\psi\|_{L^{\infty}(-\infty,0)} \le C\|\psi\|_{L^2}$ and $\lim_{x\to-\infty} \phi_{sp}(x) = 0$. Using the differential equation and an energy estimate, we then can conclude that $\|\phi_{sp}\|_{H^2} \le C\|\psi\|_{L^2}$.

Now for $\phi \equiv \phi_{sp} - (\phi_{sp}, \phi_0^{\pm})\phi_0^{\pm}$, we have $\mathcal{L}_0^{\pm}\phi = \psi$, $\phi \perp \phi_0^{\pm}$, and $\|\phi\|_{H^2} \le C\|\psi\|_{L^2}$.

Solutions to $\mathcal{L}_0^{\pm}\phi = \psi$ with $\phi \perp \phi_0^{\pm}$ are unique since $\mathcal{L}_0^{\pm}\phi = 0$ has two linearly independent solutions, $\phi_0^{\pm}$ and $\phi_0^{\pm} \int_0^x \frac{1}{\phi_0^{+}\phi_0^{-}}$, the latter being unbounded.

(3) We consider separately the cases when δ is large, small, and intermediate.

(i) First we consider $\delta \ge \delta_1 \equiv 1 + \|f_u(u_0)\|_{L^{\infty}}$. Let $\phi \in H^2$ be arbitrary and set $\psi = \mathcal{L}_0^{\pm}\phi + \delta\phi$. Then $(\mathcal{L}_0^{\pm}\phi + \delta\phi\,,\, \phi) = (\psi, \phi)$, so that $(\delta - \|f_u(u_0)\|_{L^{\infty}})\|\phi\|_{L^2} \le \|\psi\|_{L^2}$. It then follows from $\pm c_0\phi' - \phi'' = \psi - (f_u(u_0) + \delta)\phi$ that $\|\phi\|_{H^2} \le C\|\psi\|_{L^2}$ for some C independent of δ. Hence, (3.4) holds.

(ii) Next we consider $\delta \in (0, \delta_0]$, where $\delta_0 > 0$ is to be defined later. Again, set $\psi = \mathcal{L}_0^{\pm}\phi + \delta\phi$, where $\phi \in H^2$ is arbitrary. Decompose $\phi = (\phi, \phi_0^{\pm})\phi_0^{\pm} + \phi^{\perp}$. Then $\mathcal{L}_0^{\pm}\phi^{\perp} = \psi - \delta\phi$, so that, from the second part of assertion (2),

$$\|\phi^{\perp}\|_{H^2} \le C_1\{\|\psi\|_{L^2} + \delta\|\phi\|_{L^2}\}. \tag{3.6}$$

In addition, $\psi-\delta\phi \perp \phi_0^{\mp}$; that is, $(\psi,\phi_0^{\mp})=\delta(\phi,\phi_0^{\mp})=\delta(\phi,\phi_0^{\pm})(\phi_0^{+},\phi_0^{-})+\delta(\phi^{\perp},\phi_0^{\mp})$. It then follows that, denoting $\sigma=(\phi_0^{+},\phi_0^{-})=\int_{\mathbb{R}}(u_0')^2e^{-c_0x}dx/(\|u_0'\|_{L^2}\|u_0'e^{-c_0x}\|_{L^2})$,

$$\sigma|(\phi,\phi_0^{\pm})|\le\|\phi^{\perp}\|_{L^2}+\frac{1}{\delta}|(\psi,\phi_0^{\mp})|\,.$$

Adding twice (3.6) to the above, we then obtain

$$\sigma|(\phi,\phi_0^{\pm})|+\|\phi^{\perp}\|_{H^2}\le\frac{1}{\delta}|(\psi,\phi_0^{\mp})|+2C_1\|\psi\|_{L^2}+2\delta C_1\|\phi\|_{L^2}.$$

It follows, as $\sigma<1$ and $\|\phi\|_{L^2}\le\|\phi^{\perp}\|_{L^2}+|(\phi,\phi_0^{\pm})|$, that

$$(\sigma-2\delta C_1)\|\phi\|_{L^2}\le 2C_1\|\psi\|_{L^2}+\frac{1}{\delta}|(\psi,\phi_0^{\mp})|.$$

Taking $\delta_0=\sigma/(4C_1)$ we conclude that if $\delta\in(0,\delta_0]$, then $\|\phi\|_{L^2}\le\frac{4C_1}{\sigma}\{\|\psi\|_{L^2}+\frac{1}{\delta}|(\psi,\phi_0^{\pm})|\}$. Hence, (3.4) holds with C_2 independent of δ.

(iii) Finally we consider $\delta\in[\delta_0,\delta_1]$. Since $\phi_0^{\pm}$ is a positive solution to $\mathcal{L}_0^{\pm}\phi=0$, by Liouville's theorem, $(\mathcal{L}_0^{\pm}+\delta)\phi=0$ does not have any nontrivial bounded solution. Define $\tilde\Lambda^{\pm}(\delta)=\inf_{\|\phi\|_{H^2}=1}\|\mathcal{L}_0^{\pm}\phi+\delta\phi\|_{L^2}$, and $\hat\Lambda^{\pm}=\inf_{\delta\in[\delta_0,\delta_1]}\tilde\Lambda^{\pm}(\delta)$. We claim that $\hat\Lambda^{\pm}>0$, and hence (3.4) holds with C_2 independent of δ. To show that $\hat\Lambda^{\pm}>0$, take a sequence $\{(\delta_j,\phi_j)\}$ minimizing $\|\mathcal{L}_0^{\pm}\phi+\delta\phi\|_{L^2}$. Writing $\psi_j=\mathcal{L}_0^{\pm}\phi_j+\delta_j\phi_j$, we may assume that $\delta_j\to\delta\in[\delta_0,\delta_1]$, $\phi_j\to\phi$ weakly in $H^2(\mathbb{R})$ and strongly in $L^2_{loc}(\mathbb{R})$, and $\psi_j\to\psi$ weakly in $L^2(\mathbb{R})$. One can show that ϕ is a weak, and hence strong, solution to $\mathcal{L}_0^{\pm}\phi+\delta\phi=\psi$, and by the weak lower semicontinuity of the norm, $\|\psi\|_{L^2}\le\hat\Lambda^{\pm}$. If $\hat\Lambda^{\pm}=0$, then $\psi=0$ and hence $\phi=0$.

On the other hand, Cauchy's inequality applied to

$$-\|\phi_j''\|_{L^2}^2+(f_u(u_0)\phi_j,\phi_j'')\ge(\mathcal{L}_0^{\pm}\phi_j+\delta_j\phi_j,\phi_j'')=(\psi_j,\phi_j'')$$

and squaring gives

$$2\|f_u(u_0)\|_{L^\infty}^2\|\phi_j\|_{L^2}^2\ge\|\phi_j''\|_{L^2}^2-2\|\psi_j\|_{L^2}^2.$$

Using this together with (3.9) and (3.10) below, one finds

$$\int_{|x|\le m}\phi_j^2\ge\bar C_3\|\phi_j\|_{H^2}^2-\bar C_4\|\psi_j\|_{L^2}^2$$

for some positive constants $\bar C_3$ and $\bar C_4$. The constant m is defined below, but the basic idea is that because of the asymptotic values of $f_u(u_o)$, one can localize the mass of ϕ_j. Passing to the limit gives

$$\int_{|x|\le m}\phi^2\ge\bar C_3>0,$$

contradicting $\phi\equiv0$. This completes the proof. □

3.2. Bound for the inverse of $\mathcal{L}_{\varepsilon,\delta}^{\pm}$. For every positive δ and ε, we define

$$\Lambda^{\pm}(\varepsilon,\delta)=\inf_{\|\phi\|_{H^1}=1}\Big\{\|\,\mathcal{L}_{\varepsilon,\delta}^{\pm}\phi\,\|_{L^2}+\frac{1}{\delta}|(\mathcal{L}_{\varepsilon,\delta}^{\pm}\phi\,,\,\phi_0^{\mp})|\Big\},\quad\Lambda^{\pm}(\delta)=\liminf_{\varepsilon\searrow0}\Lambda^{\pm}(\varepsilon,\delta).$$

(3.7)

LEMMA 6. *There exists a positive constant C_0 such that $\Lambda^{\pm}(\delta) > 2/C_0$ for all $\delta > 0$.*

Proof. Let $\delta > 0$ be any fixed positive constant. By the definition of $\Lambda^{\pm}(\delta)$, there exists a sequence $\{(\varepsilon_j, \phi_j)\}_{j=1}^{\infty}$ in $(0,1) \times H^1(\mathbb{R})$ such that $\lim_{j\to\infty} \varepsilon_j = 0$, $\|\phi_j\|_{H^1} = 1$ for all j, and $\psi_j \equiv \mathcal{L}^{\pm}_{\varepsilon_j,\delta}\phi_j$ satisfies

$$\lim_{j\to\infty} \Big\{ \|\psi_j\|_{L^2} + \frac{1}{\delta}|(\psi_j,\, \phi_0^{\pm})| \Big\} = \Lambda^{\pm}(\delta).$$

By taking a subsequence if necessary, we can assume that there exist functions $\phi \in H^1$ and $\psi \in L^2$ such that, as $j \to \infty$,

$$\begin{aligned} \phi_j &\longrightarrow \phi \quad \text{in } L^2_{\mathrm{loc}}(\mathbb{R}) \text{ and weakly in } H^1(\mathbb{R}), \\ \psi_j &\longrightarrow \psi \quad \text{weakly in } L^2(\mathbb{R}). \end{aligned}$$

By the weak lower semicontinuity of the $L^2(\mathbb{R})$ norm, $\|\psi\|_{L^2} + \frac{1}{\delta}|(\psi, \phi_0^{\mp})| \le \Lambda^{\pm}(\delta)$.

For any test function $\zeta \in C_0^{\infty}(\mathbb{R})$, $(\psi_j, \zeta) = (\mathcal{L}^{\pm}_{\varepsilon_j,\delta}\phi_j,\, \zeta) = (\phi_j, \mathcal{L}^{\mp}_{\varepsilon_j,\delta}\zeta)$. Since $\lim_{\varepsilon\searrow 0} \|\Delta_\varepsilon\zeta - \zeta''\|_{L^2} = 0$, sending $j \to \infty$ we obtain $(\psi, \zeta) = (\phi,\, (\mathcal{L}_0^{\mp} + \delta)\zeta)$ for all $\zeta \in C_0^{\infty}(\mathbb{R})$. That is, $\phi \in H^1(\mathbb{R})$ is a weak solution to $(\mathcal{L}_0^{\pm} + \delta)\phi = \psi$. An elliptic estimate then shows that ϕ is in $H^2(\mathbb{R})$. Consequently, by Lemma 5(3),

$$\|\phi\|_{H^2} \le C_2\Big\{ \|\psi\|_{L^2} + \frac{1}{\delta}|(\psi, \phi_0^{\mp})| \Big\} \le C_2\Lambda^{\pm}(\delta). \tag{3.8}$$

It remains to find a positive lower bound of $\|\phi\|_{L^2}$.

First of all, using $(\mathcal{L}^{\pm}_{\varepsilon_j,\delta}\phi_j, \phi_j') = (\psi_j, \phi_j')$ and the identity $(\Delta_\varepsilon\phi_j, \phi_j') = 0 = (\phi_j, \phi_j')$, we obtain $\pm c_0\|\phi_j'\|_{L^2}^2 = (\psi_j, \phi_j') - (f_u(u_0)\phi_j, \phi_j')$. Cauchy's inequality then gives

$$\|f_u(u_0)\|_{L^\infty}\|\phi_j\|_{L^2} \ge |c_0|\, \|\phi_j'\|_{L^2} - \|\psi_j\|_{L^2},$$

which implies

$$2\|f_u(u_0)\|_{L^\infty}^2\|\phi_j\|_{L^2}^2 \ge c_0^2\|\phi_j'\|_{L^2}^2 - 2\|\psi_j\|_{L^2}^2. \tag{3.9}$$

Let m be a positive constant such that

$$0 < a \equiv \tfrac{1}{2}\min\{f_u(1), f_u(-1)\} = \min_{|x|\ge m}\{f_u(u_0(x))\}.$$

Using $(\psi_j, \phi_j) = (\mathcal{L}^{\pm}_{\varepsilon_j,\delta}\phi_j, \phi_j)$, the identity $(\phi_j', \phi_j) = 0$, and the fact $(-\Delta_{\varepsilon_j}\phi_j, \phi_j) \ge 0$, we obtain

$$\begin{aligned} (\psi_j, \phi_j) &\ge (f_u(u_0)\phi_j, \phi_j) \ge \min_{|x|\ge m}\{f_u(u_0)\}\int_{|x|\ge m}\phi_j^2 - \|f_u(u_0)\|_{L^\infty}\int_{|x|\le m}\phi_j^2 \\ &= a\|\phi_j\|_{L^2}^2 - (a + \|f_u(u_0)\|_{L^\infty})\int_{|x|\le m}\phi_j^2. \end{aligned}$$

Therefore,

$$(a + \|f_u(u_0)\|_{L^\infty})\int_{|x|\le m}\phi_j^2 \ge a\|\phi_j\|_{L^2}^2 - (\phi_j, \psi_j) \ge \tfrac{a}{2}\|\phi_j\|_{L^2}^2 - \tfrac{1}{2a}\|\psi_j\|_{L^2}^2. \tag{3.10}$$

Adding a small multiple ($\frac{a}{2(2\|f_u(u_0)\|_{L^\infty}^2+c_0^2)}$ in fact) of (3.9), we see that there exist positive constants C_3 and C_4, which depend on $|c_0|>0$ and f, such that

$$\int_{|x|\le m}\phi_j^2 \ge C_3\|\phi_j\|_{H^1}^2 - C_4\|\psi_j\|_{L^2}^2 = C_3 - C_4\|\psi_j\|_{L^2}^2.$$

Sending $j\to\infty$ we then conclude that

$$\int_{|x|\le m}\phi^2 \ge C_3 - C_4\Lambda^{\pm}(\delta). \tag{3.11}$$

In view of (3.8), we then obtain $\Lambda^{\pm}(\delta)^2 \ge \sqrt{C_3/(C_2^2+C_4)} \equiv 2/C_0$. This completes the proof. □

Remark 6. From the proof one sees that the condition $c_0\neq 0$ plays a key role, for it ensures the boundedness of $\{\phi_j\}$ in $H^1(\mathbb{R})$ and hence guarantees (3.11) for the weak limit.

3.3. Proof of Theorem 4. Now we are ready to complete the proof of Theorem 4.

Let $\delta>0$ be fixed. Since $\Lambda^{\pm}(\delta)\ge 2/C_0$, there exists $\varepsilon_0(\delta)>0$ such that $\Lambda(\varepsilon,\delta)\ge 1/C_0$ for every $\varepsilon\in(0,\varepsilon_0(\delta)]$. Now we consider the operator $\mathcal{L}^{\pm}_{\varepsilon,\delta}$ with $\varepsilon\in(0,\varepsilon_0(\delta)]$.

First of all, $\mathcal{L}^{\pm}_{\varepsilon,\delta}$ is a bounded operator from $H^1(\mathbb{R})$ to $L^2(\mathbb{R})$. Also, by the definition and the lower bound of $\Lambda^{\pm}(\varepsilon,\delta)$, $\mathcal{L}^{\pm}_{\varepsilon,\delta}$ is a homeomorphism from $H^1(\mathbb{R})$ to its image $\mathcal{L}^{\pm}_{\varepsilon,\delta}(H^1(\mathbb{R}))$, and the inverse $(\mathcal{L}^{\pm}_{\varepsilon,\delta})^{-1}$ from $\mathcal{L}^{\pm}_{\varepsilon,\delta}(H^1(\mathbb{R}))\subseteq L^2(\mathbb{R})$ to $H^1(\mathbb{R})$ is bounded by $1/\Lambda^{\pm}(\varepsilon,\delta)\le C_0$. As a consequence of the boundedness, we see that $\mathcal{L}^{\pm}_{\varepsilon,\delta}(H^1(\mathbb{R}))$ is closed in $L^2(\mathbb{R})$.

It remains to show that $\mathcal{L}^{\pm}_{\varepsilon,\delta}(H^1(\mathbb{R}))=L^2(\mathbb{R})$. Indeed, if this were not true, there would exist a nontrivial $\psi\in L^2(\mathbb{R})$ orthogonal to $\mathcal{L}^{\pm}_{\varepsilon,\delta}(H^1(\mathbb{R}))$, i.e., $0=(\mathcal{L}^{\pm}_{\varepsilon,\delta}\phi,\psi)=(\phi,\mathcal{L}^{\mp}_{\varepsilon,\delta}\psi)$ for all $\phi\in H^1(\mathbb{R})$. In particular, $(\phi,\mathcal{L}^{\mp}_{\epsilon,\delta}\psi)=0$ for all $\phi\in C_0^\infty(\mathbb{R})$, and therefore the weak derivative of ψ is in $L^2(\mathbb{R})$. Consequently, $\mp c_0\psi' = (\Delta_\varepsilon - f_u(u_0)-\delta)\psi$ is in $L^2(\mathbb{R})$ since $\psi\in L^2(\mathbb{R})$. Thus, $\psi\in H^1(\mathbb{R})$ and $\mathcal{L}^{\mp}_{\varepsilon,\delta}\psi=0$. By the definition and the positivity of $\Lambda(\varepsilon,\delta)$, we then conclude that $\psi=0$ and therefore obtain a contradiction. Thus $\mathcal{L}^{\pm}_{\varepsilon,\delta}(H^1(\mathbb{R}))=L^2(\mathbb{R})$, and this completes the proof.

4. Existence of traveling waves.

Proof of Theorem 1. Let δ and η be small positive constants to be determined later. We define

$$\mathbf{X}_\eta \equiv \{\phi\in H^1(\mathbb{R}) \ : \ \|\phi\|_{H^1}\le\eta\}.$$

For every $\phi\in\mathbf{X}_\eta$, let $c_\varepsilon=c_\varepsilon(\phi)$ be the (unique) constant such that $\mathcal{R}(c_\varepsilon,\phi)\perp\phi_0^-$, i.e.,

$$c_\varepsilon(\phi)\equiv c_0 + \frac{(\Delta_\varepsilon u_0 - u_0'',\,\phi_0^-)+\delta(\phi,\phi_0^-)-(\mathbf{N}(u_0,\phi),\phi_0^-)}{(u_0',\phi_0^-)+(\phi',\phi_0^-)}. \tag{4.1}$$

We define $\mathbf{T}:\mathbf{X}_\eta\subset H^1(\mathbb{R})\to H^1(\mathbb{R})$ by

$$\mathbf{T}\phi = (\mathcal{L}^{+}_{\varepsilon,\delta})^{-1}\mathcal{R}(c_\varepsilon(\phi),\phi). \tag{4.2}$$

We now show that $\mathbf{T}$ maps $\mathbf{X}_\eta$ into itself and is a contraction, so that it possesses a fixed point, which, after adding u_0, gives a solution to (1.5), thereby completing the proof.

Since $\mathcal{R}(c_\varepsilon(\phi),\phi) \perp \phi_0^-$, we derive from Theorem 4 that

$$\|\mathbf{T}\phi\|_{H^1} \le C_0\|\mathcal{R}(c_\varepsilon(\phi),\phi)\|_{L^2} \quad \text{for all } \phi \in \mathbf{X}_\eta, \text{ and} \tag{4.3}$$

$$\|\mathbf{T}\phi_1 - \mathbf{T}\phi_2\|_{H^1} \le C_0\|\mathcal{R}(c_\varepsilon(\phi_1),\phi_1) - \mathcal{R}(c_\varepsilon(\phi_2),\phi_2)\|_{L^2} \quad \text{for all } \phi_1,\phi_2 \in \mathbf{X}_\eta. \tag{4.4}$$

Now we estimate the right-hand sides.

Let $\hat\sigma$ be defined as

$$\hat\sigma = \frac{1}{2}(u_0', \phi_0^-) = \frac{\int_{\mathbb{R}} {u_0'}^2 e^{-c_0 x}}{2(\int_{\mathbb{R}} {u_0'}^2 e^{-2c_0 x})^{1/2}} > 0.$$

Then $(u_0' + \phi', \phi_0^-) = 2\hat\sigma + (\phi', \phi_0^-) \ge 2\hat\sigma - \eta \ge \hat\sigma$ if we require $\eta \le \hat\sigma$.

To estimate the nonlinear term $\mathbf{N}(u_0,\phi)$, we first recall the embedding $\|\phi\|_{L^\infty} \le \|\phi\|_{H^1}$ for every $\phi \in H^1$. Hence, setting $M = \sup_{|s|\le 1+\hat\sigma} |f_{uu}(s)|$, we have

$$|\mathbf{N}(u_0,\phi)| \le M\eta|\phi| \quad \text{and} \quad |\mathbf{N}(u_0,\phi_1) - \mathbf{N}(u_0,\phi_2)| \le M\eta|\phi_1 - \phi_2|$$

pointwise for all $\phi, \phi_1, \phi_2 \in \mathbf{X}_\eta$.

By the definition of $c_\varepsilon(\phi)$ in (4.1), for all $\phi, \phi_1, \phi_2 \in \mathbf{X}_\eta$,

$$|c_\varepsilon(\phi) - c_0| \le \hat\sigma^{-1}\{\|\Delta_\varepsilon u_0 - u_0''\|_{L^2} + (\delta + M\eta)\eta\}$$

and

$$|c_\varepsilon(\phi_1) - c_\varepsilon(\phi_2)| \le \|\phi_1 - \phi_2\|_{L^2}\hat\sigma^{-2}\{\|\Delta_\varepsilon u_0 - u_0''\|_{L^2} + (\hat\sigma + \eta)(\delta + M\eta)\}\,.$$

Therefore, using the expression for $\mathcal{R}(c_\varepsilon(\phi),\phi)$ in (2.4) we can estimate

$$\begin{aligned}\|\mathcal{R}(c_\varepsilon(\phi),\phi)\|_{L^2} &\le |c_\varepsilon(\phi) - c_0|(\|u_0'\| + \eta) + \|\Delta_\varepsilon u_0 - u_0''\|_{L^2} + \delta\eta + M\eta^2 \\ &\le \eta\{1 + \hat\sigma^{-1}(\|u_0'\| + \eta)\}\{\eta^{-1}\|\Delta_\varepsilon u_0 - u_0''\|_{L^2} + \delta + M\eta\}\end{aligned}$$

and

$$\begin{aligned}&\|\mathcal{R}(c_\varepsilon(\phi_1),\phi_1) - \mathcal{R}(c_\varepsilon(\phi_2),\phi_2)\|_{L^2} \\ &\le \{|c_\varepsilon(\phi_1) - c_0| + \delta + M\eta\}\|\phi_1 - \phi_2\|_{L^2} + (\|u_0'\|_{L^2} + \eta)|c_\varepsilon(\phi_1) - c_\varepsilon(\phi_2)| \\ &\le \|\phi_1 - \phi_2\|_{L^2}\hat\sigma^{-2}\{\hat\sigma + \eta + \|u_0'\|_{L^2}\}\{\|\Delta_\varepsilon u_0 - u_0''\|_{L^2} + (\hat\sigma + \eta)(\delta + M\eta)\}.\end{aligned}$$

It then follows from (4.3) and (4.4) that there exists a positive constant C_5, which is independent of δ, ε and $\eta \in (0, \hat\sigma]$, such that

$$\|\mathbf{T}\phi\|_{H^1} \le \eta C_5\{\eta^{-1}\|\Delta_\varepsilon u_0 - u_0''\|_{L^2} + \delta + M\eta\}$$

and

$$\|\mathbf{T}\phi_1 - \mathbf{T}\phi_2\|_{H^1} \le C_5\big\{\|\Delta_\varepsilon u_0 - u_0''\|_{L^2} + \delta + M\eta\big\}\|\phi_1 - \phi_2\|_{H^1}.$$

Now we fix

$$\delta = \frac{1}{4C_5} \qquad \text{and} \qquad \eta = \min\Big\{\hat\sigma, \frac{1}{4MC_5}\Big\}.$$

We select a small positive ε^* such that $\varepsilon^* \le \varepsilon_0(\delta)$ (cf. Theorem 4) and

$$\sup_{\varepsilon\in(0,\varepsilon^*]} \|\Delta_\varepsilon u_0 - u_0''\|_{L^2} \le \frac{\min\{1,\eta\}}{4C_5}\,.$$

We then conclude that for any fixed $\varepsilon \in (0,\varepsilon^*]$, $\mathbf{T}$ maps $\mathbf{X}_\eta$ into itself and is a contraction. This completes the proof.

Remark 7. From the proof, one sees that the solution to (1.5) is locally unique.

Part II. Stationary waves. Here we assume that $\int_{-1}^{1} f = 0$, and hence $c_0 = 0$.

5. Assumptions and results. By the assumption that f is balanced, it is the derivative of a double equal-well potential; more precisely, since $f_u(\pm1) > 0$ we may write

$$f(u) = F_u(u), \quad \text{where} \quad F(\pm1) = 0 \ \text{ and } \quad F(u) \ge m_0(1-u^2)^2 \tag{5.1}$$

for a positive constant m_0 and $u \in [-2,2]$. When u is outside $[-2,2]$, we assume that F is bounded away from zero and approaches infinity as $|u| \to \infty$. We look for a function u_ε such that

$$-\Delta_\varepsilon u_\varepsilon + f(u_\varepsilon) = 0 \quad \text{ for all } x \in \mathbb{R}, \qquad \lim_{x\to\pm\infty} u_\varepsilon(x) = \pm1. \tag{5.2}$$

Observe that if u_ε is a solution, then for every $x_0 \in \mathbb{R}$, if we define $u_n = u_\varepsilon(x_0 + n\varepsilon)$ for all $n \in \mathbb{Z}$, then $\{u_n\}_{n=-\infty}^{\infty}$ satisfies

$$\begin{cases} -\varepsilon^{-2}\sum_{k>0}\alpha_k(u_{n+k} + u_{n-k} - 2u_n) + f(u_n) = 0 \quad \text{ for all } \ n \in \mathbb{Z}, \\ \qquad\qquad \lim_{n\to\pm\infty} u_n = \pm1. \end{cases} \tag{5.3}$$

On the other hand, if a sequence $\{u_n\}$ satisfies (5.3), then the function $u_\varepsilon = \sum_n u_n\chi_n^\varepsilon$ is a solution to (5.2) where χ_n^ε is the characteristic function of the set $(\varepsilon(n-1/2), \varepsilon(n+1/2)]$, i.e.,

$$\chi_n^\varepsilon(x) = \begin{cases} 1 & \text{if } \ x \in (\varepsilon(n-1/2), \varepsilon(n+1/2)], \\ 0 & \text{otherwise.} \end{cases} \tag{5.4}$$

Hence, the solvability of (5.2) and (5.3) are equivalent.

In what follows, $\sum_{n=-\infty}^{\infty}$ is denoted by $\sum_n$. Also a sequence $\{u_n\}$ is identified with a function u defined by $u(x) = \sum_n u_n\chi_n^\varepsilon(x)$, where $\chi_n^\varepsilon(\cdot)$ is as in (5.4).

To show that (5.2) or (5.3) admits a solution, we need the stronger assumption, (A3) given in the introduction, than we needed for the nonstationary wave case. A useful consequence is given here. Let

$$B(\zeta) \equiv \sum_{k>0} \alpha_k \frac{\sin^2\frac{k\zeta}{2}}{\sin^2\frac{\zeta}{2}} = \sum_{k>0}\alpha_k\Big|\sum_{\ell=0}^{k-1} e^{\mathbf{i}\,\ell\,\zeta}\Big|^2, \quad \zeta \in \mathbb{R}. \tag{5.5}$$

Notice that $B(\cdot)$ is 2π-periodic and even. It can be expanded as a Fourier cosine series:

$$B(\zeta) = \frac{b_0}{2} + \sum_{\ell>0} b_\ell\cos(\ell\zeta) \ \text{ for } \zeta \in \mathbb{R}, \quad \text{where } \ b_\ell = \frac{1}{\pi}\int_0^{2\pi} B(\zeta)\cos(\ell\zeta)\,d\zeta. \tag{5.6}$$

We assert the following.

LEMMA 7. *Assume* (A3). *Then $B(\zeta)$ is uniformly positive, bounded, and in $H^{1/2}([0,2\pi])$; that is, there exists a positive constant B_∞ such that*

$$\frac{1}{B_\infty} \le B(\zeta) \le B_\infty \qquad \text{for all } \zeta \in \mathbb{R}, \tag{5.7}$$

$$\sum_{\ell>0} \ell\, b_\ell^2 < \infty. \tag{5.8}$$

Proof. By (A3), $B(z) > 0$ and twice differentiable on $(0, 2\pi)$ and by l'Hôpital's rule, $\lim_{z\to 0,2\pi} B(z) = 1$. This shows that B is bounded above and below by positive constants. To show the regularity, it is convenient to write the Fourier series for B as

$$\sum_{k=-\infty}^{\infty} a_k \cos(kz).$$

We have

$$\sum_{k=-\infty}^{\infty} \alpha_k(1-\cos(kz)) = 2\sum_{k=-\infty}^{\infty} a_k \cos(kz)(1-\cos(z)).$$

It follows that

$$\alpha_k = a_{k+1} + a_{k-1} - 2a_k \qquad \text{for } \ k \neq 0.$$

This may be inverted to give

$$a_k = \sum_{j>0} j\alpha_{k+j}.$$

Now, $B \in H^{\frac{1}{2}}$ is equivalent to the convergence of $\sum_{k>0} k a_k^2$. In light of the above, we have

$$\begin{aligned}
\sum_{k>0} k a_k^2 &= \sum_{k>0} k \sum_{j>0} j\alpha_{j+k} \sum_{i>0} i\alpha_{i+k} \\
&\le \sum_{k>0}\sum_{j>k}\sum_{i>k} k\frac{(j-k)}{j^2}\frac{(i-k)}{i^2}|\alpha_j| j^2 |\alpha_i| i^2 \\
&= \sum_{j>0}\sum_{i>0}\Big(\sum_{k=1}^{\min\{j-1,i-1\}} k\frac{(j-k)}{j^2}\frac{(i-k)}{i^2}\Big)|\alpha_j| j^2 |\alpha_i| i^2.
\end{aligned}$$

The proof is completed by observing that the sum over k is bounded by $\frac{3}{4}$ and the other sums are finite by (A3). ∎

We will derive some energy estimates and then prove the existence part of Theorem 2.

THEOREM 8. *Assume that $f(u)$ satisfies* (5.1) *and $\{\alpha_k\}_{k=1}^{\infty}$ satisfies* (A3). *Then for every $\varepsilon > 0$, problem* (5.2) *or problem* (5.3) *admits at least one solution.*

We use a minimization method with an energy $\mathbf{E}$ defined by

$$\mathbf{E}[u] = \mathcal{E}_{ki}[u] + \mathcal{E}_{po}[u], \quad \mathcal{E}_{po}[u] = 2\int_{\mathbb{R}} F(u)\,dx, \quad \mathcal{E}_{ki}[u] = -\int_{\mathbb{R}} u\Delta_\varepsilon u\,dx, \tag{5.9}$$

where "po" stands for potential and "ki" for kinetic. We show that $\mathbf{E}[\cdot]$ has a minimizer in the following space

$$\mathbf{X} \equiv \left\{ u = \sum_n u_n \chi_n^\varepsilon \ : \ \sum_{n=0}^{\infty} |1-u_n|^2 + \sum_{n=-1}^{-\infty} |1+u_n|^2 < \infty \right\}. \tag{5.10}$$

LEMMA 9. *If $u^\varepsilon \in \mathbf{X}$ satisfies $\mathbf{E}[u^\varepsilon] = \inf_{u\in\mathbf{X}} \mathbf{E}[u]$, then u^ε solves* (5.3).

The proof follows a standard variation technique and is omitted. We only remark that if $u \in \mathbf{X}$, then $\Delta_\varepsilon u \in L^2(\mathbb{R})$, due to the assumption that $\sum_k |\alpha_k|\, k^2 < \infty$.

6. The energy. First we write the kinetic energy $-(\Delta_\varepsilon u, u)$ in a convenient form for piecewise constant functions.

LEMMA 10. *Let $u(x) = \sum_n u_n \chi_n^\varepsilon$ and $v(x) = \sum_n v_n \chi_n^\varepsilon$. Assume that $\sum_n (u_{n+1}-u_n)^2 < \infty$ and $\sum_n (v_{n+1}-v_n)^2 < \infty$. Then*

$$(-\Delta_\varepsilon u, v) = \frac{1}{2\pi\varepsilon} \int_0^{2\pi} B(\zeta)\phi(\zeta)\overline{\psi(\zeta)}\, d\zeta, \tag{6.1}$$

where $B(\zeta)$ is as in (5.5),

$$\phi(\zeta) = \sum_n (u_{n+1}-u_n) e^{\mathbf{i}n\zeta}, \quad \text{and} \quad \psi(\zeta) = \sum_m (v_{m+1}-v_m) e^{\mathbf{i}m\zeta}\ . \tag{6.2}$$

Consequently,

$$\mathcal{E}_{ki}[u] = \frac{1}{2\pi\varepsilon} \int_0^{2\pi} B(\zeta) \Big|\phi(\zeta)\Big|^2 d\zeta.$$

Proof. By definition,

$$\begin{aligned}
(-\Delta_\varepsilon u, v) &= \frac{1}{\varepsilon^2} \sum_{k>0} \alpha_k \Big\{ -\Big(u(x+\varepsilon k) - u(x), v(x)\Big) + \Big(u(x) - u(x-\varepsilon k), v(x)\Big)\Big\} \\
&= \frac{1}{\varepsilon^2} \sum_{k>0} \alpha_k \Big(u(x+\varepsilon k) - u(x), v(x+\varepsilon k) - v(x)\Big) \\
&= \frac{1}{2\pi\varepsilon^2} \sum_{k>0} \alpha_k \int_{\mathbb{R}} \mathcal{F}[u(x+\varepsilon k) - u(x)] \overline{\mathcal{F}[v(x+\varepsilon k) - v(x)]}\, d\xi
\end{aligned}$$

by Parseval's identity, where $\mathcal{F}[u]$ is the Fourier transform of u; see (2.6). For any $h > 0$,

$$\begin{aligned}
&(1 - e^{\mathbf{i}\varepsilon\xi})\, \mathcal{F}[u(x+h) - u(x)] \\
&= \mathcal{F}[u(x+h) - u(x) - u(x+h-\varepsilon) + u(x-\varepsilon)] \\
&= \mathcal{F}[u(x+h) - u(x+h-\varepsilon)] - \mathcal{F}[u(x) - u(x-\varepsilon)] \\
&= \sum_n (u_{n+1} - u_n) \int_{\varepsilon(n+1/2)}^{\varepsilon(n+3/2)} e^{\mathbf{i}x\xi} dx (e^{-\mathbf{i}h} - 1) \\
&= \frac{(1 - e^{\mathbf{i}\varepsilon\xi}) e^{-\mathbf{i}h\xi/2} \sin(h\xi/2)}{\xi/2} \sum_n (u_{n+1} - u_n) e^{\mathbf{i}(n+1/2)\varepsilon\xi}.
\end{aligned}$$

It then follows that

$$\mathcal{F}[u(x) - u(x-h)] = e^{-\mathbf{i}h\xi/2} \frac{\sin(h\xi/2)}{\xi/2} \sum_n (u_{n+1} - u_n) e^{\mathbf{i}\varepsilon\xi(n+1/2)}.$$

Consequently, defining ϕ and ψ as in (6.2), we have

$$(-\Delta_\varepsilon u, v) = \frac{1}{2\pi}\int_{\mathbb{R}} \sum_{k>0} \alpha_k \Big(\frac{2\sin(k\varepsilon\xi/2)}{\varepsilon\xi}\Big)^2 \phi(\varepsilon\xi)\overline{\psi(\varepsilon\xi)}\, d\xi$$
$$= \frac{1}{2\pi\varepsilon}\int_{\mathbb{R}} \frac{\sum_{k>0} 4\alpha_k \sin^2(k\zeta/2)}{\zeta^2}\phi(\zeta)\overline{\psi(\zeta)}\, d\zeta$$

after the change of variables, $\zeta = \varepsilon\xi$. As all the functions, except ζ^2, in the integrand are 2π-periodic,

$$(-\Delta_\varepsilon u, v) = \frac{1}{2\pi\varepsilon}\int_0^{2\pi} \Big(\sum_{k>0} 4\alpha_k \sin^2\frac{k\zeta}{2}\Big)\phi(\zeta)\overline{\psi(\zeta)}\sum_\ell \frac{1}{(\zeta+2\pi\ell)^2}\,.$$

The assertion of the lemma thus follows from the identity $\sum_\ell \frac{1}{(\zeta+2\pi\ell)^2} = \frac{1}{4\sin^2(\zeta/2)}$ for all ζ and the definition of $B(\zeta)$ in (5.5). ∎

With the assumption on B we immediately have the following.

LEMMA 11. *For every $u \in \mathbf{X}$,*

$$\frac{1}{\varepsilon B_\infty}\sum_n (u_{n+1}-u_n)^2 \le \mathcal{E}_{ki}[u] \le \frac{B_\infty}{\varepsilon}\sum_n (u_{n+1}-u_n)^2. \tag{6.3}$$

7. An energy decomposition. When all α_k's, $k > 0$, are nonnegative, the energy of any nonmonotonic function can be decreased by removing the "bumps" of the function. In our current situation where some of the α_k's may be negative, we cannot use this modification. Indeed, an energy minimizer may not necessarily be monotonic. Hence, to show that an energy minimizer satisfies needed asymptotic behavior as $x \to \pm\infty$ requires special treatment.

For convenience we denote by m_0 a positive constant such that

$$f_u > m_0 \qquad \text{in } (-1-m_0, -1+m_0)\cup(1-m_0, 1+m_0). \tag{7.1}$$

Also, we denote, for every positive integer M,

$$\bar{b}_M \equiv \sqrt{\textstyle\sum_{\ell\ge M} \ell\, b_\ell^2}\,, \tag{7.2}$$

where $\{b_\ell\}$ are the Fourier coefficients of B. Clearly, $\lim_{M\to\infty}\bar{b}_M = 0$.

LEMMA 12. *Let M be any fixed positive integer. Assume that $\mathbf{E}[u] < \infty$ and $|u+1| \le m_0$ on $(-\varepsilon/2, \varepsilon(M+1/2)]$. Then for any $\eta \in (0,1)$,*

$$\mathbf{E}[u^r] + \mathbf{E}[u^l] \le \tfrac{(1+\bar{b}_M B_\infty)}{1-\eta}\mathbf{E}[u] + \tfrac{2B_\infty}{\eta(1-\eta)}\Big\{\frac{\max_{0\le n<M}|u_n+1|^2}{\varepsilon M} + \sum_{0\le n<M}\frac{(u_{n+1}-u_n)^2}{\varepsilon}\Big\},$$

where

$$u^r = -1+\theta(1+u), \quad u^l = -1+(1-\theta)(1+u), \quad \theta = \sum_{n\ge 0}\min\Big\{\frac{n}{M}\,, 1\Big\}\chi_n^\varepsilon.$$

Notice that $\theta = 0$ for $x \le \varepsilon/2$ and $\theta = 1$ for $x \ge (M-1/2)\varepsilon$. It then follows that $u^r = -1, u^l = u$ for $x \le \varepsilon/2$ and $u^r = u, u^l = -1$ for $x > (M-1/2)\varepsilon$. This lemma shows that if u is in a "resting" state for a large interval, i.e., both

$\varepsilon^{-1}\sum_{0\le n<M}|u_{n+1}-u_n|^2$ and $\max_{0\le n<M}|u_n+1|$ are small, then the energy of u can be decomposed as the sum of the energy of u^r and that of u^l. In particular, it eliminates the possibility of energy minimizers having transition layers in "remote" locations. This property will be a key in our proof of the existence of an energy minimizer with required asymptotics at $x=\pm\infty$.

Proof. First we compare the potential energy. Since $F(-1)=0$,

$$\mathcal{E}_{po}[u]-(\mathcal{E}_{po}[u^r]+\mathcal{E}_{po}[u^l])=\sum_{1\le n\le M-1}2\varepsilon\,\{F(u_n)-F(u_n^r)-F(u_n^l)\}.$$

For each $n\in\{1,\dots,M-1\}$, we can calculate, using $F(z)=\int_{-1}^{z}f(s)\,ds=\int_0^{1+z}f(s-1)\,ds$,

$$F(u_n)-F(u_n^r)-F(u_n^l)=\Big\{\int_0^{u_n+1}-\int_0^{\theta_n(u_n+1)}-\int_0^{(1-\theta_n)(u_n+1)}\Big\}f(s-1)\,ds$$
$$=\int_0^{u_n+1}\Big\{f(s-1)-\theta_n f(\theta_n s-1)-(1-\theta_n)f((1-\theta_n)s-1)\Big\}\,ds\ge 0,$$

since $f(-1)=0$, $f_u(s-1)>0$ for all $s\in(-m_0,m_0)$, and $1+u_n\in(-m_0,m_0)$. Therefore,

$$\mathcal{E}_{po}[u^r]+\mathcal{E}_{po}[u^l]\le\mathcal{E}_{po}[u].$$

Next we compare the kinetic energy. For simplicity, we denote $\phi(\zeta)=\sum_n(u_{n+1}-u_n)e^{\mathbf{i}n\zeta}$, $\phi^r(\zeta)=\sum_n(u^r_{n+1}-u^r_n)e^{\mathbf{i}n\zeta}$, and $\phi^l(\zeta)=\sum_n(u^l_{n+1}-u^l_n)e^{\mathbf{i}n\zeta}$. Since $u=u^r+u^l+1$, we have $\phi=\phi^r+\phi^l$. Consequently, by the expression for the kinetic energy in Lemma 10,

$$\mathcal{E}_{ki}[u]-\mathcal{E}_{ki}[u^r]-\mathcal{E}_{ki}[u^l]=\frac{1}{\pi\varepsilon}\int_0^{2\pi}B(\zeta)\,\mathrm{Re}\{\phi^r(\zeta)\overline{\phi^l(\zeta)}\}\,d\zeta, \tag{7.3}$$

where Re represents the real part of a complex valued function. Note that the summation for ϕ^r indeed runs only for $n\ge 0$ and that for ϕ^l for $n\le M-1$. Hence, we can write

$$B\,\mathrm{Re}\{\phi^r\overline{\phi^l}\}=\mathrm{I}+\mathrm{II}+\mathrm{III}+\mathrm{IV},\qquad\text{where}$$
$$\mathrm{I}\equiv\mathrm{Re}\Big\{B\overline{\phi^l}\sum_{0\le n<M}(u^r_{n+1}-u^r_n)e^{\mathbf{i}n\zeta}\Big\},$$
$$\mathrm{II}\equiv\mathrm{Re}\Big\{B\phi^r\sum_{0\le m<M}(u^l_{m+1}-u^l_m)e^{-\mathbf{i}m\zeta}\Big\},$$
$$\mathrm{III}\equiv-B\,\mathrm{Re}\Big\{\sum_{0\le n<M}(u^r_{n+1}-u^r_n)e^{\mathbf{i}n\zeta}\sum_{0\le m<M}(u^l_{m+1}-u^l_m)e^{-\mathbf{i}m\zeta}\Big\},$$
$$\mathrm{IV}\equiv\sum_{n\ge M}\sum_{m<0}(u_{n+1}-u_n)(u_{m+1}-u_m)B(\zeta)\cos([n-m]\zeta),$$

where in IV we have used $u^r_n=u_n$ for $n\ge M$, and similarly $u^l_m=u_m$ for $m\le 0$.

We now estimate the contribution of each of the terms. First, for any $\eta>0$,

$$\Big|\frac{1}{\pi\varepsilon}\int_0^{2\pi}\mathrm{I}\,d\zeta\Big|\le 2\Big(\frac{1}{2\pi\varepsilon}\int_0^{2\pi}B|\phi^l|^2\Big)^{1/2}\Big(\frac{B_\infty}{\varepsilon}\sum_{0\le n<M}(u^r_{n+1}-u^r_n)^2\Big)^{1/2}$$
$$\le\eta\,\mathcal{E}_{ki}[u^l]+\frac{B_\infty}{\eta\,\varepsilon}\sum_{0\le n<M}(u^r_{n+1}-u^r_n)^2\,.$$

Similarly,

$$\Big|\frac{1}{\pi\varepsilon}\int_0^{2\pi} \mathrm{II}\, d\zeta\Big| \le \eta\, \mathcal{E}_{ki}[u^r] + \frac{B_\infty}{\eta\,\varepsilon}\sum_{0\le n<M}(u^l_{n+1}-u^l_n)^2\,.$$

Next,

$$\begin{aligned}\Big|\frac{1}{\pi\varepsilon}\int_0^{2\pi} \mathrm{III}\Big| &\le \frac{2B_\infty}{\varepsilon}\Big(\sum_{0\le n<M}(u^r_{n+1}-u^r_n)^2\Big)^{1/2}\Big(\sum_{0\le m<M}(u^l_{m+1}-u^l_m)^2\Big)^{1/2}\\ &\le \frac{B_\infty}{\varepsilon}\Big\{\sum_{0\le n<M}(u^r_{n+1}-u^r_n)^2+\sum_{0\le m<M}(u^l_{m+1}-u^r_m)^2\Big\}.\end{aligned}$$

To estimate IV, we use the definition of the Fourier coefficients b_ℓ to obtain

$$\begin{aligned}\Big|\frac{1}{\pi\varepsilon}\int_0^{2\pi} \mathrm{IV}\Big| &\le \frac{1}{\varepsilon}\Big(\sum_{n\ge M}(u_{n+1}-u_n)^2\sum_{m<0}(u_{m+1}-u_m)^2\sum_{n\ge M}\sum_{m<0}b^2_{n-m}\Big)^{1/2}\\ &\le \frac{\tilde b_M}{2\varepsilon}\sum_n |u_{n+1}-u_n|^2 \le \tilde b_M B_\infty \mathcal{E}_{ki}[u],\end{aligned}$$

where

$$\tilde b_M = \Big(\sum_{n\ge M}\sum_{m<0} b^2_{n-m}\Big)^{1/2} = \Big(\sum_{\ell\ge M+1}(\ell-M)b^2_\ell\Big)^{1/2} \le \bar b_M\,.$$

Combining these estimates, we then obtain

$$\begin{aligned}\mathcal{E}_{ki}[u^r]+\mathcal{E}_{ki}[u^l]-\mathcal{E}_{ki}[u] &\le \eta\{\mathcal{E}_{ki}[u^r]+\mathcal{E}_{ki}[u^l]\}+\bar b_M B_\infty \mathcal{E}_{ki}[u]\\ &\quad+\frac{2B_\infty}{\eta\,\varepsilon}\sum_{0\le n<M}\Big\{(u^r_{n+1}-u^r_n)^2+(u^l_{n+1}-u^l_n)^2\Big\}.\end{aligned}$$

It remains to estimate the last term. We note that

$$u^r_{n+1}-u^r_n = \theta_{n+1}(u_{n+1}-u_n)+(\theta_{n+1}-\theta_n)(u_n+1)$$

and

$$u^l_{n+1}-u^l_n = (1-\theta_{n+1})(u_{n+1}-u_n)-(\theta_{n+1}-\theta_n)(u_n+1).$$

It then follows that

$$(u^r_{n+1}-u^r_n)^2+(u^l_{n+1}-u^l_n)^2 \le 2(u_{n+1}-u_n)^2+2(\theta_{n+1}-\theta_n)^2(u_n+1)^2.$$

Finally, using $\theta_{n+1}-\theta_n = 1/M$ for all $n = 0,\dots,M-1$ we obtain

$$\begin{aligned}(1-\eta)(\mathcal{E}_{ki}[u^r]+\mathcal{E}_{ki}[u^l]) &\le (1+\bar b_M B_\infty)\mathcal{E}_{ki}[u]\\ &\quad+\frac{4B_\infty}{\eta}\Big\{\frac{\max_{0\le n<M}|u_n+1|^2}{\varepsilon M}+\frac{1}{\varepsilon}\sum_{0\le n<M}(u_{n+1}-u_n)^2\Big\}.\end{aligned}$$

This completes the proof. □

For later applications, we provide an energy lower bound.

LEMMA 13. *There exists a positive constant e_0 such that $\mathbf{E}[u] \geq e_0$ if u changes sign at least once, i.e., $u_n u_{n+1} \leq 0$ for some $n \in \mathbb{Z}$, where $u = \sum_n u_n \chi_n^\varepsilon$.*

Proof. By translation if necessary, we can assume that $u_0 u_1 \leq 0$.

(i) If $\min\{|u_0|, |u_1|\} \leq 1/2$, then $\mathbf{E}[u] \geq 2\varepsilon(F(u_0)+F(u_1)) \geq 2\varepsilon \min_{|s|\leq 1/2} F(s) > 0$.

(ii) If $\min\{|u_0|, |u_1|\} > 1/2$, then since $u_0 u_1 \leq 0$, we have $|u_1 - u_0| \geq 1/2$. It then follows from (6.3) that $\mathbf{E}(u) \geq (B_\infty \varepsilon)^{-1} |u_1 - u_0|^2 \geq 1/(4B_\infty \varepsilon)$.

Combining both cases we then obtain the assertion of the lemma with

$$e_0 = \min\left\{ 2\varepsilon \min_{|s|\leq 1/2}\{F(s)\},\ 1/(4B_\infty\varepsilon) \right\}. \qquad \square$$

8. Existence of a stationary wave. With all these preparations, we can now prove existence of a solution.

THEOREM 14. *For every $\varepsilon > 0$, there exists at least one function $u^\varepsilon \in \mathbf{X}$ such that*

$$\mathbf{E}[u^\varepsilon] = \mathrm{E}(\varepsilon) \equiv \inf_{u\in\mathbf{X}} \mathbf{E}[u]. \tag{8.1}$$

Consequently, problem (5.2) *or* (5.3) *admits at least one solution.*

Proof. Note that $\mathrm{E}(\varepsilon)$ is finite since $\mathbf{E}[u] < \infty$ for the test function u defined by $u = 1$ for $x > \varepsilon/2$ and $u = -1$ for $x \leq \varepsilon/2$. Hence, there exists a sequence $\{u^j\}_{j=1}^\infty$ in $\mathbf{X}$ such that as $j \to \infty$, $\mathbf{E}[u^j] \searrow \mathrm{E}(\varepsilon)$. We write $u^j = \sum_n u_n^j \chi_n^\varepsilon$. By a translation if necessary, we can assume that $x = \varepsilon/2$ is the first place where u^j changes sign; i.e., $u_n^j < 0$ for all $n \leq 0$, and $u_1^j \geq 0$.

Clearly $\{u_n^j\}$ are uniformly bounded for all n and j, since for each n and j, $2F(u_n^j) \leq \frac{1}{\varepsilon}\mathbf{E}[u^j] \leq \frac{1}{\varepsilon}\mathrm{E}(\varepsilon) + 1$, and $\lim_{|s|\to\infty} F(s) = \infty$. Thus, by a diagonal limit process, we can extract a subsequence of $\{u^j\}$, which we still denote by $\{u^j\}$, such that for every n, $\lim_{j\to\infty} u_n^j = u_n^\varepsilon$ for some $u_n^\varepsilon \in \mathbb{R}$. Set $u^\varepsilon = \sum_n u_n^\varepsilon \chi_n^\varepsilon$. We claim that $u^\varepsilon \in \mathbf{X}$ and u^ε is an energy minimizer.

First, we consider the potential energy of u^ε. We have

$$\begin{aligned}\sum_n F(u_n^\varepsilon) &= \lim_{N\to\infty} \sum_{|n|\leq N} F(u_n^\varepsilon) \\ &= \lim_{N\to\infty} \liminf_{j\to\infty} \sum_{|n|\leq N} F(u_n^j) \leq \liminf_{j\to\infty} \sum_n F(u_n^j).\end{aligned}$$

Next we consider the kinetic energy. Set $\phi^j(\zeta) = \sum_n (u_{n+1}^j - u_n^j) e^{\mathbf{i}\, n\zeta}$. By (6.3),

$$\|\phi^j\|^2_{L^2((0,2\pi))} = 2\pi \sum_n |u_{n+1}^j - u_n^j|^2 \leq 2\pi\varepsilon B_\infty\, \mathbf{E}[u^j].$$

Thus, $\{\phi^j(\cdot)\}_{j=1}^\infty$ is a bounded sequence in $L^2((0,2\pi))$, and we can extract a subsequence, which we still denote by $\{\phi^j\}$, such that for some $\phi^\varepsilon \in L^2((0,2\pi))$, as $j \to \infty$,

$$\phi^j \longrightarrow \phi^\varepsilon \qquad \text{weakly in } \ L^2((0,2\pi)).$$

As weak convergence in $L^2((0,2\pi))$ implies the convergence of the Fourier coefficients, it follows that

$$\phi^\varepsilon = \sum_n (u_{n+1}^\varepsilon - u_n^\varepsilon) e^{\mathbf{i} n\zeta}.$$

Using $|\phi^\varepsilon|^2 = |\phi^j|^2 - |\phi^\varepsilon - \phi^j|^2 + 2\mathrm{Re}\{\phi^\varepsilon\overline{(\phi^\varepsilon - \phi^j)}\} \le |\phi^j|^2 + 2\mathrm{Re}\{\phi^\varepsilon\overline{(\phi^\varepsilon - \phi^j)}\}$, we conclude that, since $B(\zeta)$ is bounded and nonnegative,

$$\int_0^{2\pi} B(\zeta)|\phi^\varepsilon(\zeta)|^2 \le \liminf_{j\to\infty} \int_0^{2\pi} B(\zeta)|\phi^j(\zeta)|^2.$$

Therefore,

$$\mathbf{E}[u^\varepsilon] \le \liminf_{j\to\infty} \mathbf{E}[u^j] = \mathrm{E}(\varepsilon).$$

It remains to show that $u^\varepsilon \in \mathbf{X}$.

First of all, the boundedness of $\mathcal{E}_{po}[u^\varepsilon] = 2\varepsilon\sum_n F(u_n^\varepsilon)$ implies that $\sum_n |(u_n^\varepsilon)^2 - 1|^2 < \infty$. Consequently,

$$\lim_{|n|\to\infty} (u_n^\varepsilon)^2 = 1. \tag{8.2}$$

As $u_n^j < 0$ for all $n \le 0$ and all j, we conclude that $u_n^\varepsilon \le 0$ for all $n \le 0$, so that $\lim_{n\to-\infty} u_n^\varepsilon = -1$, and $\sum_{n\le 0} |u_n^\varepsilon + 1|^2 < \infty$.

It remains to show that $\lim_{n\to\infty} u_n^\varepsilon = 1$.

The finiteness of $\|\phi^\varepsilon\|^2_{L^2((0,2\pi))} = 2\pi\sum_n |u_{n+1}^\varepsilon - u_n^\varepsilon|^2$ implies that there are only finitely many n where $|u_{n+1} - u_n| > 1$. Consequently, from (8.2), either $\lim_{n\to\infty} u_n^\varepsilon = 1$, which concludes our proof, or $\lim_{n\to\infty} u_n^\varepsilon = -1$, which we show below is impossible.

Suppose, on the contrary, that $\lim_{n\to\infty} u_n^\varepsilon = -1$. Let $\eta \in (0,1)$ be any fixed number, and let M be any fixed positive integer. For any $\hat{n}$ sufficiently large, applying Lemma 12 (with the origin at point $\varepsilon\hat{n}$) to u^j and using $|u_{n+1}^j - u_n^j| \le |u_{n+1}^j + 1| + |u_n^j + 1|$, we obtain

$$\mathbf{E}[u^{jr}] + \mathbf{E}[u^{jl}] \le \frac{1 + \bar{b}_M B_\infty}{1-\eta}\mathbf{E}[u^j] + C(\varepsilon,\eta,M) \max_{\hat{n}\le n\le \hat{n}+M} |u_n^j + 1|^2,$$

where $u^{jr} = u, u^{jl} = -1$ for $x \ge (\hat{n} + M + 1/2)\varepsilon$ and $u^{jr} = -1, u^{jl} = u$ for $x \le (\hat{n} - 1/2)\varepsilon$.

Note that $u^{jr} \in \mathbf{X}$, so that $\mathbf{E}[u^{jr}] \ge \mathrm{E}(\varepsilon)$. Also u^{jl} experiences a sign change, so, by Lemma 13, $\mathbf{E}[u^{jl}] \ge e_0$. Thus,

$$e_0 + \mathrm{E}(\varepsilon) \le \frac{1 + \bar{b}_M B_\infty}{1-\eta}\mathbf{E}[u^j] + C(\varepsilon,\eta,M) \max_{\hat{n}\le n\le \hat{n}+M} |u_n^j + 1|^2.$$

Sending j to ∞ then gives

$$e_0 + \mathrm{E}(\varepsilon) \le \frac{1 + \bar{b}_M B_\infty}{1-\eta}\mathrm{E}(\varepsilon) + C(\varepsilon,\eta,M) \max_{\hat{n}\le n\le \hat{n}+M} |u_n^\varepsilon + 1|^2.$$

Sending $\hat{n}$ to ∞ and using $\lim_{n\to\infty} u_n^\varepsilon = -1$, we further obtain $e_0 + \mathrm{E}(\varepsilon) \le \frac{1+\bar{b}_M B_\infty}{1-\eta}\mathrm{E}(\varepsilon)$. Finally, sending $M \to \infty$ first and then $\eta \to 0$, we obtain $e_0 + \mathrm{E}(\varepsilon) \le \mathrm{E}(\varepsilon)$, a contradiction. This contradiction shows that $u^\varepsilon \in \mathbf{X}$. This concludes the proof. ∎

9. Convergence of minimizers to u_0 as $\varepsilon \to 0$. In this section, we establish the convergence part of Theorem 2.

THEOREM 15. *Let $\{u^\varepsilon\}_{\varepsilon>0}$ be the energy minimizing solutions to (5.3) obtained above, translated so that $x = \varepsilon/2$ is the first place where $u^\varepsilon = \sum_n u_n^\varepsilon \chi_n^\varepsilon$ experiences a sign change; i.e., $u_n^\varepsilon < 0$ for all $n \le 0$ and $u_1^\varepsilon \ge 0$. Then*

$$\lim_{\varepsilon \searrow 0} u_\varepsilon = u_0 \qquad \text{in } \ L^\infty(\mathbb{R}),$$
$$\lim_{\varepsilon \searrow 0} (\tilde{u}_\varepsilon - u_0) = 0 \qquad \text{in } \ H^1(\mathbb{R}),$$

where $\tilde{u}^\varepsilon$ is the "companion" of u^ε obtained by a linear interpolation of the node values of u^ε at εn, $n \in \mathbb{Z}$:

$$\tilde{u}^\varepsilon(x) = \sum_n \left\{ u_n^\varepsilon + \frac{(x - \varepsilon n)}{\varepsilon}(u_{n+1}^\varepsilon - u_n^\varepsilon) \right\} \chi_n^\varepsilon(x - \varepsilon/2).$$

To prove the theorem, we need an upper bound on the minimum energy.

LEMMA 16. *Let $\mathrm{E}(\varepsilon) = \mathbf{E}[u^\varepsilon]$ be the minimum of the energy. Then*

$$\limsup_{\varepsilon \searrow 0} \mathrm{E}(\varepsilon) \le \mathrm{E}(0) \equiv \int_{\mathbb{R}} \left\{ \tfrac{1}{2}(u_0')^2 + F(u_0) \right\} dx\,. \tag{9.1}$$

The proof follows by taking u_0 as a test function and using the fact that $\Delta_\varepsilon u_0 - u_0'' \to 0$ in $L^2(\mathbb{R})$. We omit the details.

Proof of Theorem 15. Note that $\tilde{u}^\varepsilon$ is Lipschitz continuous. Also

$$\|\tilde{u}^{\varepsilon\prime}\|_{L^2(\mathbb{R})}^2 = \varepsilon \sum_n \left(\frac{u_{n+1} - u_n}{\varepsilon} \right)^2 \le B_\infty \mathrm{E}(\varepsilon)$$

is uniformly bounded for all $\varepsilon \in (0, 1]$. In addition, $\tilde{u}^\varepsilon$ and u^ε are close in $L^2(\mathbb{R})$ since

$$\|u^\varepsilon - \tilde{u}^\varepsilon\|_{L^2}^2 = \frac{\varepsilon}{12} \sum_n (u_{n+1} - u_n)^2 \le \frac{\varepsilon^2 B_\infty}{12} \mathrm{E}(\varepsilon).$$

Hence, we can interchange $\tilde{u}^\epsilon$ and u^ϵ freely.

Since

$$2 \int_{\mathbb{R}} F(u^\varepsilon) = 2\varepsilon \sum_n F(u_n^\varepsilon) < \mathrm{E}(\varepsilon),$$

we see that $\{\tilde{u}^\varepsilon\}_{0<\varepsilon<1}$ is a bounded family in $H^1((-R, R))$ for any $R > 1$. Hence, there exists a sequence $\{\varepsilon_j\}_{j=1}^\infty$ and a function $u^0 \in H^1_{\mathrm{loc}}(\mathbb{R})$ such that as $j \to \infty$, $\varepsilon_j \to 0$ and

$$\tilde{u}^{\varepsilon_j} - u^0 \longrightarrow 0 \quad \text{weakly in } H^1(\mathbb{R}) \text{ and uniformly in } C^0([-R, R]) \text{ for every } R > 1.$$

Consequently,

$$u^\varepsilon \longrightarrow u^0 \quad \text{ in } L^\infty((-R, R)) \text{ for any } R > 1.$$

Taking any test function $\zeta(\cdot) \in C_0^\infty(\mathbb{R})$, we obtain by passing to the limit in the identity $(u^\varepsilon, -\Delta_\varepsilon \zeta) + (f(u^\varepsilon), \zeta) = 0$ that $(u^0, -\zeta'') + (f(u^0), \zeta) = 0$. That is, u^0 is a weak solution to $-(u^0)'' + f(u^0) = 0$, and consequently, since $(u^0)' \in L^2(\mathbb{R})$, u^0 is a classical solution to $-(u^0)'' + f(u^0) = 0$.

Since $u^\varepsilon \le 0$ for all $x \le 0$, we have $u^0 \le 0$ for all $x \le 0$. Also, from $u^\varepsilon(\varepsilon/2) \ge 0$ we see that $u^0(0) = 0$. Note that u_0 is the only solution to $-u'' + f(u) = 0$ with the property that (1) $u(0) = 0$, (2) $u \le 0$ for $x < 0$, and (3) $\int_{\mathbb{R}} F(u)dx < \infty$. We then conclude that $u^0 = u_0$. The uniqueness of the limit then implies that the whole family $\{u^\varepsilon\}$ converges to u_0 as $\varepsilon \searrow 0$.

Next we show that $u^\varepsilon \to u_0$ in $L^\infty(\mathbb{R})$. It suffices to show that $u^\varepsilon(x) \to \pm 1$ as $x \to \pm\infty$, uniformly in ε. By symmetry, we need only consider $x \to -\infty$.

For this purpose, let $J \gg 1$ be any large integer. First of all, there exists a small positive ε_1 such that for all $\varepsilon \in (0, \varepsilon_1]$, $|u^\varepsilon - u_0| \le 1/J$ on the interval $[-2J, 0]$. Consequently, $|u^\varepsilon + 1| \le 2/J \le m_0$ on $[-2J, -J]$, since u_0 approximates -1 exponentially fast as $x \to -\infty$.

Now consider the interval $x \in [-2J, -J]$. There exists an interval $[x^\varepsilon, x^\varepsilon + 1] \subset [-2J, -J]$ such that

$$\int_{x^\varepsilon}^{x^\varepsilon+1} |\tilde{u}^{\varepsilon\prime}|^2 \le \frac{1}{J}\int_{-2J}^{-J} |\tilde{u}^{\varepsilon\prime}|^2 \le \frac{B_\infty}{J}\mathbf{E}[u^\varepsilon].$$

That is,

$$\sum_{n\in[x^\varepsilon/\varepsilon,(x^\varepsilon+1)/\varepsilon]} \varepsilon\Big|\frac{u^\varepsilon_{n+1} - u^\varepsilon_n}{\varepsilon}\Big|^2 \le \frac{B_\infty \mathbf{E}[u^\varepsilon]}{J}.$$

Consequently, taking $\eta = 1/\sqrt{J}$ and $M = 1/\varepsilon$ (we can assume that $1/\varepsilon$ is an integer), we obtain from the energy decomposition, Lemma 12 (with origin at x^ε), that

$$\begin{aligned}(1 - J^{-1/2})(\mathbf{E}[u^{\varepsilon r}] + \mathbf{E}[u^{\varepsilon l}]) &\le (1 + \bar{b}_{1/\varepsilon}B_\infty)\mathbf{E}[u^\varepsilon] \\ &\quad + 2\sqrt{J}B_\infty\left\{\max_{x\in[x^\varepsilon,x^\varepsilon+1]}|u^\varepsilon + 1|^2 + \frac{1}{\varepsilon}\sum_{n\in[x^\varepsilon/\varepsilon,(x^\varepsilon+1)/\varepsilon]}|u^\varepsilon_{n+1} - u^\varepsilon_n|^2\right\} \\ &\le (1 + \bar{b}_{1/\varepsilon}B_\infty)\mathbf{E}[u^\varepsilon] + C/\sqrt{J}\,,\end{aligned}$$

where C is independent of $\varepsilon \in (0, \varepsilon_1]$ and J. As $u^{\varepsilon r} \in \mathbf{X}$, $\mathbf{E}[u^{\varepsilon r}] \ge \mathbf{E}[u^\varepsilon]$. It then follows that

$$\mathbf{E}[u^{\varepsilon l}] \le C(1/\sqrt{J} + \bar{b}_{1/\varepsilon}).$$

Note that $u^\varepsilon = u^{\varepsilon l}$ for $x \le x^\varepsilon$. We have

$$\begin{aligned}\mathbf{E}[u^{\varepsilon l}] &\ge \sum_{n\le x^\varepsilon/\varepsilon}\Big\{\frac{1}{B_\infty\varepsilon}|u_{n+1} - u_n|^2 + 2\varepsilon F(u^\varepsilon)\Big\} \\ &\ge \frac{2\sqrt{2}}{\sqrt{B_\infty}}\sum_{n\le x^\varepsilon/\varepsilon}|u_{n+1} - u_n|\sqrt{F(u_\varepsilon)} \\ &\ge \frac{2\sqrt{2}}{\sqrt{B_\infty}}\Big|\int_{-1}^{u^\varepsilon(x)}\sqrt{F(s)}ds - O(\sqrt{\varepsilon})\Big|\end{aligned}$$

for any $x \le x^\varepsilon$, since $\sum_{n\le m}|u^\varepsilon_{n+1} - u^\varepsilon_n|\sqrt{F(u^\varepsilon)}$ is not less than a Riemann sum for the integral $|\int_{-1}^{u^\varepsilon_m}\sqrt{F(s)}ds|$ with mesh size no bigger than $\sup_n|u^\varepsilon_{n+1} - u^\varepsilon_n| = O(\sqrt{\varepsilon})$.

We obtain, for all $x \le -2J$ and $\varepsilon \in (0, \varepsilon_1]$,

$$\Big| \int_{-1}^{u^\varepsilon(x)} \sqrt{F(s)} ds \Big| \le C(1/\sqrt{J} + \sqrt{\varepsilon} + \bar{b}_{1/\varepsilon}),$$

where C is independent of J and ε. This inequality implies that $u^\varepsilon \to -1$ as $x \to -\infty$, uniformly in ε. Thus, $u^\varepsilon \to u_0$ uniformly on $\mathbb{R}$.

Finally, we show that $\tilde{u}^\varepsilon - u_0 \to 0$ in $H^1(\mathbb{R})$. First of all, we have

$$(-\Delta_\varepsilon(u^\varepsilon - u_0), u^\varepsilon - u_0) + (f(u^\varepsilon) - f(u_0), u^\varepsilon - u_0) = (\Delta_\varepsilon u_0 - u_0'', u^\varepsilon - u_0).$$

Here, for simplicity, we shall not distinguish between the functions $u_0, u_0^\varepsilon \equiv \sum_n u_0(n\varepsilon)\chi_n^\varepsilon$, and $\tilde{u}_0^\varepsilon = \sum_n \{u_0(n\varepsilon) + [x/\varepsilon - n][u(\varepsilon n + \varepsilon) - u_0(n\varepsilon)]\}\chi_n^\varepsilon(x - \varepsilon/2)$, since with a small error added, one can change from one to the other.

Using (6.3), we have

$$\begin{aligned}(-\Delta_\varepsilon(u^\varepsilon - u_0), u^\varepsilon - u_0) &\ge \frac{1}{B_\infty\, \varepsilon} \sum_n |(u_{n+1}^\varepsilon - u_n^\varepsilon) - (u_{0(n+1)} - u_{0n})|^2 \\ &= \frac{1}{B_\infty} \int_{\mathbb{R}} |\tilde{u}^{\varepsilon\prime} - u_0'|^2\, dx.\end{aligned}$$

Let m_0 be as in (7.1). Then by the $L^\infty(\mathbb{R})$ convergence of u^ε to u_0, there is a positive constant $M > 0$ such that $\{f(u^\varepsilon) - f(u_0)\}\{u^\varepsilon - u_0\} \ge m_0(u^\varepsilon - u_0)^2$ for all $|x| \ge M$ and small enough ε. Thus,

$$(f(u^\varepsilon) - f(u_0), u^\varepsilon - u^0) \ge m_0 \|u^\varepsilon - u_0\|^2 - (m_0 + \|f_u\|_{L^\infty}) \int_{|x|<M} |u^\varepsilon - u_0|^2.$$

In conclusion, for all small enough ε,

$$\frac{1}{B_\infty} \|\tilde{u}^{\varepsilon\prime} - u_0'\|_{L^2}^2 + m_0 \|u^\varepsilon - u_0\|_{L^2}^2 \le C\|u^\varepsilon - u_0\|_{L^2(-M,M)}^2 + |(\Delta_\varepsilon u_0 - u_0'', u^\varepsilon - u_0)|.$$

Sending $\varepsilon \to 0$ and using the inequality $|(\Delta_\varepsilon u_0 - u_0'', u^\varepsilon - u_0)| \le \|\Delta_\varepsilon u_0 - u_0''\|_{L^1(\mathbb{R})} \|u^\varepsilon - u_0\|_{L^\infty}$, we conclude that $\tilde{u}^\varepsilon - u^0 \to 0$ in $H^1(\mathbb{R})$. This completes the proof. ∎

10. Nonuniqueness of stationary wave for small ε. In this section, we show that for small ε, (5.3) admits another solution, which is close to, but different from, any translation of that solution obtained in the previous sections via the energy minimization technique.

10.1. An eigenvalue estimate. First we investigate the operator

$$\mathcal{L}_\varepsilon \phi \equiv -\Delta_\varepsilon \phi + f_u(u_0)\phi \tag{10.1}$$

for functions which are constant on every interval $(\varepsilon(n - 1/2), \varepsilon(n + 1/2)]$, $n \in \mathbb{Z}$, lying in the space

$$\mathbf{X}_0 \equiv \left\{ \phi = \sum_n \phi_n \chi_n^\varepsilon \ : \ \sum_n \phi_n^2 < \infty \right\}.$$

We define

$$\begin{aligned}\Lambda(\varepsilon) &\equiv \inf_{\phi \in \mathbf{X}_0, \|\phi\|_{L^2} = 1, \phi(0) = 0} \Big\{ \| - \Delta_\varepsilon \phi + f_u(u_0)\phi\|_{L^2(\mathbb{R} \setminus (-\varepsilon/2, \varepsilon/2])} \Big\}, \\ \Lambda_0 &\equiv \liminf_{\varepsilon \searrow 0} \Lambda(\varepsilon)\ .\end{aligned}$$

LEMMA 17. $\Lambda_0 > 0$. *Consequently,* $\Lambda(\varepsilon) > \Lambda_0/2$ *for all small positive* ε.

Proof. We use the same idea as that for the nonbalanced potential case. Let $\{\varepsilon_j, \phi_j, \psi_j\}$ be a sequence such that $\lim_{j\to\infty} \varepsilon_j = 0$, $\lim_{j\to\infty} \|\psi_j\|_{L^2(\mathbb{R}\setminus(-\varepsilon_j/2,\varepsilon_j/2])} = \Lambda_0$, and for each $j \geq 1$, $\varepsilon_j > 0, \phi_j \in \mathbf{X}_0, \|\phi_j\|_{L^2} = 1$, $\phi_j = 0$ on $(-\varepsilon_j/2, \varepsilon_j/2]$, and $\psi_j = -\Delta_{\varepsilon_j}\phi_j + f_u(u_0)\phi_j$.

Using the identity $(-\Delta_{\varepsilon_j}\phi_j, \phi_j) + (f_u(u_0)\phi_j, \phi_j) = (\psi_j, \phi_j)$, that $\phi_j = 0$ on $(-\varepsilon_j/2, \varepsilon_j/2]$, and Lemma 11, we obtain

$$\frac{1}{B_\infty}\|\tilde{\phi}_j'\|^2 + \int_{\mathbb{R}} f(u_0)\phi_j^2 \leq \|\phi_j\|_{L^2}\|\psi_j\|_{L^2(\mathbb{R}\setminus(-\varepsilon_j/2,\varepsilon_j/2])},$$

where $\tilde{\phi}_j$ is the linear interpolant of ϕ at node points. It then follows that $\|\tilde{\phi}'\|_{L^2}$ is uniformly bounded. Consequently, $\|\phi_j - \tilde{\phi}_j\|_{L^2(\mathbb{R})}$ is of size $O(\varepsilon_j^2)$.

Thus, we can select a subsequence from $\{\tilde{\phi}_j, \psi_j\}$, still denoted by $\{\tilde{\phi}_j, \psi_j\}$, such that for some $\phi \in H^1(\mathbb{R})$ and $\psi \in L^2(\mathbb{R})$,

$$\begin{aligned} \tilde{\phi}_j &\longrightarrow \phi \qquad \text{in } L^2_{\text{loc}}(\mathbb{R}) \text{ and weakly in } H^1(\mathbb{R}), \\ \psi_j &\longrightarrow \psi \qquad \text{weakly in } L^2_{\text{loc}}(\mathbb{R}\setminus\{0\}). \end{aligned}$$

In addition, $\phi(0) = 0$. In the weak formulation, one can show that $-\phi'' + f(u_0)\phi = \psi$ in $\mathbb{R}\setminus\{0\}$. By an estimate similar to (3.10), (3.11), we conclude that $\Lambda_0 > 0$. This completes the proof. ∎

10.2. Energy minimizer with constraint. For every $\alpha \in (-1, 1)$ we define

$$\mathbf{X}_\alpha \equiv \left\{ u = \sum_n u_n \chi_n^\varepsilon \;:\; u_0 = \alpha,\ \sum_{n>0} |1 - u_n|^2 + \sum_{n<0} |1 + u_n|^2 < \infty \right\}.$$

Define

$$E(\alpha, \varepsilon) \equiv \inf_{u\in\mathbf{X}_\alpha} \mathbf{E}[u], \quad \mathrm{E}(\varepsilon) = \inf_{\alpha\in(-1,1)} E(\alpha, \varepsilon).$$

We note that $\mathrm{E}(\varepsilon)$ is the energy of the minimizer we studied in the previous sections.

LEMMA 18. *There exists* $\varepsilon_0 > 0$ *such that for every* $\alpha \in [-1/2, 1/2]$ *and* $\varepsilon \in (0, \varepsilon_0]$, *there exists* $u_\varepsilon^\alpha \in \mathbf{X}_\alpha$ *such that* $\mathbf{E}[u_\varepsilon^\alpha] = E(\alpha, \varepsilon)$.

Proof. The proof follows the same lines as that of Theorem 8. We can extract a (weak) limit from a minimizing sequence. This limit is a minimizer if it is in $\mathbf{X}_\alpha$. To show this, we follow the same contradiction argument as before, obtaining the inequality $\tilde{e}_0 + \mathrm{E}(\varepsilon) \leq E(\alpha, \varepsilon)$, where $\tilde{e}_0$ is the minimum energy among all functions which take the value α at the origin. To contradict $\tilde{e}_0 + \mathrm{E}(\varepsilon) \leq E(\alpha, \varepsilon)$, we have to assume that ε is small. In fact, for small ε, $\mathrm{E}(\varepsilon) = \mathrm{E}(0) + o(1)$ (Theorem 14), and $\tilde{e}_0 > C(\min\{\int_{-1}^{\alpha} \sqrt{F(s)}ds, \int_{\alpha}^{1} \sqrt{F(s)}ds\} - O(\sqrt{\varepsilon}))$. Also, taking an appropriate test function shows that $E(\alpha, \varepsilon) \leq \mathrm{E}(0) + o(1)$. Therefore, $\tilde{e}_0 + E(\varepsilon) > E(\alpha, \varepsilon)$ for all small positive ε and all $\alpha \in [-1/2, 1/2]$. This inequality shows that the limit of the minimizer has the correct asymptotics, thereby establishing the existence of a minimizer u_ε^α. We omit the details. ∎

LEMMA 19. *For every* $\delta > 0$, *there exists* $\varepsilon_1(\delta) > 0$ *such that if* $\varepsilon \in (0, \varepsilon_1(\delta)], \alpha \in [-1/2, 1/2]$ *and* $u_\varepsilon^\alpha \in \mathbf{X}_\alpha$ *is a minimizer of* $\mathbf{E}[u]$ *in* $\mathbf{X}_\alpha$, *then* $\|u_\varepsilon^\alpha - u_0(z(\alpha) + \cdot)\|_{L^2\cap L^\infty} \leq \delta$, *where* $z(\alpha)$ *is the point satisfying* $u_0(z(\alpha)) = \alpha$.

The proof follows the same lines as that of Theorem 14 and is omitted.

LEMMA 20. *There exists $\varepsilon_2 > 0$ such that for every $\varepsilon \in (0, \varepsilon_2]$ and $\alpha \in [-1/2, 1/2]$, the minimizer $u_\varepsilon^\alpha \in \mathbf{X}_\alpha$ for $\mathbf{E}[u]$ in $\mathbf{X}_\alpha$ is unique.*

Proof. Let $u_\epsilon^{\alpha,1}$ and $u_\varepsilon^{\alpha,2}$ be two minimizers. Set $\phi = u_\varepsilon^{\alpha,1} - u_\varepsilon^{\alpha,2}$. Then ϕ satisfies the equation

$$-\Delta_\varepsilon \phi + f_u(u_0(\bar{z}(\alpha) + x))\phi = r(x, \phi(x))\phi \quad \text{for all } x \in \mathbb{R} \setminus (-\varepsilon/2, \varepsilon/2],$$

where $r(x, \phi(x)) \equiv \int_0^1 [f_u(u^{\alpha,1}(x) + s\phi(x)) - f_u(u_0(\bar{z}(\alpha) + x))]ds$ and $\bar{z}(\alpha)$ is an integer multiple of ε with $|z(\alpha) - \bar{z}(\alpha)| \le \varepsilon/2$. In view of the previous lemma, we see that when $\varepsilon \le \varepsilon_1(\delta)$, $|r(x, \phi(x))| \le C\delta$, where C is a constant depending only on f. Hence, by the eigenvalue estimate in Lemma 17,

$$\|\phi\|_{L^2} \le \frac{\|r(\cdot, \phi(\cdot))\|_{L^\infty}}{\Lambda(\varepsilon)} \|\phi\|_{L^2(\mathbb{R}\setminus(-\varepsilon/2, \varepsilon/2])}.$$

From this we conclude that $\phi = 0$ if we take $\delta = \Lambda_0/(4C)$ and let ε be small enough such that $\Lambda(\varepsilon) \ge \Lambda_0/2$. This completes the proof. ∎

LEMMA 21. *For every $\varepsilon \in (0, \varepsilon_2]$, $E(\alpha, \varepsilon)$ is continuously differentiable in $\alpha \in (-1/2, 1/2)$ and*

$$\frac{d}{d\alpha} E(\alpha, \varepsilon) = 2\varepsilon\{-\Delta_\varepsilon u_\varepsilon^\alpha + f(u_\varepsilon^\alpha)\}\Big|_{x=0}. \tag{10.2}$$

Consequently, u_ε^α solves (5.2) if and only if $\frac{d}{d\alpha} E(\alpha, \varepsilon) = 0$.

Proof. Denote by $L(\alpha, \varepsilon)$ the right-hand side of (10.2). Since the uniqueness of u_ε^α implies the continuity of u_ε^α in α, $L(\alpha, \varepsilon)$ is continuous in $\alpha \in (-1/2, 1/2)$.

For any small positive h, we have

$$\begin{aligned} E(\alpha + h, \varepsilon) &\le \mathbf{E}[u_\varepsilon^\alpha + h\chi_\varepsilon^0] \\ &= E(\alpha, \varepsilon) + L(\alpha, \varepsilon)h + \varepsilon(-\Delta_\varepsilon \chi_\varepsilon^0, \chi_\varepsilon^0)h^2 + 2\varepsilon\{F(\alpha + h) - F(\alpha) - f(\alpha)h\}. \end{aligned}$$

Similarly, we have

$$E(\alpha, \varepsilon) \le E(\alpha + h, \varepsilon) - L(\alpha + h, \varepsilon)h + O(h^2).$$

It then follows that, for all small positive h,

$$L(\alpha + h, \varepsilon) + O(h) \le \frac{E(\alpha + h, \varepsilon) - E(\alpha, \varepsilon)}{h} \le L(\alpha, \varepsilon) + O(h). \tag{10.3}$$

The assertion (10.2) thus follows by letting $h \searrow 0$.

Note that the minimizer u_ε^α in $\mathbf{X}_\alpha$ satisfies $-\Delta_\varepsilon u_\varepsilon^\alpha + f(u_\varepsilon^\alpha) = 0$ in $\mathbb{R} \setminus (-\varepsilon/2, \varepsilon/2]$. Hence, u_ε^α solves (5.2) if and only if $\frac{d}{d\alpha} E(\alpha, \varepsilon) = 0$. ∎

The final part of Theorem 2 is that which gives a second stationary wave for $\varepsilon > 0$ sufficiently small.

THEOREM 22. *For every $\varepsilon \in (0, \varepsilon_2]$, problem (5.3) admits at least two solutions, u_ε^1 and u_ε^2, which differ by more than translation, and as $\varepsilon \to 0$, $\|u_\varepsilon^i - u_0\|_{L^\infty(\mathbb{R})} \to 0$, $i = 1, 2$.*

Proof. Let u_ε be one of the global minimizers given in Theorem 14. We write $u_\varepsilon = \sum_n u_n^\varepsilon \chi_n^\varepsilon$. Let $(a, b) \subset [-1/2, 1/2]$ be an interval such that for some integers n_1 and n_2, $a = u_{n_1}^\varepsilon < b = u_{n_2}^\varepsilon$.

Consider the differentiable function $E(\alpha, \varepsilon)$ for $\alpha \in [a, b]$.

If $E(\alpha,\varepsilon)$ is a constant function, then every u_ε^α is a solution to (5.2), and hence we have a continuum of solutions to (5.3).

If $E(\alpha,\varepsilon)$ is not a constant function, then as it attains the global minimum $\mathrm{E}(\varepsilon)$ at $\alpha = a$ and $\alpha = b$, there exists at least a local maximum of $E(\cdot,\varepsilon)$ attained at some $c \in (a,b)$, at which $\frac{d}{d\alpha}E(c,\varepsilon) = 0$. Consequently, the local saddle $u_\varepsilon^c \in \mathbf{X}_c$ is a solution to (5.3). Clearly, u_ε^c is different from any translation of u_ε since their energies are different. Translating u_ε^c and using Lemma 19 completes the proof. □

Remark 8. (1) We do not know if (5.2) admits a continuous solution. If this is true, then $E(\alpha,\varepsilon)$ is simply a constant function and (5.2) has a continuum of truly different solutions, even though they are obtained by translating the underlying continuous profile. It is shown in [14], for example, that starting with a particular wave profile, such as $u(x) = \tanh(\mu x)$, one can reverse engineer a nonlinearity so that this is a traveling wave for a discrete bistable equation with nearest neighbor coupling. For the profile given here, the equation is

$$cu'(x) = \frac{1}{\varepsilon^2}[u(x+1) - 2u(x) + u(x-1)] - c\mu(1-u(x)^2) - \frac{2}{\varepsilon^2}\frac{u(x)}{1+(1-u(x)^2)\sinh^2\mu} + \frac{2u(x)}{\varepsilon^2}.$$

(2) If $f(u)$ is odd and α_k decays to zero exponentially fast in k (e.g., only finitely many nonzero α_k's), one can show that $E(\alpha,\varepsilon) - E(\varepsilon)$ is small beyond any power of ε. The idea is as follows.

Write, as an asymptotic approximation,

$$\Delta_\varepsilon \sim \frac{d^2}{dx^2} + \sum_{n=2}^{\infty} \varepsilon^{2n} a_n \frac{d^{2n}}{dx^{2n}}\ .$$

One looks for an asymptotic solution

$$u_\varepsilon \sim u_0(x) + \sum_{n=1}^{\infty} \varepsilon^n U_n(x), \qquad \text{with } U_n(x) = -U_n(-x),$$

to the equation

$$-u_\varepsilon'' + f(u_\varepsilon) - \sum_{n=2}^{\infty} \varepsilon^{2n} a_n \frac{d^{2n}}{dx^{2n}} u_\varepsilon \sim 0.$$

One can show that all U_n, $n = 1, 2, \dots$, exist, and are uniquely determined, by using the symmetry assumption. For more details, see Chen et al. [5].

Now fix any order of approximation, K, desired. Set $u_\varepsilon^K = u_0 + \sum_{n=1}^K \varepsilon^k U_k(x)$. Then

$$-\Delta u_\varepsilon^K + f(u_\varepsilon^K) = O(\varepsilon^{K+1}) \qquad \text{on } \mathbb{R}.$$

For any $\alpha \in [-1/2, 1/2]$, let $z = z^{\varepsilon,K}(\alpha)$ be such that $u_\varepsilon^K(z) = \alpha$. Define $u_\varepsilon^{K,\alpha}(x) = \sum_n u_\varepsilon^K(z + \varepsilon n)\chi_n^\varepsilon$.

Consider the function $\tilde{E}(\alpha,\varepsilon) = \mathbf{E}[u_\varepsilon^{K,\alpha}]$. Since $u_\varepsilon^{K,\alpha}$ satisfies (5.3) up to order ε^{K+1}, one can show that $\frac{d}{d\alpha}\tilde{E}(\alpha,\varepsilon) = O(\varepsilon^{K+1})$ and hence is almost independent of α. On the other hand, following our proof of uniqueness of u_ε^α, one can show that $\|u_\varepsilon^\alpha - u_\varepsilon^{K,\alpha}\|_{L^2}$ is of order ε^{K+1}. Thus, $E(\alpha,\varepsilon) = \tilde{E}(\alpha,\varepsilon) + O(\varepsilon^{K+1})$. As K is arbitrary, we see that $E(\alpha,\varepsilon)$ is a constant, subject to a correction which is smaller than any power of ε, asymptotically.

Remark 9. The uniqueness of u_ε^α indeed implies that $u_\varepsilon^\alpha(\cdot)$ is monotonic in the range $[-1/2, 1/2]$. The reason is as follows. Suppose on the contrary that $u_\varepsilon^\alpha(\cdot)$ is not monotonic in this range. Then, by translation, it attains a local maximum, say $a \in (-1/2, 1/2]$, at the origin. Set $c = u_\varepsilon^a(\varepsilon)$ and $B(\alpha) = u_\varepsilon^\alpha(-\varepsilon)$, and let α vary from a to c. Then $B(a) = u_\varepsilon^a(-\varepsilon) \le a$ and $B(c) = u_\varepsilon^c(-\varepsilon) = u_\varepsilon^a(0) = a \ge c$. Thus, by the continuity of $B(\cdot)$, there exists $\hat\alpha \in [c, a]$ such that $B(\hat\alpha) = \hat\alpha$; that is, $u_\varepsilon^{\hat\alpha}(0) = u^{\hat\alpha}(-\varepsilon)$. Hence, by the uniqueness, we must have $u^{\hat\alpha}(-\varepsilon + \cdot) = u^{\hat\alpha}(\cdot)$, which is impossible. Therefore, in the range $[-1/2, 1/2]$, $u_\varepsilon^\alpha(\cdot)$ is monotonic. Of course, as ε becomes smaller, the interval $[-1/2, 1/2]$ can be made arbitrarily close to $(-1, 1)$. We are not sure whether or not u_ε^α is monotonic overall.

Acknowledgments. PWB and XC gratefully acknowledge the support and hospitality received while visiting the Research Institute for Mathematical Sciences, Kyoto University, Japan.

REFERENCES

[1] P.W. Bates and A. Chmaj, *A discrete convolution model for phase transitions,* Arch. Ration. Mech. Anal., 150 (1999), pp. 281–305.
[2] P.W. Bates, P.C. Fife, X. Ren, and X. Wang, *Traveling waves in a convolution model for phase transitions,* Arch. Ration. Mech. Anal., 138 (1997), pp. 105–136.
[3] J.W. Cahn, J. Mallet-Paret, and E.S. Van Vleck, *Traveling wave solutions for systems of ODEs on a two-dimensional spatial lattice,* SIAM J. Appl. Math., 59 (1998), pp. 455–493.
[4] X. Chen, *Existence, uniqueness, and asymptotic stability of traveling waves in nonlocal evolution equations,* Adv. Differential Equations, 2 (1997), pp. 125–160.
[5] X. Chen, C.M. Elliott, A. Gardiner, and J.J. Zhao, *Convergence of numerical solutions to the Allen–Cahn equation,* Appl. Anal., 69 (1998), pp. 47–56.
[6] A. Chmaj and X. Ren, *Homoclinic solutions of an integral equation: Existence and stability,* J. Differential Equations, 155 (1999), pp. 17–43.
[7] A. Chmaj and X. Ren, *Multiple layered solutions of the nonlocal bistable equation,* Phys. D, 147 (2000), pp. 135–154.
[8] S.-N. Chow, J. Mallet-Paret, and W. Shen, *Traveling waves in lattice dynamical systems,* J. Differential Equations, 149 (1998), pp. 248–291.
[9] S.-N. Chow, J. Mallet-Paret, and E.S. Van Vleck, *Dynamics of lattice differential equations,* Internat. J. Bifur. Chaos Appl. Sci. Engrg., 6 (1996), pp. 1605–1621.
[10] S.-N. Chow and W. Shen, *Dynamics in a discrete Nagumo equation: Spatial topological chaos,* SIAM J. Appl. Math., 55 (1995), pp. 1764–1781.
[11] A. de Masi, T. Gobron, and E. Presutti, *Traveling fronts in non local evolution equations,* Arch. Ration. Mech. Anal., 132 (1995), pp. 143–205.
[12] G.B. Ermentrout and J.B. McLeod, *Existence and uniqueness of travelling waves for a neural network,* Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 461–478.
[13] P.C. Fife and J.B. McLeod, *The approach of solutions of nonlinear diffusion equations to travelling front solutions,* Arch. Ration. Mech. Anal., 65 (1977), pp. 335–361.
[14] S. Flach, Y. Zolotaryuk, and K. Kladko, *Moving lattice kinks and pulses: An inverse method,* Phys. Rev. E, 59 (1999), pp. 6105–6115.
[15] C. Grant and E.S. Van Vleck, *Slowly migrating transition layers for the discrete Allen-Cahn and Cahn-Hilliard Equations,* Nonlinearity, 8 (1995), pp. 861–876.
[16] D. Henry, *Geometric Theory of Semilinear Parabolic Equations,* Lecture Notes in Math. 840, Springer-Verlag, New York, 1981.
[17] J.P. Keener, *Propagation and its failure in coupled systems of discrete excitable fibers,* SIAM J. Appl. Math., 47 (1987), pp. 556–572.
[18] C. Kittel, *Introduction to Solid State Physics,* 7th ed., Wiley, New York, 1996.
[19] R.S. MacKay and J.-A. Sepulchre, *Multistability in networks of weakly coupled bistable units,* Phys. D, 82 (1995), pp. 243–254.
[20] J. Mallet-Paret, *The global structure of traveling waves in spatially discrete dynamical systems,* J. Dynam. Differential Equations, 11 (1999), pp. 49–127.
[21] B. Zinner, *Existence of traveling wavefront solutions for the discrete Nagumo equation,* J. Differential Equations, 96 (1992), pp. 1–27.

SIAM J. MATH. ANAL.
Vol. 35, No. 3, pp. 547–560

SOLVING TIME-HARMONIC SCATTERING PROBLEMS BASED ON THE POLE CONDITION II: CONVERGENCE OF THE PML METHOD*

THORSTEN HOHAGE†, FRANK SCHMIDT‡, AND LIN ZSCHIEDRICH‡

Abstract. In this paper we study the PML method for Helmholtz-type scattering problems with radially symmetric potential. The PML method consists of surrounding the computational domain with a perfectly matched sponge layer. We prove that the approximate solution obtained by the PML method converges exponentially fast to the true solution in the computational domain as the thickness of the sponge layer tends to infinity. This is a generalization of results by Lassas and Somersalo based on boundary integral equation techniques. Here we use techniques based on the pole condition instead. This makes it possible to treat problems without an explicitly known fundamental solution.

Key words. transparent boundary conditions, perfectly matched layer, pole condition

AMS subject classification. 65N12

DOI. 10.1137/S0036141002406485

1. Introduction. Since the first paper on the subject by Bérenger in 1994 [1], the perfectly matched layer (PML) method has become very popular due to its accuracy, simplicity, and flexibility. In this article we explore the connections between the PML method for time-harmonic scattering problems and the methods based on the pole condition, which are discussed in [4]. We start with a brief summary of the derivation of the PML equations. Let $u(r,\hat{x})$ denote the solution to the scattering problem in a coordinate system consisting of a radial variable $r > 0$ and a vector of angular variables $\hat{x}$. The first step of the PML method consists of a complex extension of the solution $u(\cdot,\hat{x})$ along some given path $\gamma : [a,\infty) \to \mathbb{C}$, $a > 0$, which satisfies

$$\gamma(a) = a, \qquad \operatorname{Re}\gamma(r) = r \text{ and } \operatorname{Im}\gamma'(r) \geq 0 \text{ for } r > a.$$

In Cartesian coordinates the so-called Bérenger solution $u^{(\mathrm{B})}(r,\hat{x}) := u(\gamma(r),\hat{x})$ satisfies a Helmholtz-type equation with an anisotropic damping tensor. If u is outgoing, then $u^{(\mathrm{B})}(r,\hat{x})$ decays exponentially as $r \to \infty$. On the other hand, $u^{(\mathrm{B})}(r,\hat{x})$ grows exponentially if u is an incoming field. Therefore, the Sommerfeld radiation condition for u is equivalent to the boundedness of $u^{(\mathrm{B})}(r,\hat{x})$. In a second step, the boundedness condition for $u^{(\mathrm{B})}(r,\hat{x})$ is replaced by the zero Dirichlet condition $u^{(\mathrm{B})}(\rho,\hat{x}) = 0$ at some finite distance $\rho > a$. We end up with an elliptic boundary value problem on a bounded domain, which can be solved by standard finite element codes.

The analysis in this paper is based on the work of Collino and Monk [2] and Lassas and Somersalo [8]. We show that, for the Helmholtz equation with a radially symmetric potential, the solution to the PML equations converges exponentially to the true solution within the ball $\{x : |x| < a\}$ as $\rho \to \infty$. In [9] Lassas and Somersalo show

*Received by the editors April 26, 2002; accepted for publication (in revised form) October 31, 2002; published electronically August 23, 2003.
http://www.siam.org/journals/sima/35-3/40648.html
†Institut für Numerische und Angewandte Mathematik, Universität Göttingen, Lotzestr. 16–18, D-37083 Göttingen, Germany (hohage@math.uni-goettingen.de). This author was supported by DFG grant DE293/7-1.
‡Zuse Institute Berlin (ZIB), Takustr. 7, D-14195 Berlin, Germany (frank.schmidt@zib.de, zschiedrich@zib.de). The third author was supported by DFG grant SCHM 1386/1-1.

the exponential convergence of the PML method for general convex computational domains, but constant exterior potentials. In [7] the PML method for Maxwell's equations is interpreted as a complexification of the metric. Our proof proceeds along the same lines as in [8], but we replace integral equation techniques with a representation formula derived in [4]. This allows us to treat problems for which no fundamental solution is known explicitly. In particular, as shown in numerical experiments, the method converges for inhomogeneous exterior domains involving waveguide structures; see [5, 6, 10]. However, the analysis presented in this paper covers only radially symmetric potentials.

We also show that there exists a close connection between the exponential decay of Bérenger solutions and the pole condition. The former condition always implies the latter, and under certain conditions on the singularities in the Laplace domain, the converse implication holds true as well. As a consequence, the class of applications of the PML method and methods based on the pole condition is almost the same within the class of time-harmonic scattering problems. For a comparison of the numerical performance of the two methods, we refer to [6]. A potential advantage of methods based on the pole condition is the possibility to evaluate the exterior field numerically by a representation formula if the location and the type of the singularities in the Laplace domain are known. This is particularly relevant if a fundamental solution is not known explicitly. Otherwise, the exterior field can be evaluated by Green's representation formula.

The plan of this paper is as follows. In section 2 we introduce the class of problems considered in this paper and the corresponding Dirichlet-to-Neumann (DtN) map. Section 3 deals with the analytic continuation properties of the solution and the relation to the pole condition. In section 4 a more detailed derivation of the PML equations is given. Finally, in section 5 we prove our main theorem on the exponential convergence of the solutions to the PML equations as $\rho \to \infty$.

2. Helmholtz scattering problem and the DtN-operator. We are concerned with Helmholtz-type scattering problems

$$\Delta u(x) + k^2(x)\, u(x) = 0 \text{ in } \mathbb{R}^d \setminus K, \tag{2.1a}$$

$$\frac{\partial}{\partial \nu} u|_{\partial K} = f, \tag{2.1b}$$

$$\lim_{r\to\infty} r^{\frac{d-1}{2}} \left(\frac{\partial u}{\partial r} - i\kappa u \right) = 0. \tag{2.1c}$$

Here $K \subset \mathbb{R}^d$ denotes a compact smooth set, $f \in H^{-1/2}(\partial K)$, and ν is the normal vector on ∂K pointing to the interior of K. We assume that k is a bounded, continuous function which is given by

$$k^2(x) = (1 + p(|x|))\,\kappa^2 \text{ for } |x| \geq a_*.$$

Here $p\left(t^{-1}\right) = \sum_{m=2}^{\infty} p_m t^m$ has a convergence radius greater than $\frac{1}{a_p}$, $a_p \in (0, \infty]$ with $a_* > a_p$. As proved in [4] the above system has a unique solution. We are interested only in the restriction of the solution to the interior domain $\Omega_a = B_a^d \setminus \overline{K}$, where $a > a_*$. We denote $u^{(\mathrm{int})}(x) = u(x)$, $x \in \Omega_a$. $u^{(\mathrm{int})}$ is the unique solution to the variational problem

$$\int_{\Omega_a} \nabla u \nabla \overline{v}\, \mathrm{d}x - \int_{\Omega_a} k^2 u \overline{v}\, \mathrm{d}x - \int_{S_a^{d-1}} \mathrm{DtN}_a u \overline{v}\, \mathrm{d}s = \int_{\partial K} f \overline{v}\, \mathrm{d}s \tag{2.2}$$

for all $v \in H^1(\Omega_a)$ (see [4]). The corresponding boundary value problem is

$$\Delta u^{(\mathrm{int})} + k^2 u^{(\mathrm{int})} = 0 \quad \text{in } \Omega_a, \tag{2.3a}$$

$$\partial_n u^{(\mathrm{int})} = f \quad \text{on } \partial K, \tag{2.3b}$$

$$\partial_r u^{(\mathrm{int})} - \mathrm{DtN}_a u^{(\mathrm{int})} = 0 \quad \text{on } S_a^{d-1}. \tag{2.3c}$$

Here $S_a^{d-1} := \left\{x \in \mathbb{R}^d \,:\, |x| = a\right\}$ and $\mathrm{DtN}_a : H^{1/2}\left(S_a^{d-1}\right) \to H^{-1/2}\left(S_a^{d-1}\right)$ denotes the DtN map defined as follows. Given $g \in H^{1/2}\left(S_a^{d-1}\right)$

$$\mathrm{DtN}_a g = \partial_r u^{(\mathrm{ext})}|_{S_a^{d-1}},$$

where $u^{(\mathrm{ext})}$ is the unique solution to the exterior problem

$$\Delta u^{(\mathrm{ext})} + k^2 u^{(\mathrm{ext})} = 0 \text{ in } D_{a,\infty}, \tag{2.4a}$$

$$u^{(\mathrm{ext})}|_{\partial S_a^{d-1}} = g, \tag{2.4b}$$

$$\lim_{r\to\infty} r^{\frac{d-1}{2}} \left(\frac{\partial u^{(\mathrm{ext})}}{\partial r} - i\kappa u^{(\mathrm{ext})}\right) = 0. \tag{2.4c}$$

Here and in the following we use the notation

$$D_{\theta_1,\theta_2} = B_{\theta_2}^d \setminus \overline{B_{\theta_1}^d}, \qquad D_{\theta_1,\infty} = \mathbb{R}^d \setminus \overline{B_{\theta_1}^d}. \tag{2.5}$$

We call the boundary condition (2.3c) transparent because it leads to the exact solution in the interior domain without any spurious reflections. It can be seen from the above definition of the DtN_a-operator that the boundary condition (2.3c) is nonlocal. In particular, it is not given by a finite sum of differential operators acting on the boundary S_a^{d-1}. Due to the presence of the potential p, neither an integral representation nor a series representation of the DtN_a-operator is known explicitly. Therefore, it cannot be used directly for a finite element approximation of the interior problem. Nevertheless it will provide the theoretical framework of our convergence proof, where we interpret the action of the sponge layer as a perturbation of the DtN_a-operator.

3. The pole condition and analytic continuation of the exterior solution. We introduce polar coordinates $r > 0$ and $\hat{x} \in S_1^{d-1}$ in $\mathbb{R}^d$. With a slight misuse of notation, we use the same letter for exterior fields in polar and Cartesian coordinates, i.e., $u(r,\hat{x}) = u(r\hat{x})$. If u is a solution to the boundary problem (2.4), we will show that $u(\cdot,\hat{x})$ has a holomorphic extension to $\mathbb{C}_a^{++} = \{z \in \mathbb{C} \,:\, \mathrm{Re}\, z > a,\ \mathrm{Im}\, z \geq 0\}$. Recall that the Laplace operator in polar coordinates is given by $\frac{1}{r^{d-1}}\partial_r\left(r^{d-1}\partial_r\right) + \frac{1}{r^2}\Delta_{\hat{x}}$, where $\Delta_{\hat{x}}$ denotes the Laplace–Beltrami operator on the unit sphere. We replace the real coordinate r with the complex variable z and define

$$\Delta_z = \frac{1}{z^{d-1}}\frac{\partial}{\partial z} z^{d-1} \frac{\partial}{\partial z} + \frac{1}{z^2}\Delta_{\hat{x}}.$$

As usual, we define $\frac{\partial}{\partial z} f(z) := \lim_{\tilde{z}\in\mathbb{C}_a^{++}\to z} \frac{f(\tilde{z})-f(z)}{\tilde{z}-z}$. Thus, ∂_z is a one-sided derivative on the real axis. We consider nonstandard boundary value problems of the form

$$\Delta_z u(z,\hat{x}) + k^2(z)\, u(z,\hat{x}) = 0, \qquad z \in \mathbb{C}_{z_0}^{++},\ \hat{x} \in S_1^{d-1}, \tag{3.1a}$$

$$u(z_0,\cdot) = \tilde{g}, \tag{3.1b}$$

$$\lim_{\mathrm{Re}\, z\to\infty} z^{\frac{d-1}{2}} \left(\frac{\partial}{\partial z} u - i\kappa u\right) = 0, \tag{3.1c}$$

where boundary condition (3.1b) has to be understood in the sense of the trace operator. We will also have to consider the case $\operatorname{Im} z_0 > 0$, where we seek a solution defined on $\mathbb{C}^{++}_{z_0} = \{z \in \mathbb{C} : \operatorname{Re} z > \operatorname{Re} a,\ \operatorname{Im} z \geq \operatorname{Im} z_0\}$. Since the energy argument in [4, Lemma 8.1] is no longer valid in this case, we cannot guarantee uniqueness in general. However, we will show in Lemma 5.3 that uniqueness is given if z_0 satisfies certain conditions. The next theorem is a generalization of [4, Theorems 8.4 and 9.3] for complex arguments. Since the proof is almost the same, it is omitted here.

THEOREM 3.1. *Let $z_0 \in \mathbb{C}^{++}_a$ and assume that (3.1) with $\tilde{g} = 0$ has only the trivial solution. Then the following hold true:*

1. *There exists a unique solution $u \in C^2\left(\mathbb{C}^{++}_{z_0} \times S_1^{d-1}\right)$ to (3.1) for all $\tilde{g} \in H^{1/2}\left(S_1^{d-1}\right)$.*
2. *There exist functions $u_\infty \in C^\infty\left(S_1^{d-1}\right)$ and $\Psi \in C^1\left(\mathbb{R}_+ \times S_1^{d-1}\right)$ and a constant $\tilde{a} > \operatorname{Re} z_0$ such that the above solution is given by*

$$u(z,\hat{x}) = z^{\frac{1-d}{2}} e^{i\kappa z}\left(u_\infty(\hat{x}) + \int_0^\infty e^{-t(z-\tilde{a})}\Psi(t,\hat{x})\,\mathrm{d}t\right) \tag{3.2}$$

for $\operatorname{Re} z \geq \tilde{a}$. $\Psi(t,\hat{x})$ decays exponentially as $t \to \infty$. The formula (3.2) may be differentiated any number of times with respect to both z and $\hat{x}$; integration and differentiation may be interchanged. Moreover, given $m \in \{0,1\}$ and $l \in \{0,1,\dots\}$, there exists a constant $C > 0$ such that

$$\left\|u_\infty\right\|_{C^l\left(S_1^{d-1}\right)} \leq C\left\|\tilde{g}\right\|_{L^2}, \tag{3.3}$$

$$\int_0^\infty t^k \left\|\frac{\partial^m}{\partial t^m}\Psi(t,\cdot)\right\|_{C^l\left(S_1^{d-1}\right)} \mathrm{d}t \leq C\left\|\tilde{g}\right\|_{L^2}. \tag{3.4}$$

Let us consider the restriction $v_a(z) := u(z+a,\hat{x})$, $z \in \mathbb{C}^{++}_0$, of the solution to a ray with direction $\hat{x} \in S_1^{d-1}$. It follows from (3.2), (3.3), and (3.4) that

$$v_a \text{ has a holomorphic extension to } \mathbb{C}^{++}_0 \text{ and} \tag{3.5a}$$

$$\sup_{z \in \mathbb{C}^{++}_0} |e^{-i\kappa z} v_a(z)| < \infty, \tag{3.5b}$$

i.e., that $v_a(z)$ decays exponentially as $\operatorname{Im} z \to \infty$. Since, for an incoming wave, $v_a(z)$ grows exponentially as $\operatorname{Im} z \to \infty$, the exponential decay of $v_a(z)$ as $\operatorname{Im} z \to \infty$ characterizes outgoing waves. At the same time it is the foundation of the PML method.

The pole condition is an alternative characterization of outgoing waves, which is also the basis of numerical algorithms (cf. [5, 6]). For the differential equation (2.1a), (2.1c), we have shown in [4] that the Laplace transform $\hat{v}_a(s) := \int_0^\infty e^{-rs} v_a(r)\,\mathrm{d}r$, $\operatorname{Re} s > 0$, satisfies the condition

$$\hat{v}_a \text{ has a holomorphic extension to } \{s \in \mathbb{C} : \operatorname{Im} s < \tilde{\kappa}\}, \tag{3.6a}$$

$$\sup_{\operatorname{Im} s < 0} |s + i\tilde{\kappa}||\hat{v}_a(s + i\tilde{\kappa})| < \infty \tag{3.6b}$$

for $\tilde{\kappa} \leq \kappa$. For general differential equations, e.g., problems with waveguides, we do not have a proof that either (3.5) or (3.6) is an appropriate characterization of outgoing waves sufficient to show both existence and uniqueness of a solution. Let us now look at the relation between the conditions (3.5) and (3.6).

THEOREM 3.2. *Let $\tilde{\kappa} < \kappa$ and $\tilde{a} > a$. If $v_a : (0,\infty) \to \mathbb{C}$ satisfies (3.5), then $v_{\tilde{a}}(r) := v_a(r + \tilde{a} - a)$ satisfies (3.6).*

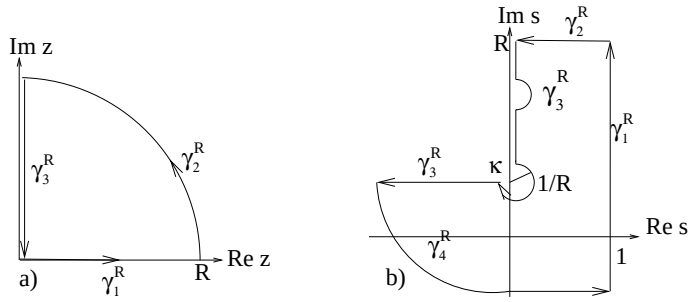


FIG. 3.1. *Relation of the pole condition and the exponential decay of the Bérenger functions.*

Proof. For the contour in Figure 3.1(a), we have $\int_{\gamma_1^R+\gamma_2^R+\gamma_3^R} e^{-sz} v_{\tilde{a}}(z)\,\mathrm{d}z = 0$ for all $R > 0$ by Cauchy's integral theorem. Since $\lim_{R\to\infty} \int_{\gamma_2^R} e^{-sz} v_{\tilde{a}}(z)\,\mathrm{d}z = 0$ for $\operatorname{Re} s > 0$ and $\operatorname{Im} s \le 0$ due to (3.5), it follows that

$$\hat{v}_{\tilde{a}}(s) = \lim_{R\to\infty} \int_{\gamma_1^R} e^{-sz} v_{\tilde{a}}(z)\,\mathrm{d}z = -\lim_{R\to\infty} \int_{\gamma_3^R} e^{-sz} v_{\tilde{a}}(z)\,\mathrm{d}z = \int_0^\infty e^{-sit} v_{\tilde{a}}(it)\, i\mathrm{d}t.$$

A partial integration yields

$$(s + i\tilde{\kappa})\hat{v}_{\tilde{a}}(s + i\tilde{\kappa}) = v_{\tilde{a}}(0) + \int_0^\infty e^{-sit + s\tilde{\kappa}t} v'_{\tilde{a}}(it)\, i\mathrm{d}t. \tag{3.7}$$

Differentiating Cauchy's formula $v_a(z) = (2\pi i)^{-1} \int_{\{|\zeta - z| = \epsilon\}} v_a(\zeta)(\zeta - z)^{-1}\,\mathrm{d}\zeta$ with $z = it + \tilde{a} - a$ and $0 < \epsilon < \tilde{a} - a$ and using assumption (3.5b), we obtain

$$|v'_{\tilde{a}}(it)| = |v'_a(it + \tilde{a} - a)| \le \sup_{|\zeta - z| \le \epsilon} \epsilon^{-1} |v_a(\zeta)| \le Ce^{-\kappa t}$$

for all $t > 0$ with some constant $C > 0$. Since $\tilde{\kappa} < \kappa$, the integrand in (3.7) decays exponentially with t for $\operatorname{Im} s \le 0$. This shows that (3.6) holds true. ∎

We immediately obtain the following.

COROLLARY 3.3. *If $u \in C(\mathbb{C}^{++}_{a_*} \times S_1^{d-1})$ is holomorphic in the first variable and satisfies*

$$\sup_{z \in \mathbb{C}^{++}_{a_*}, \hat{x} \in S^{d-1}} |e^{-i\kappa z} z^{\frac{d-1}{2}} u(z, \hat{x})| < \infty \tag{3.8}$$

for $\kappa > 0$, then u satisfies the pole condition; i.e., the functions $\hat{U}_a(s, \hat{x}) := \int_0^\infty e^{-sr}(r + a)^{(d-1)/2} u(r + a, \hat{x})\,\mathrm{d}r$, initially defined for $\Re s > 0$, have holomorphic extensions to $\{s \in \mathbb{C} : \Im s < 0\}$ for all $\hat{x} \in S^{d-1}$ and all $a > a_*$.

The converse, that the pole condition implies the exponential decay of the Bérenger function, follows from the representation formula (3.2) for problems of the form (2.1). The basic idea of the proof of this formula is a contour deformation in the Fourier inversion formula

$$v_a(r) = \frac{1}{2\pi i} \lim_{R\to\infty} \int_{\gamma_1^R} e^{sr} \hat{v}_a(s)\,\mathrm{d}s, \qquad r > 0$$

(cf. Figure 3.1(b)). We assume that v_a is smooth and satisfies (3.6). Then we have $\lim_{R\to\infty}\int_{\gamma_2^R+\gamma_4^R} e^{sz}\hat{v}_a(s)\,\mathrm{d}s = 0$, and it follows from Cauchy's integral theorem that

$$v_a(z) = -\frac{1}{2\pi i}\lim_{R\to\infty}\int_{\gamma_3^R} e^{sz}\hat{v}_a(s)\,\mathrm{d}s \tag{3.9}$$

for $z > 0$. Here the path γ_3^R is chosen such that the distance to the singularities of $\hat{v}_a(s)$ is $\geq 1/R$. If the above limit exists, not only for $z > 0$ but also for $z \in \mathbb{C}_0^{++}$, we obtain the desired extension of v_a to $\mathbb{C}_0^{++}$. However, without further information on the singularities, we cannot guarantee that this limit exists and use it to estimate $|v_a(z)|$ for $z \in \mathbb{C}_0^{++}$. Only if $\hat{v}_a$ is sufficiently well behaved can we derive the bound (3.5b).

4. The PML equations. We assume that $\gamma(r) = r\left(1 + \frac{i}{r}\int_a^r \sigma(t)\,\mathrm{d}t\right)$ with a function $\sigma \in C^1([a,\infty),[0,\infty))$ satisfying

$$\sigma(a) = 0, \tag{4.1a}$$

$$\sup_{r\geq a}\sigma(r) < \infty, \tag{4.1b}$$

$$\lim_{r\to\infty}\inf_{r'\geq r}\sigma(r') > 0. \tag{4.1c}$$

Let u denote the solution to (3.1). Then the Bérenger function defined by

$$u^{(\mathrm{B})}(r,\hat{x}) = u(\gamma(r),\hat{x}) \tag{4.2}$$

solves the boundary value problem

$$\frac{1}{\gamma^{d-1}\gamma'}\frac{\partial}{\partial r}\left(\frac{\gamma^{d-1}}{\gamma'}\frac{\partial}{\partial r}u^{(\mathrm{B})}\right) + \frac{1}{\gamma^2}\Delta_{\hat{x}}u^{(\mathrm{B})} + k^2(\gamma(r))\,u^{(\mathrm{B})} = 0, \tag{4.3a}$$

$$u^{(\mathrm{B})}(a',\cdot) = \tilde{g}. \tag{4.3b}$$

If $a' = a$ and $\tilde{g} = g$, we have $u^{(\mathrm{B})}(a\hat{x}) = u^{(\mathrm{ext})}(a\hat{x})$ and $\partial_r u^{(\mathrm{B})}(a\hat{x}) = \partial_r u^{(\mathrm{ext})}(a\hat{x})$ due to (4.1a). Thus concerning the variational formulation of the inner domain problem, $u^{(\mathrm{B})}$ is as good as $u^{(\mathrm{ext})}$.

LEMMA 4.1. *Under the assumptions of Theorem* 3.1 *with* $z_0 = \gamma(a')$*, the boundary value problem* (4.3) *has a unique solution* $u^{(\mathrm{B})}$ *in* $H^1(D_{a',\infty})$ *for all* $\tilde{g} \in H^{1/2}(S_1^{d-1})$*. It satisfies* $\|u^{(\mathrm{B})}\|_{H^1} \leq C\|\tilde{g}\|_{H^{1/2}(S_1^{d-1})}$ *with a constant* C *independent of* $\tilde{g}$*.*

Proof. Let u denote the solution to (3.1), and let $u^{(\mathrm{B})}$ be defined by (4.2). It follows from Theorem 3.1(2) that there exist constants C and $\tilde{a} > a'$ such that $\|u^{(\mathrm{B})}\|_{H^1(D_{\tilde{a},\infty})} \leq C\|\tilde{g}\|_{L^2}$. In order to show that $\|u^{(\mathrm{B})}\|_{H^1(D_{a',\tilde{a}})} \leq C\|\tilde{g}\|_{H^{1/2}}$, we consider a Fourier expansion

$$u^{(\mathrm{B})}(r,\hat{x}) = \sum_{j=1}^{\infty}\tilde{g}_j R_j(r)\varphi_j(\hat{x}). \tag{4.4}$$

Here $\{\varphi_j,\lambda_j : j\in\mathbb{N}\}$ is a complete orthonormal system of eigenfunctions and eigenvalues of the Laplace–Beltrami operator $\Delta_{\hat{x}}$ on the sphere S_1^{d-1}, $\tilde{g}_j = \int_{S_1^{d-1}}\tilde{g}\varphi_j\,\mathrm{d}s$, and $R_j(a') = 1$. By virtue of the orthogonality of the Fourier modes with respect to

the H^1-norm we have

$$\begin{aligned}\|u^{(\mathrm{B})}\|^2_{H^1(D_{a',\tilde{a}})} &= \sum_{j=1}^{\infty} |\tilde{g}_j|^2 \left\{ \|R_j(r)\varphi(\hat{x})\|^2_{L^2(D_{a',\tilde{a}})} + \|\operatorname{grad} R_j(r)\varphi(\hat{x})\|^2_{L^2(D_{a',\tilde{a}})} \right\} \\ &= \sum_{j=1}^{\infty} |\tilde{g}_j|^2 \int_{a'}^{\tilde{a}} r^{d-1} \left\{ \left(1 - \frac{\lambda_j}{r^2}\right) |R_j(r)|^2 + |R_j'(r)|^2 \right\} \mathrm{d}r.\end{aligned}$$

Since $R_j = (\frac{\gamma(\kappa a')}{\gamma(\kappa r)})^{\frac{d-1}{2}} \frac{\mathcal{H}_j(\gamma(\kappa r))}{\mathcal{H}_j(\gamma(\kappa a'))}$ in the notation of [4, Corollary 8.3], there exists a constant N such that

$$|R_j^{(l)}(r)| \le C\sqrt{-\lambda_j}^{\,l} \left|\frac{\gamma(a')}{\gamma(r)}\right|^{\sqrt{-\lambda_j}} \tag{4.5}$$

for all $j \ge N$, $a' \le r \le \tilde{a}$, and $l = 0, 1$ with a generic constant C independent of j. Plugging this into the previous equation and using the estimate $|\gamma(a')/\gamma(r)| \le |\gamma(a')|(r^2 + \operatorname{Im}\gamma(a')^2)^{-1/2}$ yields

$$\|u^{(\mathrm{B})}\|^2_{H^1(D_{a',\tilde{a}})} \le C \sum_{j=1}^{\infty} (1 + |\lambda_j|^{1/2}) |\tilde{g}_j|^2 \le C \|\tilde{g}\|^2_{H^{1/2}}$$

since $\|\tilde{g}\|^2_{H^{1/2}} = \sum_j (1 + |\lambda|)^{1/2} |\tilde{g}_j|^2$.

Assume now that $u^{(B)} \in H^1(D_{a',\infty})$ is any solution to (4.3) not necessarily related to a solution of (3.1) by (4.2). By [4, Theorem 6.4] and assumption (4.1c), the Fourier coefficients $u_j^{(\mathrm{B})}(r) := \int_{S_1^{d-1}} u^{(\mathrm{B})}(r,\cdot)\varphi_j \,\mathrm{d}s$ are linear superpositions of an exponentially decreasing function proportional to R_j and an exponentially increasing function. It follows from the orthogonality of the Fourier modes and the boundedness of $\|u^{(B)}\|_{H^1}$ that $u_j^{(\mathrm{B})}(r) = \tilde{g}_g R_j(r)$ as in (4.4). This shows uniqueness of a solution to (4.3). ∎

LEMMA 4.2. *In Cartesian coordinates* (4.3a) *has the form* $\Delta_\gamma u + k_\gamma^2 u = 0$, *where* $k_\gamma^2(r) = k^2(\gamma(r))$ *and*

$$\Delta_\gamma = \nabla \cdot A_\gamma \nabla + b_\gamma \nabla.$$

$A_\gamma \in C^1\left([a,\infty), \mathbb{C}^{d\times d}\right)$ *and* $b_\gamma \in C^0\left([a,\infty), \mathbb{C}^{1\times d}\right)$ *satisfy*

$$A_\gamma(r,\hat{x}) = G_{\hat{x}}^{\mathrm{T}} \operatorname{diag}\left(\frac{1}{(\gamma'(r))^2}, \frac{r^2}{\gamma^2(r)}, \ldots, \frac{r^2}{\gamma^2(r)}\right) G_{\hat{x}}, \tag{4.6a}$$

$$b_\gamma(r,\hat{x}) = \left(\frac{d-1}{r} - \frac{d-1}{\gamma(r)\gamma'(r)} - \frac{\gamma''(r)}{(\gamma'(r))}, 0, \ldots, 0\right) G_{\hat{x}} \tag{4.6b}$$

for any orthogonal matrix $G_{\hat{x}}$ *whose first line is* $\hat{x}$.

Proof. Let $\Delta_\gamma u$ be defined by the first two terms in (4.3a). Then

$$\begin{aligned}\int_{D_{a,\infty}} \Delta_\gamma u \overline{v} \,\mathrm{d}x = &\int_a^\infty \int_{S^{d-1}} r^{d-1} \left[\frac{1}{\gamma^{d-1}\gamma'} \frac{\partial}{\partial r}\left(\frac{\gamma^{d-1}}{\gamma'}\frac{\partial}{\partial r} u\right) + \frac{1}{\gamma^2}\Delta_{\hat{x}} u\right] \overline{v} \,\mathrm{d}s \,\mathrm{d}r \\ = &-\int_a^\infty \int_{S^{d-1}} r^{d-1} \left[\frac{1}{(\gamma')^2}\frac{\partial}{\partial r} u \frac{\partial}{\partial r}\overline{v} + \frac{1}{\gamma^2}\nabla_{\hat{x}} u \cdot \nabla_{\hat{x}}\right] \overline{v} \,\mathrm{d}s \,\mathrm{d}r \\ &-\int_a^\infty \int_{S^{d-1}} r^{d-1} \left[\left(\frac{d-1}{r} - \frac{d-1}{\gamma\gamma'} - \frac{\gamma''}{(\gamma')^3}\right)\frac{\partial}{\partial r} u\right] \overline{v} \,\mathrm{d}x\end{aligned}$$

for $u, v \in C_0^\infty(D_{a,\infty})$, where $\nabla_{\hat{x}}$ denotes the surface gradient on S_1^{d-1}. Recall that $\nabla_{\hat{x}} u(r,\hat{x})$ is the projection of $r\nabla u(r\hat{x})$ to the tangential plane, which is orthogonal to $\hat{x}$. Since this projection is given by $G_{\hat{x}}^{\mathrm{T}} \operatorname{diag}(0,1,\dots,1) G_{\hat{x}}$ we get

$$\nabla_{\hat{x}} u(r,\hat{x}) = r G_{\hat{x}}^{\mathrm{T}} \operatorname{diag}(0,1,\dots,1) G_{\hat{x}} \nabla u(r\hat{x}).$$

Analogously, $\hat{x}\frac{\partial}{\partial r} u(r,\hat{x}) G_{\hat{x}}^{\mathrm{T}} \operatorname{diag}(1,0,\dots,0) G_{\hat{x}} \nabla u(r\hat{x})$. Therefore

$$\frac{\partial}{\partial r} u(r,\hat{x}) \frac{\partial}{\partial r} \overline{v(r,\hat{x})} = (\nabla u(r\hat{x}))^{\mathrm{T}} G_{\hat{x}}^{\mathrm{T}} \operatorname{diag}(1,0,\dots,0) G_{\hat{x}} \nabla \overline{v(r\hat{x})},$$
$$\nabla_{\hat{x}} u(r,\hat{x}) \cdot \nabla_{\hat{x}} \overline{v(r,\hat{x})} = r^2 (\nabla u(r\hat{x}))^{\mathrm{T}} G_{\hat{x}}^{\mathrm{T}} \operatorname{diag}(0,1,\dots,1) G_{\hat{x}} \nabla \overline{v(r\hat{x})}.$$

Inserting this yields $\int_{D_{a,\infty}} \Delta_\gamma u \overline{v}\,\mathrm{d}x = \int_{D_{a,\infty}} A_\gamma \nabla u \cdot \nabla \overline{v} + b_\gamma \nabla u \overline{v}\,\mathrm{d}x$ for all $v \in C_0^\infty(D_{a,\infty})$. This implies the asserted form of Δ_γ. To prove the regularity of A_γ and b_γ we may choose $G_{\hat{x}}$ such that it locally depends smoothly on $\hat{x}$. ∎

LEMMA 4.3. *The operator $\frac{\gamma\gamma'}{r}\Delta_\gamma$ is strongly elliptic on $D_{a,\infty}$.*

Proof. By (4.6a) we must show that there exists $\delta > 0$ with

$$\max_{a \le r < \infty} \left\{ \operatorname{Re} \frac{\gamma(r)}{r\gamma'(r)}, \operatorname{Re} \frac{r\gamma'(r)}{\gamma(r)} \right\} > \delta > 0.$$

This follows from the assumption (4.1b) since $\operatorname{Re}(\frac{\gamma}{r\gamma'}) = \frac{1+\sigma\frac{1}{r}\int_a^r \sigma(t)\,\mathrm{d}t}{1+\sigma^2} \ge \frac{1}{1+\max\sigma^2}$ and $\operatorname{Re}(\frac{r\gamma'}{\gamma}) = \frac{1+\sigma\frac{1}{r}\int_a^r \sigma(t)\,\mathrm{d}t}{1+\left(\frac{1}{r}\int_a^r \sigma(t)\,\mathrm{d}t\right)^2} \ge \frac{1}{1+\max\sigma^2}$. ∎

So far, we have replaced the exterior Helmholtz problem (2.4) with the Bérenger problem (4.3), which is still posed on an unbounded domain. Motivated by the exponential decay of $u^{(\mathrm{B})}$, we restrict (4.3) onto a bounded domain, say $D_{a,\rho}$, $\rho > a$, and impose a zero Dirichlet boundary condition on the artificial boundary S_ρ^{d-1}. This yields the so-called PML system

$$\Delta_\gamma u + k_\gamma^2 u = 0, \quad x \in D_{a,\rho}, \tag{4.7a}$$
$$u|_{S_a^{d-1}} = g, \tag{4.7b}$$
$$u|_{S_\rho^{d-1}} = 0. \tag{4.7c}$$

We define $\mathrm{DtN}_{a,\rho}^{(\mathrm{PML})} : H^{1/2}(S_a^{d-1}) \to H^{-1/2}(S_a^{d-1})$ by

$$\mathrm{DtN}_{a,\rho}^{(\mathrm{PML})} g = \partial_r u_\rho^{(\mathrm{PML})}\Big|_{S_a^{d-1}}.$$

To derive a reformulation of (2.2) with DtN_a replaced with $\mathrm{DtN}_{a,\rho}^{(\mathrm{PML})}$ which does not involve a DtN-operator, we extend A_γ, b_γ, and k_γ to Ω_a by Id, $[0,\dots,0]$, and k^2, respectively, and introduce the Hilbert space $H_{(0)}^1(\Omega_\rho) := \{v \in H^1(\Omega_\rho) : v|_{S_\rho^{d-1}} = 0\}$. Then the variational formulation of the total PML system is

$$\int_{\Omega_\rho} (A_\gamma \nabla u) \overline{\nabla v} + b_\gamma \nabla u \overline{v} + k_{\mathrm{PML}}^2 u \overline{v}\,\mathrm{d}x = \int_{\partial K} f \overline{v}\,\mathrm{d}s \qquad \text{for all } v \in H_{(0)}^1(\Omega_\rho). \tag{4.8}$$

This problem can be solved by standard finite element codes. Using elliptic regularity results, it is easy to show that any solution to (2.2) with DtN_a replaced with $\mathrm{DtN}_{a,\rho}^{(\mathrm{PML})}$ can be extended to a solution of (4.8), and conversely the restriction of any solution to (4.8) is a solution to (2.2) with DtN_a replaced with $\mathrm{DtN}_{a,\rho}^{(\mathrm{PML})}$. In the next sections we show that (4.8) has a unique solution $u_\rho^{(\mathrm{PML})}$ for ρ large enough and suitable γ and that $u_\rho^{(\mathrm{PML})}$ converges exponentially fast to the true solution in the interior domain Ω_a.

5. Convergence analysis. In the following we are repeatedly concerned with boundary value problems of type (4.3) and (4.7) on domains D_{θ_1,θ_2}, defined in (2.5). For a compact notation of these problems we make the following definitions.

DEFINITION 5.1. *Let $a < \theta_1 < \theta_2 < \infty$ be given. We define the operators*

$$\mathcal{L}_{\theta_1,\theta_2} : H^1(D_{\theta_1,\theta_2}) \to H^{-1}(D_{\theta_1,\theta_2}) \times H^{1/2}\left(S^{d-1}_{\theta_1}\right) \times H^{1/2}\left(S^{d-1}_{\theta_2}\right)$$

and

$$\mathcal{L}_{\theta_1,\infty} : H^1_{\mathrm{loc}}(D_{\theta_1,\infty}) \to \mathcal{D}'(D_{\theta_1,\infty}) \times H^{1/2}\left(S^{d-1}_{\theta_1}\right)$$

by $u \mapsto (\Delta_\gamma u + k_\gamma^2 u, u|_{S^{d-1}_{\theta_1}}, u|_{S^{d-1}_{\theta_2}})$ and $u \mapsto (\Delta_\gamma u + k_\gamma^2 u, u|_{S^{d-1}_{\theta_1}})$, respectively.

Remark 5.2. A function $u \in H^1(D_{a,\infty})$ solves the Bérenger system (4.3) if and only if it satisfies $\mathcal{L}_{a,\infty} u = (0, g)$. A function $u \in H^1(D_{a,\rho})$ solves the PML system (4.7) if and only if it satisfies $\mathcal{L}_{a,\rho} u = (0, g, 0)$.

For the proof of the following lemma we restrict the class of admissible paths γ by the following condition: There exist constants $a' > a$ and $\sigma_0 > 0$ such that

$$\gamma(r) = (1 + i\sigma_0)\, r \tag{5.1a}$$

for $r \geq a'$. Later we will also need that

$$\kappa^2 |\alpha_0|^2 |p(\alpha_0 r)| < \frac{\min\left(1, \kappa^2\right)\sigma_0}{|\alpha_0|} \tag{5.1b}$$

for $r \geq a'$ with $\alpha_0 := (1 + i\sigma_0)$. This can always be achieved by increasing the value of a'. For simplicity we also assume $a' \geq 1$. Condition (5.1a) means that γ is a straight line in the complex plane for $r \geq a'$. It is easily checked that $\Delta_\gamma = \frac{1}{\alpha_0^2}\Delta$ for $r \geq a'$. Therefore (4.3a) is equivalent to

$$\left(\Delta + \alpha_0^2 k^2(\alpha_0 |x|)\right) u^{(\mathrm{B})} = 0, \quad |x| > a'; \tag{5.2}$$

i.e., $u^{(\mathrm{B})}$ satisfies a Helmholtz equation with a complex wave number for $|x| > a'$.

LEMMA 5.3. *The following hold true:*

1. *The operator $\mathcal{L}^{-1}_{a',\infty}$ is well defined and bounded from $\{0\} \times H^{1/2}\left(S^{d-1}_{a'}\right)$ to $H^1(D_{a',\infty})$.*
2. *The operator $\mathcal{L}_{a',\rho}$ has a bounded inverse for $\rho > a' + 1$, where a' is defined in (5.1). There exists a constant C such that $\left\|\mathcal{L}^{-1}_{a',\rho}\right\| \leq C\rho$ for all $\rho > a' + 1$.*
3. *$\mathcal{L}_{\theta_1,\theta_2}$ is a Fredholm operator with index zero for all $a < \theta_1 < \theta_2 < \infty$.*

Proof. (1) The assertion follows from Lemma 4.1 if we can show that a solution u to (3.1) with $\tilde{g} = 0$ and $z_0 = \gamma(a')$ must vanish. Let $u_j(r) := \int_{S^{d-1}_1} u(r,\cdot)\varphi_j \,\mathrm{d}s$ denote the Fourier coefficients of u. Due to (5.2) it satisfies $r^{-d+1}(r^{d-1}u_j'(r))' - \lambda_j r^{-2} u_j(r) + k^2(\alpha_0 r)u_j(r) = 0$. By (3.1b) it satisfies $u_j(a') = 0$. Moreover, by [4, Theorem 6.4] and (3.1c), $u_j(r)$ decreases exponentially as $r \to \infty$. Therefore, we can multiply the differential equation by $r^{d-1}\overline{u_j(r)}$ and integrate by parts to obtain

$$\int_{a'}^{\infty} -r^{d-1}|u_j'(r)|^2 - r^{d-3}\lambda_j |u_j(r)|^2 + \alpha_0^2 k^2(\alpha r) r^{d-1} |u_j(r)|^2 \,\mathrm{d}r = 0.$$

Taking the imaginary part of this equation gives $u_j = 0$ since

$$\mathrm{Im}\left(\alpha_0^2 k^2(\alpha_0 r)\right) \geq \kappa^2 2\sigma_0 - \kappa^2 \left|\alpha_0^2 p(\alpha_0 r)\right| > \frac{\kappa^2 \sigma_0}{|\alpha_0|} - \kappa^2 \left|\alpha_0^2 p(\alpha_0 r)\right| > 0$$

for all $r \geq a'$ by virtue of assumption (5.1b).

(2) Let $(f, h_{a'}, h_\rho) \in H^{-1}(D_{a',\rho}) \times H^{1/2}\left(S^{d-1}_{a'}\right) \times H^{1/2}\left(S^{d-1}_{\rho}\right)$. We introduce the operator $\widetilde{\mathcal{L}} : H^1(D_{a',\rho}) \to H^{-1}(D_{a,\rho})$ with

$$\widetilde{\mathcal{L}} u := \left[\Delta + (1 + p(\alpha_0|x|))\,\kappa^2\alpha_0^2\right] u.$$

Due to (5.2) we have to show that the boundary value problem

$$\widetilde{\mathcal{L}} u = \alpha_0^2 f, \tag{5.3a}$$

$$u|_{S^d_{a'}} = h_{a'}, \qquad u|_{S^d_\rho} = h_\rho \tag{5.3b}$$

has a unique solution $u \in H^1(D_{a',\rho})$ and that $\|u\| \leq C\rho \|(f, h_{a'}, h_\rho)\|$ with a constant C independent of $(f, h_{a'}, h_\rho)$ and ρ. To this end we reformulate (5.3) as a problem in $H^1_0(D_{a',\rho})$. Let $R_{a',\rho}$ denote a right inverse of the trace mapping $H^1(D_{a',\rho}) \to H^{1/2}\left(S^{d-1}_{a'}\right) \times H^{1/2}\left(S^{d-1}_{\rho}\right)$, $u \mapsto (u|_{S^{d-1}_{a'}}, u|_{S^{d-1}_{\rho}})$, and let $\mathcal{A}, \mathcal{P} : H^1_0(D_{a',\rho}) \to H^1_0(D_{a',\rho})$ be the operators defined by

$$\left\langle \left(\Delta w + \kappa^2\alpha_0^2 w\right), \overline{v}\right\rangle = (\mathcal{A}w, v),$$
$$\left\langle \kappa^2\alpha_0^2 p(\alpha_0|x|) w, \overline{v}\right\rangle = (\mathcal{P}w, v)$$

for all $w, v \in H^1_0(D_{a',\rho})$. Finally, let $\mathcal{J} : H^{-1}(D_{a',\rho}) = H^1_0(D_{a',\rho})' \to H^1_0(D_{a',\rho})$ denote the canonical isomorphism. Then u solves (5.3) if and only if $w := u - R_{a',\rho}(h_{a'}, h_\rho)$ satisfies

$$(\mathcal{A} + \mathcal{P}) w = \alpha_0^2 \mathcal{J} f - \mathcal{J}\widetilde{\mathcal{L}} R_{a',\rho}(h_{a'}, h_\rho). \tag{5.4}$$

Since

$$\operatorname{Im}\left(\frac{1}{\alpha_0}\mathcal{A}u, u\right) = \frac{\sigma_0}{1+\sigma_0^2}\left\langle \nabla u, \overline{\nabla u}\right\rangle + \kappa^2\sigma_0\langle u, \overline{u}\rangle,$$

we have

$$\frac{\min\left(1, \kappa^2\right)\sigma_0}{1+\sigma_0^2}\|u\|^2 \leq \operatorname{Im}\left(\frac{1}{\alpha_0}\mathcal{A}u, u\right) \leq \frac{1}{|\alpha_0|}\|\mathcal{A}u\|\|u\|$$

for all $u \in H^1_0(D_{a',\rho})$. This implies that $\mathcal{A}$ is boundedly invertible with $\|\mathcal{A}^{-1}\| \leq \frac{|\alpha_0|}{\min(1,\kappa^2)\sigma_0}$. Since $\|\mathcal{P}\| < \frac{\min(1,\kappa^2)\sigma_0}{|\alpha_0|}$ by virtue of assumption (5.1b), it follows that $\|\mathcal{A}^{-1}\mathcal{P}\| \leq c < 1$ with a constant c independent of ρ. Therefore, $\mathcal{A}+\mathcal{P} = \mathcal{A}(I+\mathcal{A}^{-1}\mathcal{P})$ is invertible and $\|(\mathcal{A}+\mathcal{P})^{-1}\| \leq \|A\|(1-c)^{-1} \leq C$ with a constant C independent of ρ. It follows from (5.4) that

$$\|u\| \leq C\|\mathcal{J}(\alpha_0^2 f - \widetilde{\mathcal{L}} R_{a',\rho}(h_{a'}, h_\rho))\| + \|R_{a',\rho}(h_{a'}, h_\rho)\|.$$

As both $\|\mathcal{J}\|$ and $\|\mathcal{J}\widetilde{\mathcal{L}}\|$ are uniformly bounded with respect to ρ, it remains to estimate the norm of $R_{a',\rho}$. We select a right inverse $R_{1,2}$ of the trace mapping $H^1(D_{1,2}) \to H^{1/2}\left(S^{d-1}_1\right) \times H^{1/2}\left(S^{d-1}_2\right)$, $u \mapsto (u|_{S^{d-1}_1}, u|_{S^{d-1}_2})$. Using $R_{1,2}$ we define $R_{a',\rho}$ by

$$R_{a',\rho}(h_{a'}, h_\rho)(r\hat{x}) := \left[R_{1,2}(h_{a'}(a'\cdot), h_\rho(\rho/2\cdot))\right]\left(\frac{\rho + r - 2a'}{\rho - a'}\hat{x}\right).$$

Recall that $a' \geq 1$, $\rho \geq 2$ and that

$$\left\|f\right\|_{H^{1/2}\left(S_\theta^{d-1}\right)} = \theta^{(d-1)/2} \left\| \left(1-\theta^{-2}\Delta_{\hat{x}}\right)^{1/4} f\left(\theta^{-1}\cdot\right) \right\|_{L^2\left(S_1^{d-1}\right)}.$$

Now

$$\begin{aligned}
&\left\|R_{a',\rho}\left(h_{a'},h_\rho\right)\right\| \leq C\rho^{d/2} \max\left\{ \left\|h_{a'}\left(a'\cdot\right)\right\|_{H^{1/2}\left(S_1^{d-1}\right)}, \left\|h_\rho\left(\rho/2\cdot\right)\right\|_{H^{1/2}\left(S_2^{d-1}\right)} \right\} \\
&\leq C\rho^{d/2} \max\left\{ \frac{\left(a'\right)^{1/2}}{\left(a'\right)^{(d-1)/2}} \left\|h_{a'}\right\|_{H^{1/2}\left(S_{a'}^{d-1}\right)}, \frac{2^{(d-1)/2}}{\rho^{(d-1)/2}} \left(\frac{\rho^2}{4}\right)^{1/4} \left\|h_\rho\right\|_{H^{1/2}\left(S_\rho^{d-1}\right)} \right\} \\
&\leq C\rho \max\left\{ \left\|h_{a'}\right\|_{H^{1/2}\left(S_{a'}^{d-1}\right)}, \left\|h_\rho\right\|_{H^{1/2}\left(S_\rho^{d-1}\right)} \right\}.
\end{aligned}$$

Therefore, $\left\|R_{a',\rho}\right\| \leq C\rho$ with C independent of ρ.

(3) We use Theorem 13.4 in [11]. The case $d > 2$ is clear. For $d = 2$ we must show that Δ_γ is properly elliptic. By definitions 10.5.2 and 10.5.3 in [11] and (4.6a), it suffices to show that the polynomial $P(z) = \frac{1}{(\gamma'(r))^2} + z^2 \frac{r^2}{\gamma^2(r)}$ has one root with $\operatorname{Im} z > 0$ and one root with $\operatorname{Im} z < 0$. The roots are given by $z_\pm = i\frac{\gamma(r)}{r\gamma'(r)}$. Since $\operatorname{Re}(\frac{\gamma}{r\gamma'}) = \frac{1+\sigma\frac{1}{r}\int_a^r \sigma(t)\,\mathrm{d}t}{1+\sigma^2} \geq \frac{1}{1+\max\sigma^2}$, the assertion follows from (4.1b). ∎

We emphasize that in the previous lemma we did not prove the existence of a solution to the PML system (4.7) or equivalently the existence of a solution to $\mathcal{L}_{a,\rho}u = (0,g,0)$. This will be done in the following for ρ large enough by a technique proposed by Lassas and Somersalo in [8]. The key idea is to introduce propagation operators which allow an equivalent formulation of the Bérenger and PML problems on a fixed domain. Then the PML problem can be interpreted as a perturbed Bérenger problem.

DEFINITION 5.4. *Let $a' < a'' \leq \rho$ with $a' > a$ given by* (5.1). *The* propagation operators $P_{a''}^{(\rho)}$, $P_{a''}^{(\infty)} : H^{1/2}\left(S_{a'}^{d-1}\right) \to H^{1/2}\left(S_{a''}^{d-1}\right)$ *are defined by*

$$\begin{aligned}
P_{a''}^{(\rho)} h &= \mathcal{L}_{a',\rho}^{-1}\left(0,h,0\right)|_{S_{a''}^{d-1}}, \\
P_{a''}^{(\infty)} h &= \mathcal{L}_{a',\infty}^{-1}\left(0,h\right)|_{S_{a''}^{d-1}}.
\end{aligned}$$

LEMMA 5.5. *For $a' < a'' \leq \rho$ with $a' > a$ given by* (5.1), *the following hold true:*

1. *The restriction of $u^{(\mathrm{B})}$ to $D_{a,a''}$ is the unique solution in $u \in H^1\left(D_{a,a''}\right)$ to the equation*

$$\mathcal{L}_{a,a''}u = \left(0,g,P_{a''}^{(\infty)}\left(u|_{S_{a'}^{d-1}}\right)\right). \tag{5.5}$$

2. *Let $u \in H^1\left(D_{a,\rho}\right)$ satisfy* (4.7a). *Then u satisfies* (4.7b) *and* (4.7c) *if and only if*

$$\mathcal{L}_{a,a''}u|_{D_{a,a''}} = \left(0,g,P_{a''}^{(\rho)}\left(u|_{S_{a'}^{d-1}}\right)\right). \tag{5.6}$$

Proof. (1) $u^{(\mathrm{B})}$ satisfies (5.5) by construction. Let u be any solution of (5.5) and $w = \mathcal{L}_{a',\infty}^{-1}(0,u|_{S_{a'}^{d-1}})$. Then $w|_{S_{a'}^{d-1}} = u|_{S_{a'}^{d-1}}$ and $w|_{S_{a''}^{d-1}} = u|_{S_{a''}^{d-1}}$. We conclude that $w = u = \mathcal{L}_{a',a''}^{-1}(0,w|_{S_{a'}^{d-1}}, w|_{S_{a''}^{d-1}})$ in $B_{a''} \setminus \bar{B}_{a'}$. Therefore, the function

$$W(x) = \begin{cases} u(x), & x \in B_{a''}^d \setminus \overline{B_a^d}, \\ w(x), & x \in \mathbb{R}^d \setminus \overline{B_{a''}^d}, \end{cases}$$

solves $\mathcal{L}_{a,\infty}W = (0,g)$. Hence $W = u^{(\mathrm{B})}$, and in particular $u = u^{(\mathrm{B})}|_{B_{a''}\setminus\overline{B_a^d}}$.

(2) Any solution u to (4.7) solves (5.6) by construction. Conversely, let $u \in H^1(D_{a,\rho})$ satisfy (4.7a) and (5.6) and let $w = \mathcal{L}_{a',\rho}^{-1}(0, u|_{S_{a'}^{d-1}}, 0)$. Then $w|_{S_{a'}^{d-1}} = u|_{S_{a'}^{d-1}}$ and, as a consequence of (5.6), $w|_{S_{a''}^{d-1}} = u|_{S_{a''}^{d-1}}$. It follows from the invertibility of $\mathcal{L}_{a',a''}$ that $u(x) = w(x)$ for all $x \in B_{a''}^d \setminus B_{a'}^d$. By the unique continuation principle for elliptic equations (see [3, Section 8.3]), we conclude that $u(x) = w(x)$ for $x \in B_\rho^d \setminus B_{a'}^d$. In particular u satisfies (4.7c). (4.7b) is an immediate consequence of (5.6). ∎

Again, we did not prove that (5.6) has a solution.

LEMMA 5.6. *For $a' < a''$ with $a' > a$ given by* (5.1)*, the following hold true:*

1. $P_{a''}^{(\infty)}$ *is a compact operator.*
2. *There exists a constant C such that for all $\rho > a''$ and all $h \in H^{1/2}\left(S_{a'}^{d-1}\right)$*

$$\left\|P_\rho^{(\infty)}h\right\|_{H^{1/2}} \le Ce^{-\kappa\sigma_0\rho}\left\|h\right\|_{H^{1/2}}. \tag{5.7}$$

3. *There exists a constant C such that for all $\rho > a''$*

$$\left\|P_{a''}^{(\infty)} - P_{a''}^{(\rho)}\right\| \le C\rho e^{-\kappa\sigma_0\rho}. \tag{5.8}$$

Proof. (1) By (4.4) we have $(P_{a''}^{(\infty)}h)(a''\hat{x}) = \sum_{j=1}^\infty \hat{h}_j R_j(a'')\varphi_j(\hat{x})$, where $\hat{h}_j := \int_{S_1^{d-1}} h(a'\cdot)\varphi_j \,\mathrm{d}s$. Since $\lim_{j\to\infty} R_j(a'') = 0$ due to (4.5), $P_{a''}^{(\infty)}$ is compact as an operator norm limit of the finite rank operators defined by the truncated series.

(2) Let $h \in H^{1/2}\left(S_{a'}^{d-1}\right)$ be given and define $u := \mathcal{L}_{a',\infty}^{-1}(0,h)$. By Theorem 3.1 (2), with $z_0 = \gamma(a') = (1+i\sigma_0)r$ and $\tilde{g}(\hat{x}) = h(a'\hat{x})$, we have

$$u(r\hat{x}) = \left[(1+i\sigma_0)\,r\right]^{\frac{1-d}{2}} e^{i\kappa r - \kappa\sigma_0 r}\left(u_\infty(\hat{x}) + \int_0^\infty e^{-t((1+i\sigma_0)r-\tilde{a})}\Psi(t,\hat{x})\,\mathrm{d}t\right).$$

Further it follows from the estimates (3.3) and (3.4) that there exists a constant C such that for all $z \in \mathbb{C}_{z_0}^{++}$

$$\left\|u_\infty\right\|_{H^{1/2}\left(S_1^{d-1}\right)} + \left\|\int_0^\infty e^{-t(z-\tilde{a})}\Psi(t,\cdot)\,\mathrm{d}t\right\|_{H^{1/2}\left(S_1^{d-1}\right)} \le C\left\|\tilde{g}\right\|_{L^2\left(S_1^{d-1}\right)}.$$

Using $\left\|f\right\|_{H^{1/2}\left(S_\rho^{d-1}\right)} = \rho^{(d-1)/2}\left\|\left(1-\rho^{-2}\Delta_{\hat{x}}\right)^{1/4} f\left(\rho^{-1}\cdot\right)\right\|_{L^2\left(S_1^{d-1}\right)}$, we get

$$\|P_\rho^{(\infty)}h\|_{H^{1/2}} = \|u|_{S_\rho^{d-1}}\|_{H^{1/2}} \le Ce^{-\kappa\sigma_0\rho}\left\|\tilde{g}\right\|_{L^2\left(S_1^{d-1}\right)} \le Ce^{-\kappa\sigma_0\rho}\left\|h\right\|_{H^{1/2}\left(S_{a'}^{d-1}\right)}.$$

(3) Let $h \in H^{1/2}\left(S_{a'}^{d-1}\right)$ be given. By the definition of $P_{a''}^{(\infty)}$ and $P_{a''}^{(\rho)}$, we have $P_{a''}^{(\infty)}h - P_{a''}^{(\rho)}h = \mathrm{Tr}_{S_{a''}^{d-1}}\mathcal{L}_{a',\rho}^{-1}(0,0,P_\rho^{(\infty)}h)$. Using Lemma 5.3 (3) and (5.7), we obtain

$$\left\|P_{a''}^{(\infty)}h - P_{a''}^{(\rho)}h\right\|_{H^{1/2}} = \left\|\mathrm{Tr}_{S_{a''}^{d-1}}\right\|\left\|\mathcal{L}_{a',\rho}^{-1}\right\|\left\|P_\rho^{(\infty)}h\right\| \le C\rho e^{-\kappa\sigma_0\rho}\left\|h\right\|,$$

which yields (5.8). ∎

PROPOSITION 5.7. *There exists a constant $\rho_0 > a$ such that* (4.7) *has a unique solution for $\rho \ge \rho_0$. The operator $\mathrm{DtN}_{a,\rho}^{(\mathrm{PML})}$ is well defined for $\rho \ge \rho_0$, and there exists a constant C such that*

$$\left\|\mathrm{DtN}_a - \mathrm{DtN}_{a,\rho}^{(\mathrm{PML})}\right\| \le C\rho e^{-\kappa\sigma_0\rho}. \tag{5.9}$$

Here we use the operator norm of $L\left(H^{1/2}\left(S_a^{d-1}\right), H^{-1/2}\left(S_a^{d-1}\right)\right)$.

Proof. Define the operator

$$\mathcal{K} : H^1\left(B_{a''}^d \setminus \overline{B_a^d}\right) \to H^{-1}\left(B_{a''}^d \setminus B_a^d\right) \times H^{1/2}\left(S_a^{d-1}\right) \times H^{1/2}\left(S_{a''}^{d-1}\right)$$

by $u \mapsto (0, 0, P_{a''}^{(\infty)} u|_{S_{a'}^{d-1}})$. By Lemma 5.5 (1) $u^{(\mathrm{B})}$ satisfies

$$\mathcal{L}_{a,a''} u^{(\mathrm{B})} - \mathcal{K} u^{(\mathrm{B})} = (0, g, 0)\,. \tag{5.10}$$

The operator $\mathcal{L}_{a,a''}$ is of Fredholm index zero and $\mathcal{K}$ is compact by Proposition 5.6 (1). Hence, $\mathcal{L}_{a,a''} - \mathcal{K}$ also has Fredholm index zero. Assume that u_0 satisfies (5.10) with $g = 0$. Then $\mathcal{L}_{a,a''} u_0 = (0, 0, P_{a''}^{(\infty)} u_0|_{S_{a'}^{d-1}})$, and hence by Lemma 5.5 (1), $u_0 = \mathcal{L}_{a,\infty}^{-1}(0, 0) = 0$. Thus, $(\mathcal{L}_{a,a''} - \mathcal{K})^{-1}$ exists.

Next, consider the system (5.6). The same argument as above yields an equation

$$\left(\mathcal{L}_{a,a''} - \widetilde{\mathcal{K}}\right) u_\rho^{(\mathrm{PML})} = (0, g, 0) \tag{5.11}$$

with $\widetilde{\mathcal{K}} u = (0, 0, P_{a'',\rho} \mathrm{Tr}_{a'})$. Applying $(\mathcal{L}_{a,a''} - \mathcal{K})^{-1}$ to both sides of (5.11) yields

$$\left(I + \left(\mathcal{K} - \widetilde{\mathcal{K}}\right)\right) u_\rho^{(\mathrm{PML})} = u^{(\mathrm{B})}. \tag{5.12}$$

As $(\mathcal{K} - \widetilde{\mathcal{K}})u = (0, 0, (P_{a''}^{(\infty)} - P_{a''}^{(\rho)})\mathrm{Tr}_{a'})$ it follows from Proposition 5.6 (3) that $\|\mathcal{K} - \widetilde{\mathcal{K}}\| \le C\rho e^{-\kappa\sigma_0\rho}$. Therefore (5.12) is solvable by a Neumann series for ρ large enough, and we conclude that

$$\left\|u^{(\mathrm{B})} - u_\rho^{(\mathrm{PML})}\right\|_{H^1\left(D_{a,a''}\right)} \le \frac{C\rho e^{-\kappa\sigma_0\rho}}{1 - C\rho e^{-\kappa\sigma_0\rho}} \left\|u^{(\mathrm{B})}\right\| \le C\rho e^{-\kappa\sigma_0\rho} \|g\|. \tag{5.13}$$

To finish the proof, it remains to show that any $u \in H^1\left(D_{a,a''}\right)$ satisfying $\nabla \cdot A_\gamma \nabla u + b_\gamma \nabla u + k_\gamma^2 u = 0$ has a normal derivative $\partial_r u|_{S_a^{d-1}} \in H^{-1/2}\left(S_a^{d-1}\right)$ which satisfies $\left\|\partial_r u|_{S_a^{d-1}}\right\|_{H^{-1/2}} \le C\|u\|_{H^1}$ with some constant $C > 0$ independent of u. To see this we choose a right inverse $R_a : H^{1/2}\left(S_a^{d-1}\right) \to H^1\left(D_{a,a''}\right)$ of the trace operator $\varphi \mapsto \varphi|_{S_a^{d-1}}$ satisfying $\mathrm{supp}\, R_a\varphi \subset D_{a,\frac{a+a''}{2}}$ for all $\varphi \in H^{1/2}\left(S_a^{d-1}\right)$. Given $\varphi \in H^{1/2}\left(S_a^{d-1}\right)$, we multiply the differential equation by $R_a\varphi$, integrate, and formally use the Gauss divergence theorem and the identity $A_\gamma(a, \hat{x}) = \mathrm{Id}$ to obtain

$$\left\langle \partial_r u|_{S_a^{d-1}}, \varphi \right\rangle = \int_{D_{a,\rho}} -A_\gamma \nabla u \cdot \overline{\nabla R_a\varphi} + b_\gamma \cdot \nabla u \overline{R_a\varphi} + k_\gamma^2 u \overline{R_a\varphi}\, \mathrm{d}x.$$

Since the right-hand side of this equation is bounded by $C\|u\|_{H^1}\|\varphi\|_{H^{1/2}}$, we have proved the existence of $\partial_r u \in H^{-1/2}\left(S_a^{d-1}\right)$ and the asserted bound. Hence,

$$\left\|\mathrm{DtN}_a g - \mathrm{DtN}_{a,\rho}^{(\mathrm{PML})} g\right\|_{H^{-1/2}} \le C\left\|u_\rho^{(\mathrm{PML})} - u^{(\mathrm{B})}\right\|_{H^1\left(D_{a,a''}\right)}.$$

Together with (5.13) this implies the estimate (5.9). ∎

THEOREM 5.8. *After possibly increasing the constant* ρ_0 *in Proposition* 5.7*, there exists a constant* $C > 0$ *such that the variational problem* (4.8) *has a unique solution* $u_\rho^{(\mathrm{PML})} \in H^1_{(0)}(\Omega_\rho)$ *for all* $\rho \ge \rho_0$ *satisfying*

$$\left\|u^{(\mathrm{B})} - u_\rho^{(\mathrm{PML})}\right\|_{H^1(\Omega_a)} \le C\rho e^{-\kappa\sigma_0\rho}. \tag{5.14}$$

Proof. We introduce the operators $\mathcal{L}, \mathcal{G}, \mathcal{G}_\rho : H^1(\Omega) \to H^1(\Omega)$ with

$$\begin{aligned}(\mathcal{L}u, v) &= \langle \operatorname{grad} u, \operatorname{grad} \overline{v} \rangle - \langle k^2 u, \overline{v} \rangle,\\ (\mathcal{G}u, v) &= \int_{S_a^{d-1}} \mathrm{DtN}_a u \overline{v} \, \mathrm{d}s,\\ (\mathcal{G}_\rho u, v) &= \int_{S_a^{d-1}} \mathrm{DtN}_{a,\rho}^{\mathrm{PML}} u \overline{v} \, \mathrm{d}s\end{aligned}$$

for all $u, v \in H^1(\Omega_a)$. Moreover, we define $F \in H^1(\Omega_a)$ by $(F, v) = \int_{\partial K} f \overline{v} \, \mathrm{d}s$, $v \in H^1(\Omega_a)$. Then $(\mathcal{L}+\mathcal{G})u^{(\mathrm{int})} = F$ and $(\mathcal{L}+\mathcal{G}_\rho) u_\rho^{(\mathrm{PML})}\big|_{\Omega_a} = F$. Since we know that $\mathcal{L}+\mathcal{G}$ is boundedly invertible (cf. [4]), it follows by a Neumann series argument that there exist constants $C, \epsilon > 0$ such that $\mathcal{L}+\mathcal{G}_\rho$ is invertible for $\|\mathcal{G} - \mathcal{G}_\rho\|_{H^1} < \epsilon$ and

$$\|u_\rho^{(\mathrm{PML})}\big|_{\Omega_a} - u\|_{H^1} \le C \|\mathcal{G} - \mathcal{G}_\rho\|_{H^1} \le C \|\mathrm{DtN}_a - \mathrm{DtN}_{a,\rho}^{(\mathrm{PML})}\|.$$

Now the assertion follows from Proposition 5.7. ∎

REFERENCES

[1] J. P. Bérenger, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 114 (1994), pp. 185–200.

[2] F. Collino and P. Monk, *The perfectly matched layer in curvilinear coordinates*, Comput. Methods Appl. Mech. Engrg., 164 (1998), pp. 157–171.

[3] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering*, Springer Verlag, Berlin, Heidelberg, New York, 1992.

[4] T. Hohage, F. Schmidt, and L. Zschiedrich, *Solving time-harmonic scattering problems based on the pole condition* I: *Theory*, SIAM J. Math. Anal., 35 (2003), pp. 183–210.

[5] T. Hohage, F. Schmidt, and L. Zschiedrich, *Solving time-harmonic scattering problems based on the pole condition* III: *Numerical algorithms*, in preparation.

[6] T. Hohage, F. Schmidt, and L. Zschiedrich, *A new method for the solution of scattering problems*, in Proceedings of the JEE'02 Symposium, B. Michielsen and F. Decavèle, eds., Toulouse, 2002, ONERA, pp. 251–256.

[7] M. Lassas, J. Liukkonen, and E. Somersalo, *Complex Riemannian metric and absorbing boundary conditions*, J. Math. Pures Appl., 80 (2001), pp. 739–768.

[8] M. Lassas and E. Somersalo, *On the existence and convergence of the solution of PML equations*, Computing, 60 (1998), pp. 228–241.

[9] M. Lassas and E. Somersalo, *Analysis of the PML equations in general convex geometry*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 1183–1207.

[10] F. Schmidt, *Solution of interior-exterior Helmholtz-type problems based on the pole condition: Theory and algorithms*, Habilitation thesis, Freie Universität Berlin, Berlin, Germany, 2001.

[11] J. Wloka, *Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1987.

SIAM J. MATH. ANAL.
Vol. 35, No. 3, pp. 561–578

A PARABOLIC CROSS-DIFFUSION SYSTEM FOR GRANULAR MATERIALS*

GONZALO GALIANO†, ANSGAR JÜNGEL‡, AND JULIÁN VELASCO†

Abstract. We analyze a cross-diffusion system of parabolic equations for the relative concentration and the dynamic repose angle of a mixture of two different granular materials in a long rotating drum. The main feature of the system is the ability to describe the axial segregation of the two granular components. The existence of global-in-time weak solutions is shown for arbitrary large cross-diffusion by using entropy-type inequalities and approximation arguments. The uniqueness of solutions is proved if cross-diffusion is not too large. Furthermore, we derive a sufficient condition on the parameters to have nonsegregation. Finally, numerical simulations illustrate the long-time coarsening of the segregation bands in the drum.

Key words. strongly nonlinear parabolic system, cross-diffusion, segregation, existence of weak solutions, uniqueness of solutions, entropy-type estimates

AMS subject classifications. 35K55, 76T25

DOI. 10.1137/S0036141002409386

1. Introduction.

One important feature of granular materials, consisting of different components, is their ability to segregate under external agitation rather than further mixing [21]. Consider a long cylinder rotating about its longitudinal axis which is partially filled with a mixture of two different kinds of granular particles. The mixture of grains may exhibit both radial and axial size segregation. Roughly speaking, radial segregation occurs during the first few revolutions of the drum and is often followed by slow axial segregation. Axial segregation leads to either a stable array of concentration bands or, after a very long time, to complete segregation [2, 3, 24].

In this paper we are interested in the existence analysis of a specific model for granular materials derived in [3]. Consider a mixture of two kinds of particles with volume concentrations $u_1,\ u_2 \in [0,1]$ placed in a horizontal long narrow rotating cylinder of length $L > 0$. Let $u = u_1 - u_2 \in [-1,1]$ be the relative concentration of the mixture. Introduce the so-called dynamic angle of repose θ as the arctangent of the average slope of the free surface of the mixture, which is assumed to be flat (see Figure 1.1). The variables u and θ are assumed to be constant in each cross section of the drum and depend therefore only on the axial coordinate $z \in \Omega = (0,L)$ and on the time $t > 0$.

*Received by the editors June 8, 2002; accepted for publication (in revised form) February 7, 2003; published electronically August 23, 2003. This work was partially supported by the Spanish-German Bilateral Project Acciones Integradas–DAAD.

http://www.siam.org/journals/sima/35-3/40938.html

†Departamento de Matemáticas, Universidad de Oviedo, c/ Calvo Sotelo s/n, 33007 Oviedo, Spain (galiano@correo.uniovi.es, julian@orion.ciencias.uniovi.es). The first author was partially supported by the European RTN Project, grant HPRN-CT2002-00274. The first and the third authors were supported by the Spanish D.G.I. Project BFM2000-1324.

‡Fachbereich Mathematik und Statistik, Universität Konstanz, 78457 Konstanz, Germany (ansgar.juengel@uni-konstanz.de). This author was supported by the Deutsche Forschungsgemeinschaft, grants JU 359/3 (Gerhard-Hess Program) and JU 359/5 (Priority Program "Multiscale Problems"), the European IHP Project "Hyperbolic and Kinetic Equations" (grant HPRN-CT-2002-00282), and the AFF Project of the University of Konstanz.

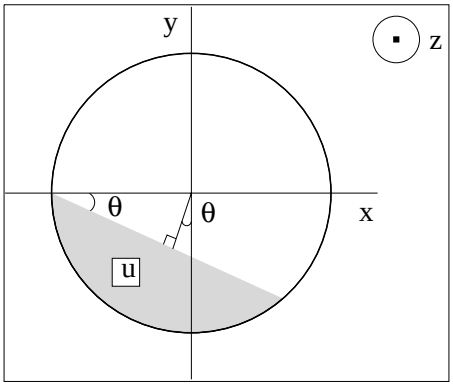


FIG. 1.1. *Relative concentration u and dynamical angle of repose θ in the geometry of the cross section of a rotating drum. The gray region indicates the mixture partially filling the drum. The variables u and θ are assumed to be constant in each cross section.*

In [3] the following cross-diffusion system for the evolution of u and θ was derived:

$$u_t - (\nu u_z - (1-u^2)\theta_z)_z = 0, \tag{1.1}$$

$$\theta_t - (\gamma u + \theta)_{zz} + \theta = \mu u \quad \text{in } Q_T := \Omega \times (0,T), \tag{1.2}$$

where the subindices denote partial derivatives. The model (1.1)–(1.2) is obtained by averaging the mass conservation laws for the two components of the granular matter over the cross section of the cylinder, under the main assumptions that the mass of grains in each cross section of the drum remains constant and that the grains separate predominantly near the surface of the drum, whereas in the bulk of the drum particles are equally advected by the bulk flow (see [3] for details of the derivation).

The positive constant ν is related to the Fickian diffusion constants arising in the surface fluxes of the two materials. The constant $\gamma > 0$ is proportional to the difference of the Fickian diffusivities. Finally, μ is related to the difference of the static repose angles of the two kind of particles.

We impose as in [3, 20] periodic boundary conditions and initial conditions for u and θ, as we are not interested in effects due to the boundary conditions:

$$\begin{aligned} &u(0,\cdot) = u(L,\cdot), \quad u_z(0,\cdot) = u_z(L,\cdot) \\ &\theta(0,\cdot) = \theta(L,\cdot), \quad \theta_z(0,\cdot) = \theta_z(L,\cdot) \\ &u(\cdot,0) = u_0, \qquad \theta(\cdot,0) = \theta_0 \end{aligned} \qquad \begin{aligned} &\text{in } (0,T), \\ &\\ &\text{in } \Omega. \end{aligned} \tag{1.3}$$

In the physical literature, periodic boundary conditions have been employed in numerical simulations of the dynamics of the granular materials in order to eliminate boundary effects [3, 20]. The subsequent analysis also works for no-flux and Dirichlet boundary conditions (with appropriate changes of the obtained estimates).

We remark that the problem is intrinsically one-dimensional in space since the equations are obtained by averaging over the cross section. For a two-dimensional model we refer, for instance, to [10].

The terms $((1-u^2)\theta_z)_z$ and γu_{zz} in (1.1)–(1.2) are called *cross-diffusion* terms [17]. We remark that segregation effects due to cross-diffusion are well known in population dynamics, and related cross-diffusion systems have been studied in mathematical biology (see, e.g., [18, 22]).

Segregation phenomena of granular material in rotating drums have been intensively investigated in the physical literature. For instance, radial segregation has been investigated numerically using particle methods [10] and analytically using leading-order analysis [6] or shock-wave analysis [14]. Axial segregation has been simulated, for instance, in [2, 3, 20] and analyzed in [3, 5, 16]. For more references, particularly for experimental studies, we refer to the monograph [21] and the review paper [19].

Mathematically, the evolution problem (1.1)–(1.2) has a full and nonsymmetric diffusion matrix:

$$A := \begin{pmatrix} \nu & -(1-u^2) \\ \gamma & 1 \end{pmatrix}.$$

Problems with full diffusion matrices also arise, for instance, in semiconductor theory [7], population dynamics [18], and nonequilibrium thermodynamics [9]. As a consequence, no classical maximum principle arguments and no regularity theory as for single equations are generally available for such problems.

Notice that there are values for u and the parameters ν and γ for which the above matrix A is *not* positive definite in the sense that $x^\top Ax < 0$ may hold for some x. The ellipticity of the system (1.1)–(1.2) is guaranteed if $4\nu > \gamma$ (and $|u| \le 1$). For these values, the existence of global-in-time solutions of (1.1)–(1.3) can be proved using standard techniques. The question arises if it is possible to prove the existence of global weak solutions for *any* values of the parameters $\nu > 0$ and $\gamma > 0$. In this paper we give a positive answer to this question.

The key of the existence analysis is the observation that the system (1.1)–(1.2) possesses a functional whose time derivative is uniformly bounded in time if $|u| < 1$. Indeed, using the functions $\phi(u)$, where

$$\phi(s) := \frac{\gamma}{2}\log\frac{1+s}{1-s} \qquad \text{for } -1 < s < 1,$$

and θ in the weak formulation of (1.1) and (1.2), respectively, and adding the resulting equations leads to the inequality

$$\frac{d}{dt}\int_0^L \left(\Phi(u) + \frac{1}{2}\theta^2\right)dz + \int_0^L (\gamma\nu u_z^2 + \theta_z^2)dz = \int_0^L (\mu u\theta - \theta^2)dz \le c, \tag{1.4}$$

where $c > 0$ depends only on μ and L. Here the function $\Phi(s) := \frac{\gamma}{2}(1-s)\log(1-s) + \frac{\gamma}{2}(1+s)\log(1+s) \ge 0$ is the primitive of ϕ such that $\Phi(0) = 0$. Observe that this estimate is purely formal since the values $|u| = 1$ are possible.

The estimate (1.4) has an important consequence. With the change of unknowns $u = g(v)$, where g is the inverse of ϕ, i.e., $g : \mathbb{R} \to (-1, 1)$ is given by

$$g(s) := \frac{e^{2s/\gamma} - 1}{e^{2s/\gamma} + 1}, \tag{1.5}$$

the system (1.1)–(1.2) becomes, for $|u| < 1$,

$$g(v)_t - (\nu g'(v)v_z - (1-g(v)^2)\theta_z)_z = 0, \tag{1.6}$$

$$\theta_t - (\gamma g'(v)v_z + \theta_z)_z + \theta = \mu g(v). \tag{1.7}$$

Since $\gamma g' = 1 - g^2$, the diffusion matrix of the transformed problem

$$B := \begin{pmatrix} \nu g'(v) & -(1-g(v)^2) \\ \gamma g'(v) & 1 \end{pmatrix} \tag{1.8}$$

satisfies for *any* values of $\nu > 0$ and $\gamma > 0$

$$(x,y)B(x,y)^\top = \nu g'(v)|x|^2 + |y|^2 \geq 0 \quad \forall x, y \in \mathbb{R}.$$

The fact that the above transformation of variables leads to a system of elliptic equations for all values of the parameters can be related to some analytical work on more general equations. Indeed, this fact is in some sense related to the equivalence between the existence of an entropy and the symmetrizability of hyperbolic conservation laws or parabolic systems [8, 15]. Using the definition of the (generalized) "entropy"

$$\eta(s) := g(s)s - \chi(s) + \chi(0) \tag{1.9}$$

from [4] (first used in [1]), where $\chi' = g$, gives $\eta(v) = \Phi(g(v)) = \Phi(u)$, with Φ as above. In this sense, the functional $\Phi(u(t)) + \theta(t)^2/2$ can be interpreted as an "entropy" for the system (1.1)–(1.2) as long as $|u| < 1$. However, notice that the matrix B is *not* symmetric but satisfies the inequality $x^\top Bx > 0$ for all $x \neq 0$, which is sufficient for our existence analysis. The question of whether this observation leads to an existence theory for elliptic systems with general full diffusion matrices is under investigation [12].

In order to make the above "entropy" estimate rigorous, we have to overcome the difficulties near the points where $|u| = 1$. For the transformed problem (1.6)–(1.7) this difficulty translates into the fact that the matrix B does not satisfy the *uniform* positive definiteness condition. Therefore, we have to approximate (1.6)–(1.7) appropriately; see section 2.

Our main existence result is as follows.

THEOREM 1.1. *Let γ, $\nu > 0$, $\mu \geq 0$, and u_0, $\theta_0 \in L^2(\Omega)$ with $-1 \leq u_0 \leq 1$ in Ω. For any $T > 0$, there exists a weak solution (u, θ) of (1.1)–(1.2) such that*

$$\begin{gathered} u, \theta \in H^1(0,T;(H^1_{\rm per}(\Omega))') \cap L^2(0,T;H^1_{\rm per}(\Omega)), \\ -1 \leq u \leq 1 \quad \text{in } Q_T = \Omega \times (0,T). \end{gathered} \tag{1.10}$$

As explained above, the main difficulties of the proof of this theorem are that the system (1.1)–(1.2) is generally not positive definite and no maximum principle to show $|u| \leq 1$ is available. Nevertheless, we are able to prove the existence of solutions for *any* values of ν and γ and thus for *arbitrary* large cross-diffusion.

The proof consists of three steps. First, instead of using the transformation g, we make a change of unknowns which takes into account the singular points $|u| = 1$ (section 2.1). Then the parabolic problem is discretized in time by a recursive sequence of elliptic equations which can be solved each by Schauder's fixed point theorem (section 2.2). Finally, a priori bounds independent of the time discretization parameter are obtained from an inequality similar to (1.4), and standard compactness results lead to the existence of a solution of the original problem (1.1)–(1.2) (section 2.3). The bound on u can be proved by using Stampacchia's truncation method in the approximate problem.

We notice that for $\gamma = 0$, the diffusion matrix for (1.1)–(1.2) becomes tridiagonal, and thus the problem can be solved, for instance, by methods employed in chemotaxis problems [11].

Besides the existence analysis we show two additional results. We prove the uniqueness of solutions in a slightly smaller class of functions if the cross-diffusion is not too large (section 3).

THEOREM 1.2. *Let $\gamma < 4\nu$. Then under the assumptions of Theorem* 1.1 *there exists at most one solution (u, θ) of* (1.1)–(1.2) *in the class of functions satisfying* (1.10) *and $\theta \in L^\infty(0, T; H^1_{\rm per}(\Omega))$.*

Furthermore, we derive a sufficient condition on the parameters in order to get nonsegregation, i.e., convergence of the transient solutions to the constant steady state given by

$$\bar{u} = \frac{1}{L}\int_0^L u_0(z)dz, \quad \bar{\theta} = \frac{1}{L}\int_0^L \theta_0(z)dz.$$

The rate of convergence turns out to be exponential (section 4).

THEOREM 1.3. *Let the assumptions of Theorem* 1.1 *hold and assume that $|u_0| \le c < 1$ in Ω for some $c < 1$, $\mu\bar{u} = \bar{\theta}$ and*

$$\frac{\nu\gamma}{\mu^2} > \frac{L^4}{8(L^2+1)}. \tag{1.11}$$

Then there exist constants $c_0 > 0$, depending on u_0, θ_0, and constants $\delta_1, \delta_2 > 0$, depending on the parameters, such that for all $t > 0$,

$$\|u(t) - \bar{u}\|_{L^2(\Omega)} \le c_0 e^{-\delta_1 t}, \quad \|\theta(t) - \bar{\theta}\|_{L^2(\Omega)} \le c_0 e^{-\delta_2 t}.$$

The constants c_0 and δ_1, δ_2 are defined in (4.1) and (4.4), respectively. The proof of the above result is based on careful estimates using the "entropy" (1.9). Aranson, Tsimring, and Vinokur [3] have determined from linear stability theory that the condition $\mu > \nu$ is necessary to have size segregation. The assumption (1.11) shows that the condition $\mu > \nu$ need *not* be sufficient. In fact, there are parameter values for which *both* $\mu > \nu$ *and* (1.11) hold; i.e., the granular materials are not segregating (see section 5).

Clearly, the dynamics of granular segregation pattern is of much greater interest for the applications than nonsegregation conditions. Therefore, our result must be seen as a first step in the understanding of segregation dynamics.

Finally, we present in section 5 some numerical examples illustrating the segregation or nonsegregation behavior.

2. Proof of Theorem 1.1.

2.1. Ideas of the proof. In this section we present and explain the approximations needed in the proof of Theorem 1.1. As already mentioned in the introduction, the function g provides an "entropy" estimate only if $|u| < 1$. Since $u = \pm 1$ is possible, we use another change of unknowns which includes the points $u = \pm 1$. Let the assumptions of Theorem 1.1 hold and let $\alpha > 1$. Define the transformation $u = g_\alpha(v)$ with $g_\alpha : [-s_\alpha, s_\alpha] \to [-1, 1]$, given by

$$g_\alpha(s) := \alpha\frac{e^{2\alpha s/\gamma} - 1}{e^{2\alpha s/\gamma} + 1} \quad \text{and} \quad s_\alpha := \frac{\gamma}{2\alpha}\log\frac{\alpha+1}{\alpha-1}. \tag{2.1}$$

Observe that for $\alpha \to 1$, g_α equals g on $\mathbb{R}$; see (1.5). As the range of g_α is $[-1, 1]$, the critical points $u = \pm 1$ are included in that transformation. In the following we fix some $\alpha > 1$ and again write g for g_α.

With this change of unknown we obtain the system (1.6)–(1.7), with periodic boundary conditions for v and θ and initial conditions

$$v(\cdot, 0) = v_0 := g^{-1}(u_0), \quad \theta(\cdot, 0) = \theta_0 \qquad \text{in } \Omega. \tag{2.2}$$

The new diffusion matrix B is given by (1.8). It holds for any $(x,y)\in\mathbb{R}^2$

$$\begin{aligned}(x,y)B(x,y)^\top &= \nu g'(v)x^2+y^2+\big(\gamma g'(v)-(1-g(v)^2)\big)\,xy\\ &= \nu g'(v)x^2+y^2+(\alpha^2-1)xy.\end{aligned}$$

Clearly, if $\alpha=1$, the matrix satisfies $x^\top Bx>0$ for all $x\neq 0$, and it seems reasonable that this will also be the case for $\alpha>1$ sufficiently close to 1. In fact, let (v,θ) be a weak solution to (1.1)–(1.2) and use v and θ as test functions in the weak formulation of (1.6)–(1.7), respectively, to obtain the identity

$$\begin{aligned}&\int_\Omega\left(G(v(t))+\frac{1}{2}\theta(t)^2\right)dz+\int_0^t\int_\Omega\left(\nu g'(v)^2v_z^2+\theta_z^2+\theta^2\right)dzdt\\ &\quad=\int_\Omega\left(G(v_0)+\frac{1}{2}\theta_0^2\right)dz-(\alpha^2-1)\int_0^t\int_\Omega v_z\theta_z dzdt+\int_0^t\int_\Omega\mu g(v)\theta dzdt,\end{aligned}$$

where G is defined by $G'(s)=sg'(s)$ and $G(0)=0$, i.e.,

$$G(s)=\frac{2\alpha s}{\gamma}\,\frac{e^{2\alpha s/\gamma}}{e^{2\alpha s/\gamma}+1}+\log\frac{2}{e^{2\alpha s/\gamma}+1}. \tag{2.3}$$

Since $|g|$ is bounded by 1 and $g'\geq(\alpha^2-1)/\gamma$ in $[-s_\alpha,s_\alpha]$ (see Lemma 2.2), we can estimate

$$\begin{aligned}&\int_\Omega\left(G(v(t))+\frac{1}{2}\theta(t)^2\right)dz+\int_0^t\int_\Omega\left(\frac{\nu}{\gamma}(\alpha^2-1)v_z^2+\theta_z^2\right)dzdt\\ &\quad\leq\int_\Omega\left(G(v_0)+\frac{1}{2}\theta_0^2\right)dz-(\alpha^2-1)\int_0^t\int_\Omega v_z\theta_z dzdt+\int_0^t\int_\Omega(\mu|\theta|-\theta^2)dzdt,\end{aligned} \tag{2.4}$$

as long as $-s_\alpha\leq v\leq s_\alpha$ in Q_t. Choosing $\alpha>1$ small enough and applying Young's inequality, it is possible to control the second integral on the right-hand side by the integrals on the left-hand side. This gives the estimates $v_z\in L^2(0,T;L^2(\Omega))$ and $\theta\in L^2(0,T;H^1_{\rm per}(\Omega))$. The inequality (2.4) is made rigorous in Lemma 2.6 for a time-discretized version of (1.6)–(1.7).

Still there remain two difficulties: the elliptic operator corresponding to (1.6)–(1.7) is not uniformly elliptic (since g' is only positive, but not uniformly positive in $\mathbb{R}$), and we have to deal with time derivatives in $g(v)$ (instead of having time *and* space derivatives in v). The first difficulty can be overcome by adding a small number $\varepsilon>0$ to the diffusion term containing $\nu g'(v)$ and passing to the limit $\varepsilon\to 0$ after solving the approximate problem. To overcome the second difficulty we approximate the system by a semidiscrete problem in time (backward Euler method). This method is also interesting from a numerical point of view; see, e.g., [13].

2.2. A semidiscrete problem. The main objective of this section is to prove that for given $\tau>0$ and $(\tilde w,\tilde\theta)\in(H^1_{\rm per}(\Omega))^2$, there exists a solution $(w,\xi)\in(H^1_{\rm per}(\Omega))^2$, satisfying $-s_\alpha\leq w\leq s_\alpha$ in Ω, of the problem

$$\frac{1}{\tau}(g(w)-g(\tilde w))-\big(\nu g'(w)w_z-(1-g(w)^2)\xi_z\big)_z=0, \tag{2.5}$$

$$\frac{1}{\tau}(\xi-\tilde\theta)-(\gamma g'(w)w_z+\xi_z)_z+\xi=\mu g(w)\quad\text{in }\Omega. \tag{2.6}$$

This system is a time-discretized version of (1.6)–(1.7). The function $g(s)$ is defined as in (2.1), but we allow for arguments $s \in \mathbb{R}$. We shall use the following notion of weak solution.

DEFINITION 2.1. *The pair* (w, ξ) *is called a* weak solution *of* (2.5)–(2.6) *if* $(w, \xi) \in (H^1_{\rm per}(\Omega))^2$, $-s_\alpha \le w \le s_\alpha$ *in* Ω, *the initial conditions in* (1.3) *are satisfied in the sense of* $(H^1_{\rm per}(\Omega))'$, *and for every* $(\varphi, \psi) \in (H^1_{\rm per}(\Omega))^2$ *we have*

$$\frac{1}{\tau}\int_\Omega (g(w) - g(\tilde w))\varphi dz + \int_\Omega \big(\nu g'(w)w_z - (1-g(w)^2)\xi_z\big)\varphi_z dz = 0, \tag{2.7}$$

$$\frac{1}{\tau}\int_\Omega (\xi - \tilde\theta)\psi dz + \int_\Omega (\gamma g'(w)w_z + \xi_z)\psi_z dz + \int_\Omega \xi\psi dz = \mu\int_\Omega g(w)\psi dz. \tag{2.8}$$

As explained in section 2.1, we approximate the system (2.5)–(2.6) by a system where an additional ellipticity constant $\varepsilon > 0$ is introduced: Find $(w, \xi) \in (H^1_{\rm per}(\Omega))^2$ such that in Ω

$$\frac{1}{\tau}(g(w) - g(\tilde w)) - \big((\nu g'(w) + \varepsilon)w_z - (1-g(w)^2)_+\xi_z\big)_z + \varepsilon w = 0, \tag{2.9}$$

$$\frac{1}{\tau}(\xi - \tilde\theta) - (\gamma g'(w)w_z + \xi_z)_z + \xi = \mu g(w), \tag{2.10}$$

where $s_+ = \max\{0, s\}$.

The following properties of the function g can be easily shown.

LEMMA 2.2. *The function* $g : \mathbb{R} \to (-\alpha, \alpha)$ *defined by* (2.1) *satisfies* $g \in C^\infty(\mathbb{R}) \cap W^{1,\infty}(\mathbb{R})$ *and*

$$0 < g' \le \alpha^2/\gamma \quad \text{in } \mathbb{R}, \qquad g' \ge (\alpha^2 - 1)/\gamma \quad \text{in } [-s_\alpha, s_\alpha]. \tag{2.11}$$

Fix $\alpha > 1$ *such that* $2(\alpha^2 - 1) \le \nu/2\gamma$ *and define* $h_1, h_2 : \mathbb{R} \to \mathbb{R}$ *by*

$$h_1 := \nu g' - \delta|\gamma g' - (1-g^2)_+|, \quad h_2 := 1 - \frac{1}{\delta}|\gamma g' - (1-g^2)_+|,$$

with $2(\alpha^2 - 1) \le \delta \le \nu/2\gamma$. *Then*

$$h_1 > 0, \quad h_2 \ge 1/2 \quad \text{in } \mathbb{R}, \quad \text{and} \quad h_1 \ge \frac{\nu}{2\gamma}(\alpha^2 - 1) \quad \text{in } [-s_\alpha, s_\alpha]. \tag{2.12}$$

We prove the existence of a solution of (2.9)–(2.10) using Schauder's fixed point theorem. In order to define the fixed point operator, we consider first the following linearized problem: Let $(\hat w, \hat\xi) \in (L^2(\Omega))^2$ be given and find $(w, \xi) \in (H^1_{\rm per}(\Omega))^2$ such that

$$-\big((\nu g'(\hat w) + \varepsilon)w_z - (1-g(\hat w)^2)_+\xi_z\big)_z + \varepsilon w = \frac{1}{\tau}(g(\tilde w) - g(\hat w)), \tag{2.13}$$

$$-(\gamma g'(\hat w)w_z + \xi_z)_z + \xi = \mu g(\hat w) + \frac{1}{\tau}(\tilde\theta - \hat\xi) \tag{2.14}$$

in Ω. The definition of a weak solution of problem (2.13)–(2.14) is similar to Definition 2.1.

LEMMA 2.3. *Let* $(\tilde w, \tilde\theta) \in (H^1_{\rm per}(\Omega))^2$ *and* $(\hat w, \hat\xi) \in (L^2(\Omega))^2$ *be given. Then there exists a unique weak solution of problem* (2.13)–(2.14).

Proof. We define the bilinear form $a : (H^1_{\rm per}(\Omega))^2 \times (H^1_{\rm per}(\Omega))^2 \to \mathbb{R}$,

$$a((w,\xi),(\varphi,\psi)) := \int_\Omega \big[\big((\nu g'(\hat w)+\varepsilon)w_z - (1-g(\hat w)^2)_+\xi_z\big)\varphi_z + \varepsilon w\varphi\big]\,dz \\ + \int_\Omega \big((\gamma g'(\hat w)w_z + \xi_z)\psi_z + \xi\psi\big)dz,$$

and the linear functional $f : (L^2(\Omega))^2 \to \mathbb{R}$,

$$f(\varphi,\psi) := \frac{1}{\tau}\int_\Omega \big((g(\tilde w)-g(\hat w))\varphi + (\tilde\theta-\hat\xi)\psi\big) + \mu\int_\Omega g(\hat w)\psi.$$

In order to apply the Lax–Milgram lemma, we have to check that a is continuous and coercive in $(H^1_{\rm per}(\Omega))^2 \times (H^1_{\rm per}(\Omega))^2$ and that f is continuous in $(L^2(\Omega))^2$. The continuity of a and f follows easily from the pointwise bounds of g and g' and the regularity of $\tilde w$, $\tilde\theta$, $\hat w$, and $\hat\xi$. For the coercivity of a we estimate

$$\begin{aligned} a((w,\xi),(w,\xi)) &= \int_\Omega \big((\nu g'(\hat w)+\varepsilon)|w_z|^2 + |\xi_z|^2 + \varepsilon|w|^2 + |\xi|^2\big)dz \\ &\quad + \int_\Omega \big((\gamma g'(\hat w) - (1-g(\hat w)^2)_+)w_z\xi_z\big)dz \\ &\ge \int_\Omega \big((\varepsilon + h_1(\hat w))|w_z|^2 + h_2(\hat w)|\xi_z|^2 + \varepsilon|w|^2 + |\xi|^2\big)dz \end{aligned}$$

using Young's inequality, where the functions h_1 and h_2 are defined in Lemma 2.2. The bounds (2.12) then imply that

$$a((w,\xi),(w,\xi)) \ge \min\{\varepsilon, 1/2\}\left(\|w\|^2_{H^1_{\rm per}(\Omega)} + \|\xi\|^2_{H^1_{\rm per}(\Omega)}\right),$$

and the coercivity of a is proved. ∎

LEMMA 2.4. *Let $(\tilde w,\tilde\theta) \in (H^1_{\rm per}(\Omega))^2$. Then there exists a unique weak solution of problem* (2.9)–(2.10).

Proof. We use the Schauder fixed point theorem. For this define the map $S : (L^2(\Omega))^2 \to (L^2(\Omega))^2$ by $S(\hat w,\hat\xi) = (w,\xi)$, where (w,ξ) is the weak solution of (2.13)–(2.14). We have to check that S is continuous and compact and that the set

$$\Lambda := \big\{u \in (L^2(\Omega))^2 : u = \lambda S(u)\big\}$$

for $\lambda \in [0,1]$ is bounded. The continuity of S follows by standard arguments. The compactness of S is just a consequence of the compactness of the embedding $H^1_{\rm per}(\Omega) \subset L^2(\Omega)$.

It remains to show that Λ is bounded. If $\lambda = 0$, then $\Lambda = \{(0,0)\}$ is trivially bounded. For $\lambda \in (0,1]$, the equation $S(\hat w,\hat\xi) = \frac{1}{\lambda}(\hat w,\hat\xi)$ is equivalent to

$$\int_\Omega \left(\big((\nu g'(\hat w)+\varepsilon)\hat w_z - (1-g(\hat w)^2)_+\hat\xi_z\big)_z\varphi_z + \varepsilon\hat w\varphi\right)dz = \frac{\lambda}{\tau}\int_\Omega (g(\tilde w)-g(\hat w))\varphi dz,$$

$$\int_\Omega \left((\gamma g'(\hat w)\hat w_z + \hat\xi_z)\psi_z + \hat\xi\psi\right)dz = \lambda\int_\Omega \left(\mu g(\hat w) + \frac{1}{\tau}(\tilde\theta-\hat\xi)\right)\psi dz.$$

Using $(\varphi,\psi)=(\hat{w},\hat{\xi})$ as a test function, adding the resulting integral identities, and applying Young's inequality as in (2.12), we obtain

$$\int_\Omega \big((\varepsilon+h_1(\hat{w}))|\hat{w}_z|^2+h_2(\hat{w})|\hat{\xi}_z|^2+\varepsilon|\hat{w}|^2+|\hat{\xi}|^2\big)dz = \frac{\lambda}{\tau}\int_\Omega (g(\tilde{w})-g(\hat{w}))\hat{w}dz + \lambda\int_\Omega\left(\mu g(\hat{w})+\frac{1}{\tau}(\tilde{\theta}-\hat{\xi})\right)\hat{\xi}dz.$$

Using again Young's inequality on the right-hand side of this equation and employing the estimate (2.12), we deduce

$$\int_\Omega \big(\varepsilon(|\hat{w}_z|^2+|\hat{w}|^2)+|\hat{\xi}_z|^2+|\hat{\xi}|^2\big)dz \le \frac{\lambda^2}{\tau^2\varepsilon}\int_\Omega (g(\tilde{w})-g(\hat{w}))^2dz+\frac{2\lambda^2}{\tau^2}\int_\Omega\tilde{\theta}^2dz + 2(\lambda\mu)^2\int_\Omega |g(\hat{w})|^2dz,$$

and since $g\in L^\infty(\mathbb{R})$, the assertion follows. ∎

In the following we derive uniform bounds for the solution of (2.9)–(2.10) which allow us to pass to the limit $\varepsilon\to 0$. This proves the existence of a solution of (2.5)–(2.6). We need the following auxiliary result, whose proof is standard.

LEMMA 2.5. *Let $\varphi\in C(\mathbb{R})$ be nondecreasing with $\varphi(0)=0$ and define $\Phi\in C^1(\mathbb{R})$ by $\Phi(s):=\int_0^s g'(\sigma)\varphi(\sigma)d\sigma$. Then it holds for all $s,t\in\mathbb{R}$ that*

$$\Phi(s)-\Phi(t)\le (g(s)-g(t))\varphi(s). \tag{2.15}$$

LEMMA 2.6. *Let $(\tilde{w},\tilde{\xi})\in (H^1_{\rm per}(\Omega))^2$ be such that $-s_\alpha\le\tilde{w}\le s_\alpha$ in Ω and let $(w_\varepsilon,\xi_\varepsilon)\in(H^1_{\rm per}(\Omega))^2$ be a solution of (2.9)–(2.10). Then the following estimates hold:*

$$-s_\alpha\le w_\varepsilon\le s_\alpha \quad \text{in } \Omega, \tag{2.16}$$

$$\int_\Omega\Big(G(w_\varepsilon)+\frac{1}{2}\xi_\varepsilon^2\Big)dz+C\tau\int_\Omega(|w_{\varepsilon z}|^2+|\xi_{\varepsilon z}|^2+|\xi_\varepsilon|^2)dz \le \int_\Omega\Big(G(\tilde{w})+\frac{1}{2}\tilde{\xi}^2\Big)dz+C'\tau \tag{2.17}$$

for some positive constants C, C' independent of ε and τ, and for G defined in (2.3).

In addition, there exists a subsequence of $(w_\varepsilon,\xi_\varepsilon)$ (not relabeled) such that $(w_\varepsilon,\xi_\varepsilon)\rightharpoonup(w,\xi)$ weakly in $(H^1_{\rm per}(\Omega))^2$ and strongly in $(L^2(\Omega))^2$ as $\varepsilon\to 0$, and (w,ξ) is a weak solution of problem (2.5)–(2.6).

Proof. We use $\varphi(w_\varepsilon):=\max(w_\varepsilon-s_\alpha,0)$ as a test function in the weak formulation of (2.9). Since φ is increasing and $\varphi(0)=0$ we can employ Lemma 2.5. Let Φ be defined as in Lemma 2.5. Then, together with the identities $(1-g(s)^2)_+\varphi'(s)=0$ for all $s\in\mathbb{R}$ and $\Phi(\tilde{w})=0$, we obtain

$$0\ge\frac{1}{\tau}\int_\Omega(g(w_\varepsilon)-g(\tilde{w}))\varphi(w_\varepsilon)dx\ge\int_\Omega(\Phi(w_\varepsilon)-\Phi(\tilde{w}))dx=\int_\Omega\Phi(w_\varepsilon)dx.$$

This implies $\Phi(w_\varepsilon)=0$ and therefore $w_\varepsilon\le s_\alpha$ in Ω. In a similar way we deduce $w_\varepsilon\ge -s_\alpha$ in Ω. Observe that these bounds imply that $(1-g(w_\varepsilon)^2)_+=1-g(w_\varepsilon)^2$ in Ω.

Now we use $(w_\varepsilon, \xi_\varepsilon)$ as a test function in the weak formulation of problem (2.9)–(2.10). Adding the corresponding integral identities and using property (2.15) we get, after multiplication by τ,

$$\int_\Omega \Big(G(w_\varepsilon) + \frac{1}{2}\xi_\varepsilon^2\Big)dz + \tau\int_\Omega \big(h_1(w_\varepsilon)|w_{\varepsilon z}|^2 + h_2(w_\varepsilon)|\xi_{\varepsilon z}|^2 + |\xi_\varepsilon|^2\big)dz \\ \le \mu\tau\int_\Omega g(w_\varepsilon)\xi_\varepsilon dz + \int_\Omega \Big(G(\tilde{w}) + \frac{1}{2}\tilde{\xi}^2\Big)dz.$$

Applying Young's inequality and the bounds (2.11) and (2.12) for g', h_1, and h_2, we deduce (2.17).

Finally, the uniform estimates (2.16) and (2.17) imply the existence of a subsequence (not relabeled) of $(w_\varepsilon, \xi_\varepsilon)$ and of a pair $(w, \xi) \in (H^1_{\rm per}(\Omega))^2$ such that, as $\varepsilon \to 0$,

$$w_\varepsilon \overset{*}{\rightharpoonup} w \quad \text{weakly* in } L^\infty(\Omega), \tag{2.18}$$

$$\begin{aligned} w_{\varepsilon z} &\rightharpoonup w_z \quad \text{weakly in } L^2(\Omega), \\ \xi_\varepsilon &\rightharpoonup \xi \quad \text{weakly in } H^1_{\rm per}(\Omega). \end{aligned} \tag{2.19}$$

In fact, the convergences (2.18) and (2.19) imply $w_\varepsilon \rightharpoonup w$ weakly in $H^1_{\rm per}(\Omega)$ and thus, by the compactness of the embedding $H^1_{\rm per}(\Omega) \subset L^2(\Omega)$, we deduce for a subsequence, as $\varepsilon \to 0$, $w_\varepsilon \to w$, and $\xi_\varepsilon \to \xi$ strongly in $L^2(\Omega)$ and a.e. in Ω. These convergence results and the continuity of g and g' allow us to pass to the limit $\varepsilon \to 0$ in the weak formulation of problem (2.9)–(2.10) and to identify (w, ξ) as a weak solution of (2.5)–(2.6). ∎

2.3. End of the proof of Theorem 1.1. Let $T > 0$ and $N \in \mathbb{N}$ be given and let $\tau = T/N$ be the time step. We define recursively pairs $(v^k, \theta^k) \in (H^1_{\rm per}(\Omega))^2$, $k = 1, \ldots, N$, as the weak solution of the problem (2.5)–(2.6) corresponding to the data $(\tilde{w}, \tilde{\theta}) = (v^{k-1}, \theta^{k-1})$, and with $(v^0, \theta^0) = (v_0, \theta_0)$. Then we define the piecewise constant functions

$$v^\tau(x,t) := v^k(x) \quad \text{and} \quad \theta^\tau(x,t) := \theta^k(x) \quad \text{if } (x,t) \in \Omega \times ((k-1)\tau, k\tau]$$

for $k = 1, \ldots, N$ and introduce the discrete entropies

$$\eta^k := \int_\Omega \Big(G(v^k) + \frac{1}{2}|\theta^k|^2\Big)dz, \quad \eta^\tau(t) := \int_\Omega \Big(G(v^\tau(\cdot,t)) + \frac{1}{2}|\theta^\tau(\cdot,t)|^2\Big)dz. \tag{2.20}$$

We have the following consequence of Lemma 2.6.

COROLLARY 2.7. *There exist uniform bounds with respect to τ for the norms*

$$\|\eta^\tau\|_{L^\infty(0,T)}, \quad \|v^\tau\|_{L^2(0,T;H^1_{\rm per}(\Omega))}, \quad \|g(v^\tau)\|_{L^2(0,T;H^1_{\rm per}(\Omega))}, \quad \text{and} \quad \|\theta^\tau\|_{L^2(0,T;H^1_{\rm per}(\Omega))}.$$

In addition,

$$-s_\alpha \le v^\tau \le s_\alpha \quad \text{in } Q_T = \Omega \times (0,T). \tag{2.21}$$

Proof. From the "entropy" inequality (2.17) we obtain

$$\eta^m - \eta^0 = \sum_{k=1}^m (\eta^k - \eta^{k-1}) \le C'm\tau - C\tau\sum_{k=1}^m \int_\Omega (|v_z^k|^2 + |\theta_z^k|^2 + |\theta^k|^2)dz$$

for $m = 1, \dots, N$. Taking the maximum over m yields

$$\|\eta^\tau\|_{L^\infty(0,T)} + C\int_{Q_T} \big(|v_z^\tau|^2 + |\theta_z^\tau|^2 + |\theta^\tau|^2\big)dzdt \le \eta^0 + C'T.$$

Since both g and g' are smooth and bounded we also deduce the estimate for the norm $\|g(v^\tau)\|_{L^2(0,T;H^1_{\rm per}(\Omega))}$. Finally, (2.21) follows directly from (2.16). ∎

We need uniform estimates of the time derivatives. For this, we introduce the shift operator and linear interpolations in time. For $t \in ((k-1)\tau, k\tau]$, $k = 1, \dots, N$, we define $\sigma_\tau v^\tau(\cdot,t) := v^{k-1}$ and $\sigma_\tau\theta^\tau(\cdot,t) := \theta^{k-1}$ in Ω. Setting $\delta t := (t/\tau - (k-1)) \in [0,1]$, we introduce

$$\tilde g^\tau := g(\sigma_\tau v^\tau) + \delta t\big(g(v^\tau) - g(\sigma_\tau v^\tau)\big), \quad \tilde\theta^\tau := \sigma_\tau\theta^\tau + \delta t\big(\theta^\tau - \sigma_\tau\theta^\tau\big) \tag{2.22}$$

in Q_T.

LEMMA 2.8. *There exist uniform bounds with respect to τ for the norms*

$$\|\tilde g_t^\tau\|_{L^2(0,T;(H^1_{\rm per}(\Omega))')}, \quad \|\tilde g^\tau\|_{L^2(0,T;H^1_{\rm per}(\Omega))\cap L^\infty(Q_T)},$$
$$\|\tilde\theta_t^\tau\|_{L^2(0,T;(H^1_{\rm per}(\Omega))')}, \quad \textit{and} \quad \|\tilde\theta^\tau\|_{L^2(0,T;H^1_{\rm per}(\Omega))}.$$

Proof. From the definition (2.22) of $\tilde g^\tau$ and (2.5) we compute

$$\tilde g_t^\tau = \frac{1}{\tau}\big(g(v^\tau) - g(\sigma_\tau v^\tau)\big) = \big(\nu g'(v^\tau)v_z^\tau - (1 - g(v^\tau)^2)\theta_z^\tau\big)_z.$$

Using the boundedness of g' in $\mathbb{R}$ and Corollary 2.7 we obtain a uniform bound for $\|\tilde g_t^\tau\|_{L^2((0,T;H^1_{\rm per})')}$. Moreover, since g is bounded, it is clear that $\tilde g^\tau \in L^\infty(Q_T)$ for any $\tau \ge 0$. We also have

$$\tilde g_z^\tau = \delta t g'(v^\tau)v_z^\tau + (1-\delta t)g'(\sigma_\tau v^\tau)(\sigma_\tau v^\tau)_z. \tag{2.23}$$

Since $(\sigma_\tau v^\tau)_z = \sigma_\tau v_z^\tau$, the $L^\infty(Q_T)$ bound for $\tilde g^\tau$ together with (2.23) and Corollary 2.7 implies a uniform bound for $\|\tilde g^\tau\|_{L^2(0,T;H^1_{\rm per}(\Omega))}$. In a similar way we obtain uniform estimates for $\tilde\theta^\tau$. ∎

Proof of Theorem 1.1. The functions v^τ, θ^τ, $\tilde g^\tau$, $\tilde\theta^\tau$ satisfy the weak formulation

$$\int_0^T \langle \tilde g_t^\tau, \varphi\rangle dt + \int_{Q_T} \big(\nu g'(v^\tau)v_z^\tau - (1-g(v^\tau)^2)\theta_z^\tau\big)\varphi_z dzdt = 0, \tag{2.24}$$

$$\int_0^T \langle \tilde\theta_t, \psi\rangle dt + \int_{Q_T} \big(\gamma g'(v^\tau)v_z^\tau + \theta_z^\tau\big)\psi_z dzdt + \int_{Q_T} \theta^\tau\psi dydt$$
$$= \mu\int_{Q_T} g(v^\tau)\psi dzdt \tag{2.25}$$

for any $\varphi, \psi \in L^2(0,T;H^1_{\rm per}(\Omega))$. The estimates of Lemma 2.8 allow us to extract a subsequence (not relabeled) such that, as $\tau \to 0$,

$$\tilde g_t^\tau \rightharpoonup u_t \quad \text{weakly in } L^2(0,T;(H^1_{\rm per}(\Omega))'), \tag{2.26}$$

$$\tilde g^\tau \rightharpoonup u \quad \text{weakly in } L^2(0,T;H^1_{\rm per}(\Omega)), \tag{2.27}$$

$$\tilde g^\tau \overset{*}{\rightharpoonup} u \quad \text{weakly* in } L^\infty(Q_T),$$

$$\tilde\theta_t^\tau \rightharpoonup \theta_t \quad \text{weakly in } L^2(0,T;(H^1_{\rm per}(\Omega))'), \tag{2.28}$$

$$\tilde\theta^\tau \rightharpoonup \theta \quad \text{weakly in } L^2(0,T;H^1_{\rm per}(\Omega)). \tag{2.29}$$

The compact embedding $H^1_{\rm per}(\Omega)\subset L^\infty(\Omega)$, the convergence results (2.26)–(2.29), and Aubin's lemma [23] imply, up to a subsequence,

$$\begin{aligned}(2.30)\qquad \tilde g^\tau &\to u \quad \text{strongly in } L^2(0,T;L^\infty(\Omega)),\\ \tilde\theta^\tau &\to \theta \quad \text{strongly in } L^2(0,T;L^\infty(\Omega)).\end{aligned}$$

Moreover, Corollary 2.7 yields the existence of a subsequence such that

$$\begin{aligned}(2.31)\qquad v^\tau &\rightharpoonup v \quad \text{weakly in } L^2(0,T;H^1_{\rm per}(\Omega)),\\ v^\tau &\overset{*}{\rightharpoonup} v \quad \text{weakly* in } L^\infty(Q_T),\\ g(v^\tau) &\rightharpoonup \hat u \quad \text{weakly in } L^2(0,T;H^1_{\rm per}(\Omega)),\\ \theta^\tau &\rightharpoonup \hat\theta \quad \text{weakly in } L^2(0,T;H^1_{\rm per}(\Omega)).\end{aligned}$$

It holds that $\tilde g^\tau - g(v^\tau)=\tau(\delta t-1)\tilde g_t^\tau$, and therefore, by Lemma 2.8,

$$(2.32)\qquad \|\tilde g^\tau - g(v^\tau)\|_{L^2(0,T;(H^1_{\rm per})')}\to 0 \quad \text{as } \tau\to 0.$$

Hence, $u=\hat u$. In a similar way we obtain $\theta=\hat\theta$. Finally,

$$\begin{aligned}&\|g(v^\tau)-u\|_{L^1(0,T;L^2(\Omega))}\\ &\le \|g(v^\tau)-\tilde g^\tau\|_{L^1(0,T;L^2(\Omega))}+\|\tilde g^\tau-u\|_{L^1(0,T;L^2(\Omega))}\\ &\le \|g(v^\tau)-\tilde g^\tau\|^{1/2}_{L^1(0,T;(H^1_{\rm per}(\Omega))')}\|g(v^\tau)-\tilde g^\tau\|^{1/2}_{L^1(0,T;H^1_{\rm per}(\Omega))}\\ &\quad +\|\tilde g^\tau-u\|_{L^1(0,T;L^2(\Omega))}\\ &\le C\|g(v^\tau)-\tilde g^\tau\|^{1/2}_{L^2(0,T;(H^1_{\rm per}(\Omega))')}+\|\tilde g^\tau-u\|_{L^1(0,T;L^2(\Omega))}\\ (2.33)\qquad &\to 0\end{aligned}$$

as $\tau\to 0$. Therefore, $g(v^\tau)\to u$ strongly in $L^1(0,T;L^2(\Omega))$ and a.e. in Q_T. Now, letting $\tau\to 0$ in (2.24)–(2.25), we obtain, for $\varphi,\psi\in L^2(0,T;H^1_{\rm per}(\Omega))$,

$$(2.34)\qquad \int_0^T\langle u_t,\varphi\rangle dt+\int_{Q_T}\big(\nu u_z-(1-u^2)\theta_z\big)\varphi_z dzdt=0,$$

$$(2.35)\qquad \int_0^T\langle \theta_t,\psi\rangle dt+\int_{Q_T}\big(\gamma u_z+\theta_z\big)\psi_z dzdt+\int_{Q_T}\theta\psi dz=\mu\int_{Q_T}u\psi dzdt.$$

This proves Theorem 1.1. □

3. Proof of Theorem 1.2. Let (u_1,θ_1) and (u_2,θ_2) be two weak solutions of (1.1)–(1.3) with the same initial data satisfying (1.10) and $\theta_1\in L^\infty(0,T;H^1_{\rm per}(\Omega))$. Set $Q_t=\Omega\times(0,t)$. The equations satisfied by $u=u_1-u_2$ and $\theta=\theta_1-\theta_2$ read

$$(3.1)\qquad u_t-\nu u_{zz}+\theta_{zz}=\big((u_1+u_2)u\theta_{1z}+u_2^2\theta_z\big)_z,$$

$$(3.2)\qquad \theta_t-\theta_{zz}+\theta=\gamma u_{zz}+\mu u.$$

Take u and θ as test functions in the weak formulations of (3.1) and (3.2), respectively, and add (3.2), multiplied by some number $a>0$, and (3.1) to obtain

$$\begin{aligned}&\frac12\int_\Omega(u(t)^2+a\theta(t)^2)dz+\int_{Q_t}(\nu u_z^2+a\theta_z^2+a\theta^2)dzdt\\ (3.3)\qquad &=\int_{Q_t}(1-a\gamma-u_2^2)u_z\theta_z dzdt+a\mu\int_{Q_t}u\theta dzdt-\int_{Q_t}(u_1+u_2)u\theta_{1z}u_z dzdt.\end{aligned}$$

We apply Young's inequality to the second integral on the right-hand side:

$$a\mu\int_{Q_t} u\theta dzdt \le \frac{a\mu^2}{2}\int_{Q_t} u^2 dzdt + \frac{a}{2}\int_{Q_t}\theta^2 dzdt.$$

For the third integral on the right-hand side of (3.3) we use the Gagliardo–Nirenberg inequality

$$\|u\|_{L^\infty(\Omega)} \le C_0\|u\|_{H^1(\Omega)}^{1/2}\|u\|_{L^2(\Omega)}^{1/2} \quad \forall u\in H^1(0,L)$$

and the Young inequality

$$x^{1/2}y^{3/2} \le \frac{\varepsilon}{2}x^2 + C(\varepsilon)y^2 \quad \forall x,y\ge 0,\ \varepsilon>0.$$

Then, with the abbreviation $C_1 = 2C_0\|\theta_{1z}\|_{L^\infty(0,T;L^2(\Omega))} < \infty$ and $|u_1|, |u_2| \le 1$,

$$\begin{aligned}
&\int_{Q_t}(u_1+u_2)u\theta_{1z}u_z dzdt \\
&\le 2\|u\|_{L^2(0,t;L^\infty(\Omega))}\|\theta_{1z}\|_{L^\infty(0,t;L^2(\Omega))}\|u_z\|_{L^2(0,t;L^2(\Omega))} \\
&\le C_1\|u\|_{L^2(0,t;L^2(\Omega))}^{1/2}\left(\|u\|_{L^2(Q_t)}^2 + \|u_z\|_{L^2(Q_t)}^2\right)^{1/4}\|u_z\|_{L^2(0,t;L^2(\Omega))} \\
&\le C_1\left(\|u\|_{L^2(Q_t)}\|u_z\|_{L^2(Q_t)} + \|u\|_{L^2(Q_t)}^{1/2}\|u_z\|_{L^2(Q_t)}^{3/2}\right) \\
&\le \frac{\varepsilon}{2}\|u_z\|_{L^2(Q_t)}^2 + \frac{C_1^2}{2\varepsilon}\|u\|_{L^2(Q_t)}^2 + \frac{\varepsilon}{2}\|u_z\|_{L^2(Q_t)}^2 + C(\varepsilon)C_1^4\|u\|_{L^2(Q_t)}^2.
\end{aligned}$$

With these inequalities we can estimate (3.3) as

$$\begin{aligned}
&\frac12\left(\|u(t)\|_{L^2(\Omega)}^2 + a\|\theta(t)\|_{L^2(\Omega)}^2\right) + \frac{a}{2}\|\theta\|_{L^2(Q_t)}^2 \\
&\le -\int_{Q_t}\left(-(|1-a\gamma|+1)|u_z||\theta_z| + (\nu-\varepsilon)u_z^2 + a\theta_z^2\right) \\
&\quad + \left(\frac{a\mu^2}{2} + \frac{C_1^2}{2\varepsilon} + C(\varepsilon)C_1^4\right)\|u\|_{L^2(Q_t)}^2. \qquad (3.4)
\end{aligned}$$

It can be easily seen that the quadratic form

$$A(x,y) = -(|1-a\gamma|+1)xy + (\nu-\varepsilon)x^2 + ay^2, \quad x,y\ge 0,$$

is positive definite if we choose $a = 1/\gamma$ and $\varepsilon = \nu - \gamma/4 > 0$ (since $\gamma < 4\nu$ by assumption). Then Gronwall's lemma applied to (3.4) implies that $u(t) = \theta(t) = 0$ in Ω for any $t > 0$. This proves Theorem 1.2. □

4. Proof of Theorem 1.3. Let (u,θ) be a weak solution of (1.1)–(1.3) given by Theorem 1.1. Let $\alpha > 1$ and set

$$c_0 = \frac12\int_0^L\left(\gamma(u_0+1)\ln\frac{1+u_0}{1+\bar u} + \gamma(1-u_0)\ln\frac{1-u_0}{1-\bar u} + (\theta_0-\bar\theta)\right)dz. \qquad (4.1)$$

Notice that c_0 is well defined even if $u_0(z) = \pm 1$. For the proof of Theorem 1.3 we need the following lemma.

LEMMA 4.1. *Define the function* $\psi : [-1,1] \to \mathbb{R}$ *by*

$$\psi(u) = \frac{\gamma}{2\alpha} \ln\left(\frac{\alpha+u}{\alpha+\bar{u}}\,\frac{\alpha-\bar{u}}{\alpha-u}\right).$$

Then the function $\Psi : [-1,1] \to \mathbb{R}$*, defined by*

$$\Psi(u) = \frac{\gamma}{2\alpha}(\alpha+u)\ln\frac{\alpha+u}{\alpha+\bar{u}} + \frac{\gamma}{2\alpha}(\alpha-u)\ln\frac{\alpha-u}{\alpha-\bar{u}},$$

satisfies for all $u \in [-1,1]$

$$\Psi'(u) = \psi(u), \quad \Psi''(u) = \frac{\gamma}{\alpha^2-u^2}, \quad \Psi(u) \ge \frac{\gamma}{2\alpha^2}(u-\bar{u})^2.$$

The lemma follows from Taylor expansion around $\bar{u}$:

$$\Psi(u) = \Psi(\bar{u}) + \Psi'(\bar{u})(u-\bar{u}) + \frac{1}{2}\Psi''(\xi)(u-\bar{u})^2 \ge \frac{\gamma}{2\alpha^2}(u-\bar{u})^2.$$

Proof of Theorem 1.3. We use $\psi(u) \in L^\infty(Q_T) \cap L^2(0,T;H^1_{\rm per}(\Omega))$ and $\theta - \bar{\theta} \in L^2(0,T;H^1_{\rm per}(\Omega))$ as test functions in the weak formulation of (1.1) and (1.2), respectively, and add the resulting equations:

$$\begin{aligned}(4.2)\qquad &\int_\Omega \left(\Psi(u(t)) + \frac{1}{2}(\theta(t)-\bar{\theta})^2\right)dz + \int_{Q_t}(\nu\psi'(u)u_z^2 + \theta_z^2)dzdt \\ &= \int_\Omega \left(\Psi(u_0) + \frac{1}{2}(\theta_0-\bar{\theta})^2\right)dz + \int_{Q_t}((1-u^2)\psi'(u)-\gamma)u_z\theta_z dzdt \\ &\quad + \int_{Q_t}(\mu u-\theta)(\theta-\bar{\theta})dzdt.\end{aligned}$$

For the second integral on the right-hand side we use Young's inequality:

$$\begin{aligned}&\int_{Q_t}((1-u^2)\psi'(u)-\gamma)u_z\theta_z dzdt = \gamma\int_{Q_t}\frac{1-\alpha^2}{\alpha^2-u^2}u_z\theta_z dzdt \\ &\le \frac{\nu\gamma}{2}(\alpha^2-1)^{1/2}\int_{Q_t}\frac{u_z^2}{\alpha^2-u^2}dzdt + \frac{\gamma}{2\nu}(\alpha^2-1)^{3/2}\int_{Q_t}\frac{\theta_z^2}{\alpha^2-u^2}dzdt \\ &\le \frac{\nu\gamma}{2}(\alpha^2-1)^{1/2}\int_{Q_t}\frac{u_z^2}{\alpha^2-u^2}dzdt + \frac{\gamma}{2\nu}(\alpha^2-1)^{1/2}\int_{Q_t}\theta_z^2 dzdt.\end{aligned}$$

Since $\mu\bar{u} = \bar{\theta}$, the last integral on the right-hand side of (4.2) becomes

$$\begin{aligned}\int_{Q_t}(\mu u-\theta)(\theta-\bar{\theta})dzdt &= \mu\int_{Q_t}(u-\bar{u})(\theta-\bar{\theta})dzdt - \int_{Q_t}(\theta-\bar{\theta})^2 dzdt \\ &\le \frac{\mu^2\delta}{2}\int_{Q_t}(u-\bar{u})^2 dzdt + \left(\frac{1}{2\delta}-1\right)\int_{Q_t}(\theta-\bar{\theta})^2 dzdt,\end{aligned}$$

where we choose

$$\frac{L^2}{2(L^2+2)} < \delta < \frac{4\nu\gamma}{\mu^2 L^2}.$$

This is possible by assumption (1.11). We employ Lemma 4.1 to estimate the first integral on the left-hand side of (4.2):

$$\int_\Omega \left(\Psi(u(t)) + \frac{1}{2}(\theta(t)-\bar\theta)^2\right) dz \geq \int_\Omega \left(\frac{\gamma}{2\alpha^2}(u(t)-\bar u)^2 + \frac{1}{2}(\theta(t)-\bar\theta)^2\right) dz.$$

Finally, the second term on the left-hand side of (4.2) can be estimated by using the Poincaré inequality

$$\|v-\bar v\|_{L^2(\Omega)} \leq \frac{L}{\sqrt{2}}\|v_z\|_{L^2(\Omega)} \quad \forall v \in H^1_{\rm per}(\Omega) \text{ with } \bar v = \int_0^L v(z)dz.$$

We obtain

$$\int_{Q_t} (\nu\psi'(u)u_z^2 + \theta_z^2)dzdt \geq \int_{Q_t} \left(\frac{2\nu\gamma}{L^2}\,\frac{(u-\bar u)^2}{\alpha^2-u^2} + \frac{2}{L^2}(\theta-\bar\theta)^2\right) dzdt.$$

Putting the above estimates together, we infer from (4.2) that

$$\begin{aligned}(4.3)\qquad &\int_\Omega \left(\frac{\gamma}{2\alpha^2}(u(t)-\bar u)^2 + \frac{1}{2}(\theta(t)-\bar\theta)^2\right) dz \\ &\leq c_0^2 + \int_{Q_t}\left(\frac{\mu^2\delta}{2} - \frac{2\nu\gamma}{L^2} + \frac{\nu\gamma}{L^2}(\alpha^2-1)^{1/2}\right)\frac{(u-\bar u)^2}{\alpha^2-u^2}dzdt \\ &\quad + \left(\frac{1}{2\delta} - \frac{L^2+2}{L^2} + \frac{\gamma}{\nu L^2}(\alpha^2-1)^{1/2}\right)\int_{Q_t}(\theta-\bar\theta)^2 dzdt.\end{aligned}$$

Observing that

$$\frac{(u-\bar u)^2}{\alpha^2-u^2} \geq \frac{(u-\bar u)^2}{\alpha^2},$$

we can let $\alpha \to 1$ in (4.3) to obtain

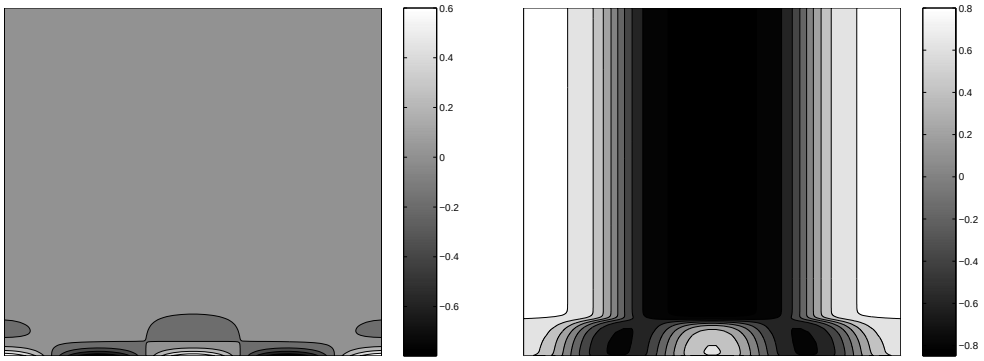
$$\begin{aligned}\frac{1}{2}\int_\Omega \left(\gamma(u(t)-\bar u)^2 + (\theta(t)-\bar\theta)^2\right) dz \leq c_0^2 &- \int_{Q_t}\left(\frac{2\nu\gamma}{L^2} - \frac{\mu^2\delta}{2}\right)(u-\bar u)^2 dzdt \\ &- \left(\frac{L^2+2}{L^2} - \frac{1}{2\delta}\right)\int_{Q_t}(\theta-\bar\theta)^2 dzdt.\end{aligned}$$

Defining

$$(4.4)\qquad \delta_1 = \frac{4\nu}{L^2} - \frac{\mu^2\delta}{\gamma} > 0, \quad \delta_2 = \frac{2(L^2+2)}{L^2} - \frac{1}{\delta} > 0,$$

the theorem follows from Gronwall's lemma. ∎

5. Numerical examples. In this section we illustrate by numerical experiments the long-time coarsening of the segregation bands in a drum. For the numerical discretization, we use a time-discretized version of (1.6)–(1.7) (backward Euler scheme), as motivated by the existence analysis of section 2,instead of discretizing (1.1)–(1.2)



(a) $L = 1$, $0 \le t \le 0.159$. (b) $L = 30$, $0 \le t \le 3049$.

FIG. 5.1. $\gamma = 2$, $\mu = 3$, $\nu = 2$, $u_0(z) = 0.8\cos(4\pi z/L)$, $N = 50$.

directly. The space discretization is performed by using finite differences. The nonlinear system is solved by a simple fixed point strategy.

In the following examples, we illustrate the segregation behavior of the component u of the solutions of (1.1)–(1.2). The behavior relies on three important conditions. First, condition (1.11) ensures the convergence of u to a constant steady state. Second, the authors of [3] conjectured that the condition $\mu > \nu$ is a necessary condition to have segregation. This conjecture arises from a linear stability analysis sketched in [3], showing that perturbations of the form $\exp(\lambda t + 2\pi z/\ell)$, where $\lambda \in \mathbb{R}$ and $\ell > 0$ is the wavelength of the perturbations, are unstable if $\mu > \nu + 4\pi^2(\nu+\gamma)/\ell^2$. Therefore, this instability is captured only if the length L of the domain satisfies the third condition

$$L > 2\pi\sqrt{(\gamma+\nu)/(\mu-\nu)}. \tag{5.1}$$

In Figure 5.1 we present the behavior of u in the (z,t)-plane for two different domain lengths. The number of grid points is $N = 50$. The parameters in Figure 5.1(a) satisfy $\mu > \nu$ and (1.11) but *not* (5.1). We observe convergence of u to a constant steady state. We expect this behavior in view of Theorem 1.3. This example shows that the condition $\mu > \nu$ is *not* sufficient for segregation. The parameters in Figure 5.1(b) satisfy the segregation condition (5.1) but not (1.11). The granular materials segregate since the length of the cylinder is large enough, as claimed by the linear stability analysis.

Figure 5.2 shows that (1.11) is a sufficient but not necessary condition to have nonsegregation. Indeed, the parameters are chosen such that (1.11) is not satisfied, but the granular materials do not segregate.

A more detailed view of the same segregation phenomena as above but with a larger number of bands is presented in Figure 5.3. The parameters do not satisfy (1.11) but (5.1) holds. Thus, we expect segregation. The initial short-wave perturbations produce decaying standing waves (Figure 5.3(a)). The segregated bands emerge, and we observe metastable long-wave bands. Finally, after a long time, the system segregates again (Figure 5.3(b)). This illustrates the very slow coarsening of the band structure (see [3]).

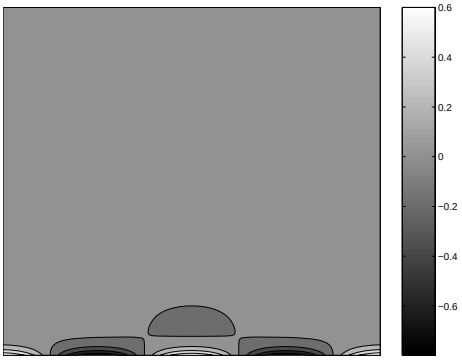
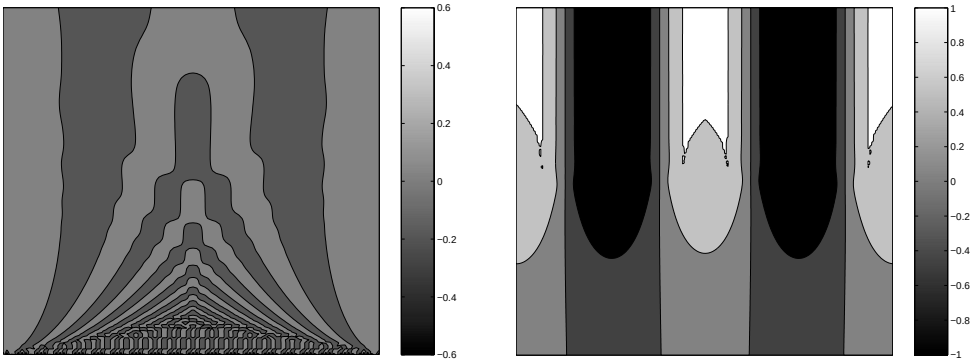


FIG. 5.2. $L = 4$, $\gamma = 2$, $\mu = 2$, $\nu = 3$, $u_0(z) = 0.8\cos(4\pi z/L)$, $N = 50$.



(a) $0 \le t \le 1.06$. (b) $4.8 \le t \le 8.9$.

FIG. 5.3. $\gamma = 100$, $\mu = 40$, $\nu = 0.5$, $L = 30$, $u_0(z) = 0.75\cos(80\pi z/L)$, $N = 1000$.

Acknowledgment. The authors thank the anonymous referees for their suggestions.

REFERENCES

[1] H. W. ALT AND S. LUCKHAUS, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311–341.

[2] I. ARANSON AND L. TSIMRING, *Dynamics of axial separation in long rotating drums*, Phys. Rev. Lett., 82 (1992), pp. 4643–4646.

[3] I. ARANSON, L. TSIMRING, AND V. VINOKUR, *Continuum theory of axial segregation in a long rotating drum*, Phys. Rev. E, 60 (1999), pp. 1975–1987.

[4] J.A. CARRILLO, A. JÜNGEL, P. MARKOWICH, G. TOSCANI, AND A. UNTERREITER, *Entropy dissipation methods for degenerate parabolic problems and generalized Sobolev inequalities*, Monatsh. Math., 133 (2001), pp. 1–82.

[5] K. CHOO, M. BAKER, T. MOLTENO, AND S. MORRIS, *The dynamics of granular segregation patterns in a long drum mixer*, Phys. Rev. E, 58 (1998), pp. 6115–6123.

[6] S. CHAKRABORTY, P. NOTT, AND R. PRAKASH, *Analysis of radial segregation of granular mixtures in a rotating drum*, Eur. Phys. J. E, 1 (2000), pp. 265–273.

[7] P. Degond, S. Génieys, and A. Jüngel, *An existence and uniqueness result for the stationary energy-transport model in semiconductor theory*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 29–34.
[8] P. Degond, S. Génieys, and A. Jüngel, *Symmetrization and entropy inequality for general diffusion systems*, C. R. Acad. Sci. Paris Sér. I Math., 325 (1997), pp. 963–968.
[9] P. Degond, S. Génieys, and A. Jüngel, *A system of parabolic equations in nonequilibrium thermodynamics including thermal and electric effects*, J. Math. Pure Appl., 76 (1997), pp. 991–1015.
[10] C. Dury and G. Ristow, *Radial segregation in a two-dimensional rotating drum*, J. Phys. I France, 7 (1997), pp. 737–745.
[11] H. Gajewski and H. Zacharias, *Global behaviour of a reaction-diffusion system modelling chemotaxis*, Math. Nachr., 195 (1998), pp. 77–114.
[12] G. Galiano and A. Jüngel, Work in preparation, 2002.
[13] G. Galiano, A. Jüngel, and M. Garzón, *Semi-discretization in time and numerical convergence of a nonlinear cross-diffusion population model*, Numer. Math., 93 (2003), pp. 655–673.
[14] J. Gray, *Granular flow in partially filled slowly rotating drums*, J. Fluid Mech., 441 (2001), pp. 1–29.
[15] S. Kawashima and Y. Shizuta, *On the normal form of the symmetric hyperbolic-parabolic systems associated with the conservation laws*, Tohoku Math. J. (2), 40 (1988), pp. 449–464.
[16] D. Levine, *Axial segregation of granular materials*, Chaos, 9 (1999), pp. 573–580.
[17] Y. Lou and W.-M. Ni, *Diffusion, self-diffusion and cross-diffusion*, J. Differential Equations, 131 (1996), pp. 79–131.
[18] M. Mimura and K. Kawasaki, *Spatial segregation in competitive interaction-diffusion equations*, J. Math. Biol., 9 (1980), pp. 49–64.
[19] J. Ottino and D. Khakhar, *Mixing and segregation of granular materials*, in Annual Review of Fluid Mechanics, Vol. 32, J. Lumley et al., eds., Palo Alto, CA, 2000, pp. 55–91.
[20] D. Rapaport, *Simulational studies of axial granular segregation in a rotating cylinder*, Phys. Rev. E, 65 (2002), 061306/1-11.
[21] G. Ristow, *Pattern Formation in Granular Materials*, Springer-Verlag, New York, 2000.
[22] N. Shigesada, K. Kawasaki, and E. Teramoto, *Spatial segregation of interacting species*, J. Theor. Biol., 79 (1979), pp. 83–99.
[23] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Math. Pura Appl., 146 (1987), pp. 65–96.
[24] O. Zik, S. Lipson, S. Shtrikman, and J. Stavans, *Rotationally induced segregation of granular materials*, Phys. Rev. Lett., 73 (1994), pp. 644–647.

SIAM J. MATH. ANAL.
Vol. 35, No. 3, pp. 579–595

ON THE CAUCHY PROBLEM FOR A NONLINEAR KOLMOGOROV EQUATION*

ANDREA PASCUCCI† AND SERGIO POLIDORO†

Abstract. We consider the Cauchy problem related to the partial differential equation

$$Lu \equiv \Delta_x u + h(u)\partial_y u - \partial_t u = f(\cdot, u),$$

where $(x, y, t) \in \mathbb{R}^N \times \mathbb{R} \times]0, T[$, which arises in mathematical finance and in the theory of diffusion processes. We study the regularity of solutions regarding L as a perturbation of an operator of Kolmogorov type. We prove the existence of local classical solutions and give some sufficient conditions for global existence.

Key words. nonlinear degenerate Kolmogorov equation, interior regularity, Hörmander operators

AMS subject classifications. 35K57, 35K65, 35K70

DOI. 10.1137/S0036141001399349

1. Introduction.

In this paper we study the Cauchy problem

$$Lu = f(\cdot, u) \qquad \text{in } S_T \equiv \mathbb{R}^{N+1} \times]0, T[, \tag{1.1}$$

$$u(\cdot, 0) = g \qquad \text{in } \mathbb{R}^{N+1}, \tag{1.2}$$

where L is the nonlinear operator defined by

$$Lu = \Delta_x u + h(u)\partial_y u - \partial_t u, \tag{1.3}$$

$(x, y, t) = z$ denotes the point in $\mathbb{R}^N \times \mathbb{R} \times \mathbb{R}$, and Δ_x is the Laplace operator acting in the variable $x \in \mathbb{R}^N$. We assume that f, g, and h are globally Lipschitz continuous functions.

One of the main features of operator (1.3) is the strong degeneracy of its characteristic form due to the lack of diffusion in the y-direction, so that (1.1)–(1.2) may include the Cauchy problem for the Burgers equation, when $h(u) = u$, $g = g(y)$, and $f \equiv 0$. On the other hand, L can be considered as nonlinear version of the operator

$$K = \Delta_x + x_1\partial_y - \partial_t, \tag{1.4}$$

which was introduced by Kolmogorov [17] and has been extensively studied by Kuptsov [12] and Lanconelli and Polidoro [14]. Among the known results of K, we recall that every solution to $Ku = 0$ is smooth; thus we may expect some regularity properties also for the solutions to (1.1).

Problem (1.1)–(1.2) arises in mathematical finance as well as in the study of nonlinear physical phenomena such as the combined effects of diffusion and convection of matter.

*Received by the editors December 10, 2001; accepted for publication (in revised form) March 21, 2003; published electronically October 2, 2003. This research was supported by University of Bologna funds for selected research topics.

http://www.siam.org/journals/sima/35-3/39934.html

†Dipartimento di Matematica, Università di Bologna, Piazza di Porta S. Donato 5, 40127 Bologna, Italy (pascucci@dm.unibo.it, polidoro@dm.unibo.it).

Escobedo, Vazquez, and Zuazua [8] prove that there exists a unique distributional solution to (1.1)–(1.2) satisfying an entropy condition that generalizes the one by Kruzhkov [11]. This solution is characterized in the *vanishing viscosity* sense; i.e., it is the limit of a sequence of classical solutions to Cauchy problems related to the regularized operator

$$L_\varepsilon u = \Delta_x u + \varepsilon^2 \partial_y^2 u + h(u)\partial_y u - \partial_t u. \tag{1.5}$$

Vol'pert and Hudjaev [19] prove similar existence and uniqueness results in a space of bounded variation functions whose spatial derivatives are square integrable with respect to (w.r.t.) a suitable weight. In this framework, it is natural to consider bounded and integrable initial data g and nonlinearities of the form $h(u) = u^{p-1}$ for $p \in]1, \frac{N+2}{N+1}[$.

Our paper is mainly motivated by the theory of agents' decisions under risk, arising in mathematical finance. The problem is the representation of agents' preferences over consumption processes. Antonelli, Barucci, and Mancino [1] propose a utility functional that takes into account aspects of decision making such as the agents' habit formation, which is described as a smoothed average of past consumption and expected utility. In that model the processes utility and habit are described by a system of backward-forward stochastic differential equations. The solution of such a system, as a function of consumption and time, satisfies the Cauchy problem (1.1)–(1.2). Our regularity assumption on f, g, h is required by the financial model, since these functions appear in the backward-forward system as Lipschitz continuous coefficients.

In the paper by Antonelli and Pascucci [2] an existence result, in the case $N = 1$, is proved by probabilistic techniques that exploit the properties of the solutions to the system of backward-forward stochastic differential equations related to (1.1)–(1.2). In [2], the existence of a viscosity solution, in the sense of [7], is proved. The solution is defined in a suitably small strip $\mathbb{R}^2 \times [0, T]$ and satisfies the following conditions:

$$\begin{aligned} |u(x,y,t) - u(x',y',t)| &\le c_0(|x - x'| + |y - y'|), \\ |u(x,y,t) - u(x,y,t')| &\le c_0(1 + |(x,y)|)|t - t'|^{\frac{1}{2}} \end{aligned} \tag{1.6}$$

for every $(x, y), (x', y') \in \mathbb{R}^2, t, t' \in [0, T]$, where c_0 is a positive constant that depends on the Lipschitz constants of f, g, and h. Concerning the regularity of u, we remark that the results by Caffarelli and Cabré [3] and Wang [20, 21] do not apply to our operator.

In this paper we prove the existence of a classical solution u to problem (1.1)–(1.2) by combining the analysis on Lie groups with the standard techniques for the Cauchy problem related to degenerate parabolic equations. We say that u is a classical solution if $\partial_{x_j x_k} u$, $j, k = 1, \dots, N$, the directional derivative

$$z \longmapsto Yu(z) = \frac{\partial u}{\partial \nu_z}(z), \qquad \nu(z) = (0, h(u(z)), -1),$$

are continuous functions, and (1.1)–(1.2) are verified at every point. Our main result is the following.

THEOREM 1.1. *There exists a positive T and a unique function $u \in C(S_T)$, verifying estimates* (1.6) *on $\overline{S}_T$, which is a classical solution to* (1.1)–(1.2).

We stress that the regularity stated above is natural for the problem under consideration. Indeed, although Yu is the sum of the more simple terms $h(u)\partial_y u$ and

$\partial_t u$, it is not true in general that they are continuous functions. Further regularity properties of solutions can be obtained under additional conditions. For instance, in [5, 6] in collaboration with Citti, we considered the nonlinear equation in three variables,

$$\partial_{xx} u + u\partial_y u - \partial_t u = 0, \tag{1.7}$$

which is a special case of (1.1). Assuming a hypothesis formally analogous to the classical Hörmander condition, we proved that the viscosity solution u of (1.7) constructed in [2] actually is a C^∞ classical solution.

In this paper we give a direct proof of the existence of a classical solution to the Cauchy problem (1.1)–(1.2) by using analytical methods. The regularity part in Theorem 1.1 is based on a modification of the classical freezing method, introduced by Citti in [4] for the study of the Levi equation. More precisely, for any $\bar{z} \in S_T$, we approximate L by the linear operator

$$L_{\bar{z}} = \Delta_x + (h(u(\bar{z})) + x_1 - \bar{x}_1)\,\partial_y - \partial_t, \tag{1.8}$$

and we represent a solution u in terms of its fundamental solution. Note that up to a straightforward change of coordinates, $L_{\bar{z}}$ is the Kolmogorov operator (1.4), and hence an explicit expression of the fundamental solution of $L_{\bar{z}}$ is available. Also note that $L_{\bar{z}}$ is a good approximation of L in the sense that, by (1.6), we have

$$|Lu(z) - L_{\bar{z}} u(z)| = |u(z) - u(\bar{z}) - (x_1 - \bar{x}_1)|\,|\partial_y u(z)| \le c_0 d(\bar{z}, z),$$

where $d(\bar{z}, z)$ is the standard parabolic distance.

The existence part of Theorem 1.1 relies on the Bernstein technique. We explicitly note that the nonlinearity in (1.3) is not monotone; therefore a maximum principle for the operator $Lv + h'(u)v^2$, which occurs when we differentiate both sides of (1.1) w.r.t. y, does not hold unless we assume condition (1.6).

We end this introduction with a remark about the existence of global solutions. We first note that the space of functions characterized by conditions (1.6) is, in some sense, optimal for the existence of local classical solutions. Indeed the linear growth of the initial data g does not allow, in general, solutions that are defined at every time $t > 0$, as the following example given in [2] shows. Consider the problem (1.7), with $f \equiv 0$ and $g(x, y) = x + y$: a direct computation shows that $u(x, y, t) = \frac{x+y}{1-t}$ is the unique solution to the problem and blows up as $t \to 1$. This fact is expected since, if u grows as a linear function, then its Cole–Hopf transformed function grows as $\exp(y^2)$, which is the critical case for the parabolic Cauchy problem. Next we give a simple sufficient condition for the global existence of classical solutions.

THEOREM 1.2. *Let f, g, and h be globally Lipschitz continuous functions. Suppose that g is nonincreasing w.r.t. y, that f is nondecreasing w.r.t. y, and that there exists $c_0 \in\ ]0, c_1]$ such that*

$$c_0(u - v) \le h(u) - h(v) \tag{1.9}$$

for every $u, v \in \mathbb{R}$. Then the Cauchy problem (1.1)–(1.2) has a solution u that is defined in $\mathbb{R}^{N+1} \times \mathbb{R}^+$.

This paper is organized as follows. In section 2 we prove Theorem 1.1, and in section 3 we prove the existence of a viscosity solution of (1.1)–(1.2). Section 4 is devoted to the proof of Theorem 1.2.

2. Classical solutions. In this section we prove Theorem 1.1. We first state an existence and uniqueness result of a strong solution u to problem (1.1)–(1.2). And then we prove that u is a solution in the classical sense. We say that a continuous function u is a strong solution to (1.1)–(1.2) if $u \in H^1_{\rm loc}(S_T)$, $\partial_{x_j x_k} u \in L^2_{\rm loc}(S_T)$, $j,k = 1,\dots,N$, it satisfies equation (1.1) a.e., and it assumes the initial datum g.

THEOREM 2.1. *If T is suitably small, there exists a unique strong solution of* (1.1)–(1.2) *verifying estimates* (1.6) *on* $\overline{S}_T$.

The proof of Theorem 2.1 is postponed to section 3. We remark that in the above statement, we consider the term

$$Yu = h(u)\partial_y u - \partial_t u$$

as a sum of weak derivatives. Here we aim to prove that Yu is a continuous function and that it coincides with the directional derivative w.r.t. the vector $\nu_z = (0, h(u), -1)$, namely,

$$Y(u(z)) = \frac{\partial u}{\partial \nu_z}(z) \qquad \forall z \in S_T. \tag{2.1}$$

In what follows, when we consider a function F that depends on many variables, to avoid any ambiguity we shall systematically write the directional derivative introduced in (2.1) as

$$Y(z)F(\cdot,\zeta) = \frac{\partial F(\cdot,\zeta)}{\partial \nu_z}(z).$$

Our technique is inspired by the recent paper [6], where, in collaboration with Citti, we developed some ideas for a general study of a nonlinear equation of the form (1.4). We recall the following lemma, which has been proved in Lemma 3.1 of [6], for the Cauchy problem (1.7). We state the lemma for the operator (1.3) and omit the proof, since it is analogous to the one given in [6].

LEMMA 2.2. *Let v be a continuous function defined in S_T. Assume that its weak derivatives $v_y, v_t \in L^2_{loc}$ and that the limit*

$$\lim_{\delta \to 0} \frac{v(z + \delta \nu_z) - v(z)}{\delta}$$

exists and is uniform w.r.t. z in compact subsets of S_T. Then

$$\frac{\partial v}{\partial \nu_z}(z) = (h(u)\partial_y v - \partial_t v)(z) \qquad \text{a.e. } z \in S_T.$$

We next prove Theorem 1.1 by using a representation formula of the strong solution u in terms of the fundamental solution of the operator $L_{\bar z}$ introduced in (1.8). We define the first order operators (vector fields)

$$X_j = \partial_{x_j},\ j = 1,\dots,N, \qquad Y_{\bar z} = (h(u(\bar z)) + x_1 - \bar x_1)\,\partial_y - \partial_t. \tag{2.2}$$

Thus we can rewrite the operator $L_{\bar z}$ in the standard form

$$L_{\bar z} = \sum_{j=1}^{N} X_j^2 + Y_{\bar z}. \tag{2.3}$$

Let us recall some preliminary facts about real analysis on nilpotent Lie groups. More details about this topic can be found in [15] and [18]. We define on $\mathbb{R}^{N+2}$ the composition law

$$\theta \oplus \theta' = \left(\theta_1 + \theta'_1, \dots, \theta_N + \theta'_N, \theta_{N+1} + \theta'_{N+1}, \theta_{N+2} + \theta'_{N+2} + \frac{1}{2}(\theta_1\theta'_{N+1} - \theta_{N+1}\theta'_1)\right)$$

and the dilations group

$$\delta_\lambda(\theta) = (\lambda\theta_1, \dots, \lambda\theta_N, \lambda^2\theta_{N+1}, \lambda^3\theta_{N+2}), \qquad \lambda > 0.$$

We remark that $G = (\mathbb{R}^{N+2}, \oplus)$ is a nilpotent stratified Lie group which, in the case $N = 1$, coincides with the standard Heisenberg group. Since the Jacobian $J\delta_\lambda$ equals λ^{N+5}, the homogeneous dimension of G w.r.t. $(\delta_\lambda)_{\lambda>0}$ is the natural number $Q = N + 5$. A norm which is homogeneous w.r.t. this dilations group is given by

$$\|\theta\| = \left(|\theta_1|^6 + \dots + |\theta_N|^6 + |\theta_{N+1}|^3 + |\theta_{N+2}|^2\right)^{\frac{1}{6}}.$$

Let $\nabla_{\bar z} = (X_1, \dots, X_N, Y_{\bar z}, \partial_y)$ be the gradient naturally associated to $L_{\bar z}$ and consider any $z \in \mathbb{R}^{N+2}$. The exponential map

$$E_{\bar z}(\theta, z) = \exp(\langle \theta, \nabla_{\bar z}\rangle)(z)$$

is a global diffeomorphism and induces a Lie group structure on $\mathbb{R}^{N+2}$ whose product is defined by

$$\zeta \circ z = E_{\bar z}\left(\left(E^{-1}_{\bar z}(\zeta, 0) \oplus E^{-1}_{\bar z}(z, 0)\right), 0\right),$$

and it can be explicitly computed as

$$\zeta \circ z = (x + \xi, y + \eta - t\xi_1, t + \tau).$$

Moreover, a control distance $d_{\bar z}$ in $(\mathbb{R}^{N+2}, \circ)$ is defined by

$$\begin{aligned} d_{\bar z}(z, \zeta) &= \|E^{-1}_{\bar z}\left(\zeta^{-1} \circ z, 0\right)\| \\ &= \left(|x - \xi|^6 + |t - \tau|^3 + \left|y - \eta + (t - \tau)\left(h(u(\bar z)) + \frac{x_1 - \xi_1 - 2\bar x_1}{2}\right)\right|^2\right)^{\frac{1}{6}}, \end{aligned} \tag{2.4}$$

where ζ^{-1} is the inverse in the group law "$\circ$". We denote by $\Gamma_{\bar z}(z, \zeta)$ the fundamental solution of $L_{\bar z}$ with pole in ζ and evaluated at z. We refer to [12, 14, 13, 9] for known results about $\Gamma_{\bar z}$. The following bound holds:

$$\Gamma_{\bar z}(z, \zeta) = \Gamma_{\bar z}(\zeta^{-1} \circ z, 0) \le c d_{\bar z}(z, \zeta)^{-Q+2}, \tag{2.5}$$

where the constant c continuously depends on $\bar z$. We are now in a position to prove Theorem 1.1.

Proof of Theorem 1.1. By Theorem 2.1 there exists a unique strong solution of (1.1)–(1.2) verifying (1.6) in $\overline{S}_T$ for T suitably small. In order to prove that u is a classical solution, we represent it in terms of the fundamental solution $\Gamma_{\bar z}$:

$$(u\varphi)(z) = \int_{S_T} \Gamma_{\bar z}(z, \zeta)\left(U_{1,\bar z}(\zeta) - U_{2,\bar z}(\zeta)\right) d\zeta \equiv I_1(z) - I_2(z) \tag{2.6}$$

for every $\varphi \in C_0^\infty(S_T)$, where

$$U_{1,\bar z} = \varphi f(\cdot,u) + uL_{\bar z}\varphi + 2\langle \nabla_x u, \nabla_x \varphi\rangle,$$
$$U_{2,\bar z} = (h(u) - h(u(\bar z)) - (x_1 - \bar x_1))\,\partial_y u\varphi$$

are bounded functions with compact support. Therefore it is straightforward to prove that $u\varphi \in C^{1+\alpha}_{d_{\bar z}}$, $\alpha \in]0,1[$, where $C^{k+\alpha}_{d_{\bar z}}$ denotes the space of Hölder continuous functions w.r.t. the control distance $d_{\bar z}$. In particular, by choosing $\varphi \equiv 1$ in a compact neighborhood K of $\bar z$, we have that

$$X_j u(z) = \int X_j(z)\Gamma_{\bar z}(\cdot,\zeta)\,(U_{1,\bar z}(\zeta) - U_{2,\bar z}(\zeta))\,d\zeta, \qquad z \in K,\ j = 1,\dots,N,$$

and

$$|X_j u(z) - X_j u(\zeta)| \le c d_{\bar z}(z,\zeta)^\alpha \qquad \forall z,\zeta \in K,\ \alpha \in]0,1[. \tag{2.7}$$

This proves the Hölder continuity of the first order derivatives of u. Let us now consider the second order derivatives $X_j X_h u$, $j,k,= 1,\dots,N$, and Yu.

We next prove the existence of the directional derivative $Yu(\bar z)$ by studying separately the terms I_1, I_2. Since Y is the unique nonlinear vector field to be considered, the proof of our result for the derivatives $X_j X_h u$ is simpler and will be omitted.

The term I_2. We set

$$J(\bar z) = \int_{S_T} Y(\bar z)\Gamma_{\bar z}(\cdot,\zeta)U_{2,\bar z}(\zeta)d\zeta.$$

We remark that J is well defined and continuous since, by (1.6), we have

$$|U_{2,\bar z}(\zeta)| \le c\ d_{\bar z}(\bar z,\zeta). \tag{2.8}$$

We denote by $\chi \in C^\infty([0,+\infty[,[0,1])$ a cut-off function such that

$$\chi(s) = 0 \ \text{ for } \ 0 \le s \le \frac{1}{2}, \qquad \chi(s) = 1 \ \text{ for } \ s \ge 1,$$

and we set

$$I_{2,\delta}(z) = \int_{S_T} \Gamma_{\bar z}(z,\zeta)\chi\left(\frac{d_{\bar z}(\bar z,\zeta)}{\bar c\ \delta^{\frac12}}\right) U_{2,\bar z}(\zeta)d\zeta, \qquad \bar c,\delta > 0.$$

In what follows we shall assume $d_{\bar z}(\bar z, z) \le \delta^{\frac12}$; then by the triangular inequality

$$d_{\bar z}(\bar z,\zeta) \le c\ \left(d_{\bar z}(\bar z, z) + d_{\bar z}(z,\zeta)\right), \tag{2.9}$$

we can choose $\bar c$ suitably large so that

$$\chi\left(\frac{d_{\bar z}(\bar z,\zeta)}{\bar c\ \delta^{\frac12}}\right) = 0 \qquad \text{if } \ d_{\bar z}(z,\zeta) < \delta^{\frac12},$$

and, as a consequence, $I_{2,\delta}$ is smooth for any $\delta > 0$. We claim that

$$\sup_{d_{\bar z}(\bar z,z) \le \delta^{\frac12}} |I_{2,\delta}(z) - I_2(z)| \le c\ \delta^{\frac32}, \tag{2.10}$$

$$\sup_{d_{\bar z}(\bar z,z) \le \delta^{\frac12}} |Y_{\bar z} I_{2,\delta}(z) - J(\bar z)| \le c\ \delta^{\frac12}|\log(\delta)| \tag{2.11}$$

for some positive constant c. We postpone the proof of (2.10)–(2.11) to the end.

Let us now compute the derivative $\frac{\partial I_2}{\partial \nu_{\bar z}}(\bar z)$. For every positive δ we have

$$\left|\frac{I_2(\bar z+\delta\nu_{\bar z})-I_2(\bar z)}{\delta}-J(\bar z)\right| \le \left|\frac{I_{2,\delta}(\bar z+\delta\nu_{\bar z})-I_{2,\delta}(\bar z)}{\delta}-J(\bar z)\right| + \left|\frac{I_2(\bar z+\delta\nu_{\bar z})-I_{2,\delta}(\bar z+\delta\nu_{\bar z})}{\delta}\right| + \left|\frac{I_2(\bar z)-I_{2,\delta}(\bar z)}{\delta}\right|.$$

We first note that, using the expression (2.4), we find $d_{\bar z}(\bar z,\bar z+\delta\nu_{\bar z})=\delta^{\frac12}$. Thus, by (2.10) and by the mean value theorem, there exists a $\delta_0\in\,]0,\delta[$ such that

$$\begin{aligned}\left|\frac{I_2(\bar z+\delta\nu_{\bar z})-I_2(\bar z)}{\delta}-J(\bar z)\right| &\le \left|\left(h(u(\bar z))\partial_y I_{2,\delta}-\partial_t I_{2,\delta}\right)(\bar z+\delta_0\nu_{\bar z})-J(\bar z)\right|+c\delta^{\frac12}\\ &= |Y_{\bar z}I_{2,\delta}(\bar z+\delta_0\nu_{\bar z})-J(\bar z)|+c\delta^{\frac12}\le c\,\delta^{\frac12}|\log\delta|,\end{aligned}$$

where the last inequality follows from (2.11). Therefore we have

$$\frac{\partial I_2}{\partial \nu_z}(z)=J(z),$$

and, by Lemma 2.2, we get (2.1).

We are left with the proof of (2.10)–(2.11). We assume $d_{\bar z}(\bar z,z)\le\delta^{\frac12}$. By (2.8) and (2.5), we have

$$|I_{2,\delta}(z)-I_2(z)|\le c\int_{d_{\bar z}(z,\zeta)<\delta^{\frac12}} d_{\bar z}(z,\zeta)^{-Q+2}d_{\bar z}(\bar z,\zeta)d\zeta$$

(since, by (2.9), $d_{\bar z}(\bar z,\zeta)<c\delta^{\frac12}$, and by using the homogeneous polar coordinates)

$$\le\ \delta^{\frac12}\int_{\varrho<\delta^{\frac12}}\varrho^{-Q+2+Q-1}d\varrho=c\delta^{\frac32}.$$

This proves (2.10). Next we recall the following estimate which immediately follows by the mean value theorem:

$$|Y_{\bar z}(z)\Gamma_{\bar z}(\cdot,\zeta)-Y_{\bar z}(\bar z)\Gamma_{\bar z}(\cdot,\zeta)|\le c d_{\bar z}(\bar z,z)d_{\bar z}(\bar z,\zeta)^{-Q-1} \tag{2.12}$$

for $d_{\bar z}(\bar z,\zeta)\ge\bar c\,d_{\bar z}(\bar z,z)$. Then we have

$$\begin{aligned}|Y_{\bar z}I_{2,\delta}(z)-J(\bar z)| &\le \int|Y_{\bar z}(z)\Gamma_{\bar z}(\cdot,\zeta)-Y_{\bar z}(\bar z)\Gamma_{\bar z}(\cdot,\zeta)|\,\chi\left(\frac{d_{\bar z}(\bar z,\zeta)}{\bar c\ \delta^{\frac12}}\right)|U_{2,\bar z}(\zeta)|d\zeta\\ &\quad+\int_{d_{\bar z}(\bar z,\zeta)<\delta^{\frac12}}|Y_{\bar z}(\bar z)\Gamma_{\bar z}(\cdot,\zeta)U_{2,\bar z}(\zeta)|d\zeta\end{aligned}$$

(by (2.12) and since the second term can be estimated as before)

$$\le c\delta^{\frac12}\int_{d_{\bar z}(\bar z,\zeta)>\bar c\delta^{\frac12}} d_{\bar z}(\bar z,\zeta)^{-Q-1}|U_{2,\bar z}(\zeta)|d\zeta+c\delta^{\frac12}=c\delta^{\frac12}|\log(\delta)|.$$

This concludes the proof of (2.11).

The term I_1. Let $G(z,\zeta) = g(\zeta^{-1} \circ z)$, where g is a smooth function. A direct computation gives

$$Y_{\bar z}(z)G(\cdot,\zeta) = R_{\bar z}(\zeta)G(z,\cdot), \tag{2.13}$$

where

$$R_{\bar z}(\zeta) = -Y_{\bar z}(\zeta) - (x_1 - \xi_1)\partial_\eta$$

(see [16, p. 295] for a general statement of this result). We aim to prove that

$$Y(\bar z)I_1 = \int_\Omega Y(\bar z)\Gamma_{\bar z}(\cdot,\zeta)\left(U_{1,\bar z}(\zeta) - U_{1,\bar z}(\bar z)\right)d\zeta + U_{1,\bar z}(\bar z)\int_{\partial\Omega}\Gamma_{\bar z}(\bar z,\zeta)\langle R_{\bar z}(\zeta),\nu(\zeta)\rangle d\sigma, \tag{2.14}$$

where ν is the outer normal to the set $\Omega = \mathrm{supp}(\varphi)$, for which we assume that the divergence theorem holds. By (2.5), the homogeneity of the fundamental solution, and the Hölder continuity of $U_{1,\bar z}$, the function

$$V(z) = \int_\Omega Y(z)\Gamma_{\bar z}(\cdot,\zeta)\left(U_{1,\bar z}(\zeta) - U_{1,z}(z)\right)d\zeta + U_{1,\bar z}(z)\int_{\partial\Omega}\Gamma_{\bar z}(z,\zeta)\langle R_{\bar z}(\zeta),\nu(\zeta)\rangle d\sigma(\zeta) \tag{2.15}$$

is well defined. Let K be a compact subset of Ω. We set, for $\delta > 0$,

$$I_{1,\delta}(z) = \int_\Omega \Gamma_{\bar z}(z,\zeta)\chi\left(\frac{d_{\bar z}(z,\zeta)}{\delta}\right)U_{1,\bar z}(\zeta)d\zeta,$$

where χ is the cut-off function previously introduced. We choose δ suitably small so that

$$\chi\left(\frac{d_{\bar z}(z,\zeta)}{\delta}\right) = 1 \tag{2.16}$$

for any $z \in K, \zeta \in \partial\Omega$. Clearly $I_{1,\delta}$ is a smooth function, and differentiating we get

$$\begin{aligned} Y_{\bar z}(z)I_{1,\delta} &= \int_\Omega Y_{\bar z}(z)\left(\Gamma_{\bar z}(\cdot,\zeta)\chi\left(\frac{d_{\bar z}(\cdot,\zeta)}{\delta}\right)\right)\left(U_{1,\bar z}(\zeta) - U_{1,\bar z}(z)\right)d\zeta \\ &\quad + U_{1,\bar z}(z)\int_\Omega Y_{\bar z}(z)\left(\Gamma_{\bar z}(\cdot,\zeta)\chi\left(\frac{d_{\bar z}(\cdot,\zeta)}{\delta}\right)\right)d\zeta. \end{aligned} \tag{2.17}$$

By (2.13) and the divergence theorem, we have

$$\begin{aligned} \int_\Omega Y_{\bar z}(z)\left(\Gamma_{\bar z}(\cdot,\zeta)\chi\left(\frac{d_{\bar z}(\cdot,\zeta)}{\delta}\right)\right)d\zeta &= \int_\Omega R_{\bar z}(\zeta)\left(\Gamma_{\bar z}(z,\cdot)\chi\left(\frac{d_{\bar z}(z,\cdot)}{\delta}\right)\right)d\zeta \\ &= \int_{\partial\Omega}\Gamma_{\bar z}(z,\zeta)\chi\left(\frac{d_{\bar z}(z,\zeta)}{\delta}\right)\langle R_{\bar z}(\zeta),\nu(\zeta)\rangle d\sigma(\zeta). \end{aligned} \tag{2.18}$$

Then, by (2.18) and (2.16), the last terms in (2.17) and (2.15) are equal. Hence we get

$$\begin{aligned} &|V(z) - Y_{\bar z}(z)I_{1,\delta}| \\ &= \left|\int_{\delta_{\bar z}(z,\zeta)\le\delta} Y_{\bar z}(z)\left(\Gamma_{\bar z}(\cdot,\zeta)\left(1 - \chi\left(\frac{d_{\bar z}(\cdot,\zeta)}{\delta}\right)\right)\right)\left(U_{1,\bar z}(\zeta) - U_{1,\bar z}(z)\right)d\zeta\right| \\ &\le C\int_{\delta_{\bar z}(z,\zeta)\le\delta}\left(d_{\bar z}(z,\zeta)^{-Q} + \Gamma_{\bar z}(z,\zeta)\frac{d_{\bar z}(z,\zeta)^{-1}}{\delta}\right)d_{\bar z}(z,\zeta)^\alpha d\zeta \le C\delta^\alpha. \end{aligned}$$

Since the constant C continuously depends on $\bar{z}$, we have that $Y_{\bar{z}}(z)I_{1,\delta}$ converges to V as $\delta \to 0$ uniformly on K. Since $I_{1,\delta}$ converges to I_1 we get (2.14). This completes the proof of Theorem 1.1. □

3. A priori estimates. In this section we prove Theorem 2.1 by using a modification of the classical Bernstein method. Here we adopt the notation of [10, Chap. 3], which we briefly recall for the reader's convenience. Given a bounded domain Ω in $\mathbb{R}^{N+2}$ and $\alpha \in \ ]0,1[$, $\overline{C}_\alpha(\Omega)$ denotes the space of Hölder continuous functions w.r.t. the parabolic distance

$$d(z,z') \equiv |x-x'| + |y-y'| + |t-t'|^{\frac{1}{2}},$$

i.e., the family of all functions u on Ω for which

$$\overline{|u|}_\alpha^\Omega = \overline{|u|}_\alpha = |u|_0 + \sup_\Omega \frac{|u(z)-u(z')|}{d(z,z')^\alpha} < \infty,$$

where $|u|_0^\Omega = |u|_0 = \sup_\Omega |u|$. The spaces of Hölder continuous functions $\overline{C}_{k+\alpha}$, $k \in \mathbb{N}$, are defined straightforwardly. We set

$$B_r = \{(x,y) \in \mathbb{R}^N \times \mathbb{R} \ \mid \ |(x,y)| < r\}, \qquad S_{r,T} = B_r \times \ ]0,T[, \quad T,r>0. \tag{3.1}$$

The "parabolic" boundary of the cylinder $S_{r,T}$ is defined by

$$\partial_p S_{r,T} = (B_r \times \{0\}) \cup (\partial B_r \times [0,T]). \tag{3.2}$$

Given two points $z, z' \in S_{r,T}$ in (3.1), we denote by d_z the distance from z to the parabolic boundary $\partial_p S_{r,T}$ (cf. (3.2)), and $d_{zz'} = \min\{d_z, d_{z'}\}$. We set

$$|u|_\alpha^{S_{r,T}} = |u|_\alpha = |u|_0 + \sup_{S_{r,T}} d_{zz'}^\alpha \frac{|u(z)-u(z')|}{d(z,z')^\alpha}.$$

The space of all functions u with finite norm $|u|_\alpha^{S_{r,T}}$ is denoted by $C_\alpha(S_{r,T})$. The spaces $C_{k+\alpha}$ of Hölder continuous functions of higher order are defined analogously. We say that $u \in C_{k+\alpha,\mathrm{loc}}(S_T)$ if $u \in C_{k+\alpha}(S_{r,T})$ for every $r>0$.

We consider the Cauchy problem

$$L^\varepsilon u = f(\cdot,u) \qquad \text{in } \ S_T \equiv \mathbb{R}^{N+1} \times \ ]0,T[, \tag{3.3}$$

$$u(\cdot,0) = g \qquad \text{in } \ \mathbb{R}^{N+1}, \tag{3.4}$$

where L^ε, $\varepsilon > 0$, is the regularized operator in (1.5). We assume that the functions f, g, h are globally Lipschitz continuous; then there exists a positive constant c_1 such that

$$\begin{aligned} &c_1 \geq \max\{\text{Lipschitz constants of } f,g,h\},\\ &|h(v)| \leq c_1\sqrt{1+v^2}, \qquad |g(x,y)| \leq c_1\sqrt{1+|(x,y)|^2},\\ &|f(x,y,t,v)| \leq c_1\sqrt{1+|(x,y,t,v)|^2}, \qquad (x,y,t,v) \in S_T \times \mathbb{R}. \end{aligned} \tag{3.5}$$

The following result holds.

THEOREM 3.1. *There exist two positive constants T, c that depend only on the constant c_1 in (3.5) such that for every $\varepsilon > 0$ and $\alpha \in \ ]0,1[$ the Cauchy problem*

(3.3)–(3.4) *has a unique solution* $u^\varepsilon \in C_{2+\alpha,\mathrm{loc}}(S_T) \cap C\left(\overline{S}_T\right)$ *verifying the following* ε*-uniform estimates:*

$$|u^\varepsilon_{x_i}|_0, |u^\varepsilon_y|_0 \le 4c_1, \qquad i = 1, \dots, N, \tag{3.6}$$

$$|u^\varepsilon(x,y,t+s) - u^\varepsilon(x,y,t)| \le c\sqrt{1+|(x,y)|^2}\ |s|^{\frac{1}{2}}, \tag{3.7}$$

$$|u^\varepsilon(x,y,t)| \le 2c_1\sqrt{1+|(x,y,t)|^2} \qquad \forall (x,y,t) \in \overline{S}_T. \tag{3.8}$$

Before proving Theorem 3.1, we introduce some further notation. If $\chi = \chi(x,y) \in C_0^\infty\left(\mathbb{R}^{N+1}\right)$ is a cut-off function such that $\chi = 1$ in $B_{\frac{1}{2}}$ and $\mathrm{supp}(\chi) \subset B_1$, we set

$$\chi_n(x,y) = \chi\left(\frac{x}{n}, \frac{y}{n}\right), \quad f_n = f\chi_n, \quad g_n(\cdot,t) = g\chi_n, \quad h_n(\cdot,v) = h(v)\chi_n, \quad n \in \mathbb{N}, \tag{3.9}$$

so that, by (3.5) and readjusting the constant c_1 if necessary, we have

$$|\nabla\chi_n|_0 \le \frac{|\nabla\chi|_0}{n}, \qquad |\nabla g_n| \le c_1,$$
$$|\nabla_{x,y} f_n(x,y,t,v)| \le |\chi_n \nabla_{x,y} f| + \frac{c_1\,|\nabla\chi|_0}{n}\sqrt{1+n^2+T^2+v^2} \le c_1$$

if $\frac{|v|}{n}$ is bounded and $t \in [0,T]$.

Finally, fixing $n \in \mathbb{N}$ and $\varepsilon > 0$, we consider the linearized Cauchy–Dirichlet problem

$$L^{\varepsilon,n}_v u \equiv \Delta_x u + \varepsilon^2 u_{yy} + h_n(\cdot,v)\partial_y u - \partial_t u = f_n(\cdot,v) \qquad \text{in } S_{n,T}, \tag{3.10}$$

$$u = g_n \qquad \text{in } \partial_p S_{n,T}. \tag{3.11}$$

Given $\alpha \in]0,1[$, we assume that the coefficient v in (3.10)–(3.11) belongs to the space $\overline{C}_{1+\alpha}(S_{n,T})$ and satisfies the estimates

$$|v(x,y,t)| \le 2c_1\sqrt{1+|(x,y)|^2} \qquad \text{in } S_{n,T}, \tag{3.12}$$

$$|v_{x_i}|_0 \le 4c_1, \qquad i = 1, \dots, N, \tag{3.13}$$

$$|v_y|_0 \le 4c_1. \tag{3.14}$$

Then a classical solution $u \in \overline{C}_{2+\alpha}(S_{n,T})$ to (3.10)–(3.11) exists by known results (see, for example, [10, Chap. 3, Thm. 7], since $h_n(\cdot,v), f_n(\cdot,v) \in \overline{C}_{1+\alpha}(S_{n,T})$, $g_n \in C^\infty\left(\overline{S}_{n,T}\right)$, and the compatibility condition $L^{\varepsilon,n}_v g_n = f_n = 0$ holds on ∂B_n. Once we have given the following ε-uniform a priori estimates, the proof of Theorem 3.1 is rather standard.

LEMMA 3.2. *Under the above assumptions, there exists* $T > 0$ *such that, for any* $n \in \mathbb{N}$*, every classical solution of* (3.10)–(3.11) *verifies* (3.12)–(3.14).

Proof. Let u be a classical solution of (3.10)–(3.11). We prove estimate (3.12) for u by applying the maximum principle to the functions $H \pm u$, where H is defined as

$$H(x,y,t) = (c_1 + \mu t)\sqrt{1+|(x,y)|^2}$$

and μ is to be suitably fixed. Keeping in mind (3.5) and (3.12), it is easily verified that

$$\begin{aligned} L^{\varepsilon,n}_v H(x,y,t) &\le \frac{(1+\varepsilon^2)(c_1+\mu T)}{\sqrt{1+|(x,y)|^2}} + ((c_1+\mu T)\,c_1 - \mu)\sqrt{1+|(x,y)|^2} \\ &\le -|f_n(x,y,t,v(x,y,t))| \end{aligned}$$

if $\mu, \frac{1}{T}$ are suitably large. On the other hand, by (3.5), $H|_{\partial_p S_{n,T}} \geq |g_n|$. Therefore, by the maximum principle, we infer that

$$|u| \leq H \leq 2c_1\sqrt{1+|(x,y)|^2} \qquad \text{if } T \leq \frac{c_1}{\mu}.$$

Next we prove estimate (3.14) for the y-derivative of u. Our method is based on the maximum principle. We start by proving a gradient estimate for u on the parabolic boundary of $S_{n,T}$. Since $u \in \overline{C}_{2+\alpha}(S_{n,T})$, it is clear that $\nabla_{x,y}u = \nabla_{x,y}g_n$ in $B_n \times \{0\}$. In order to estimate $\nabla_{x,y}u$ on $\partial B_n \times]0,T[$, we employ the classical argument of the barrier functions on the cylinder $Q \equiv S_{n,T} \setminus S_{\frac{n}{2},T}$. More precisely, given $(x_0,y_0,t_0) \in \partial B_n \times]0,T[$, we set

$$w(x,y) = 4c_1\langle (x-x_0, y-y_0), \nu\rangle,$$

where ν is the inner normal to Q at (x_0,y_0,t_0). Then we have

$$L_v^{\varepsilon,n}(w \pm u) = \pm f_n(\cdot, v) = 0 \qquad \text{in } Q,$$

since f_n and h_n vanish on Q. On the other hand, it is straightforward to verify that $|u| \leq w$ on $\partial_p Q$. Therefore, by the maximum principle, we get $|u| \leq w$ and, in particular,

$$|\nabla_{x,y}u(x_0,y_0,t_0)| \leq |\nabla_{x,y}w(x_0,y_0)| \leq 4c_1. \tag{3.15}$$

Now we are in a position to prove estimate (3.14) for u. We differentiate equation (3.10) w.r.t. the variable y and then multiply it by $e^{-2\lambda t}u_y$. Denoting $\omega = \left(e^{-\lambda t}u_y\right)^2$, we obtain

$$\begin{aligned} L_v^{\varepsilon}\omega &= e^{-2\lambda t}L_v^{\varepsilon}u_y^2 + 2\lambda\omega \\ &= 2\left(e^{-2\lambda t}\left(|\nabla_x u_y|^2 + \varepsilon^2 u_{yy}^2 + u_y\left((f_n)_y + (f_n)_v v_y\right)\right) + \omega\left(\lambda - h'(v)v_y\right)\right) \\ &\geq 2\left(e^{-2\lambda t}u_y\left((f_n)_y + (f_n)_v v_y\right) + \omega\left(\lambda - h'(v)v_y\right)\right). \end{aligned} \tag{3.16}$$

Hence, by setting $w = \omega - (4c_1)^2$, we get from (3.16)

$$L_v^{\varepsilon}w \geq 2\sqrt{\omega}\left(-|(f_n)_y| - |v_y(f_n)_v| + \sqrt{\omega}\left(\lambda - |h'v_y|\right)\right)$$

(by (3.5), (3.14), and by the elementary inequality $\sqrt{\omega} \geq \frac{\sqrt{2}}{2}(4c_1 + \operatorname{sgn}(w)\sqrt{|w|})$)

$$\geq \sqrt{2\omega}\left(\sqrt{2}c_1\left(2\sqrt{2}\left(\lambda - 4c_1^2\right) - 4c_1 - 1\right) + \left(\lambda - 4c_1^2\right)\operatorname{sgn}(w)\sqrt{|w|}\right)$$

(for $\lambda = \lambda(c_1)$ suitably large)

$$\geq c\sqrt{\omega|w|}\operatorname{sgn}(w) \tag{3.17}$$

for some positive constant $c = c(c_1)$. By contradiction, we want to prove that $w \leq 0$ in $S_{n,T}$. It will follow that

$$|u_y| \leq c_1 e^{\lambda t},$$

which implies (3.16) if $T = T(c_1) > 0$ is sufficiently small. Let z_0 be the maximum of w on $\overline{Q}_T$. If $w(z_0) > 0$, then $z_0 \in S_{n,T} \setminus \partial_p S_{n,T}$, since by (3.15) $w \leq 0$ on $\partial_p S_{n,T}$. This leads to a contradiction, since by (3.17)

$$0 \geq L_v^{\varepsilon}w(z_0) \geq c\sqrt{\omega(z_0)w(z_0)} > 0.$$

This concludes the proof of (3.14). By a similar technique, we prove estimate (3.13) of the x-derivatives of u:

$$|u_{x_k}|_0 \leq 4c_1, \qquad k = 1, \dots, N.$$

We set

$$\omega = \left(e^{-\lambda t} u_{x_k}\right)^2, \qquad w = \omega - (4c_1)^2.$$

Differentiating (3.10) w.r.t. x_k and multiplying it by $e^{-2\lambda t} u_{x_k}$, we get

$$\begin{aligned} L_v^\varepsilon w &= e^{-2\lambda t} L_v^\varepsilon u_{x_k}^2 + 2\lambda\omega \\ &= 2\left(e^{-2\lambda t} u_{x_k} \left((f_n)_{x_k} + v_{x_k}\left((f_n)_v - u_y h'\right)\right) + \lambda\omega\right) \end{aligned}$$

(by (3.5), (3.13), and estimate (3.14) of u_y previously proved)

$$\geq c\sqrt{\omega|w|}\operatorname{sgn}(w),$$

if $\lambda = \lambda(c_1)$ is suitably large, for some positive constant c which depends only on c_1. As before, we infer that $w \leq 0$, which yields (3.13). □

We are in a position to prove Theorem 3.1.

Proof of Theorem 3.1. In order to prove the existence of a unique classical solution to (3.3)–(3.4), we consider, for every $\varepsilon > 0$ and $n \in \mathbb{N}$, the Cauchy–Dirichlet problem

$$\Delta_x u + \varepsilon^2 u_{yy} + h_n(\cdot, u)\partial_y u - \partial_t u = f_n(\cdot, u) \qquad \text{in } S_{n,T}, \tag{3.18}$$

$$u = g_n \qquad \text{in } \partial_p S_{n,T}. \tag{3.19}$$

We split the proof into four steps: We first use Schauder's fixed point theorem to solve the above problem. Then we let n go to infinity under the assumption that the coefficients are smooth. Next we prove estimates (3.6), (3.7), and (3.8). Finally we remove the smoothness assumption.

First step. Assume that f, g, h are C^∞ functions. We fix $\alpha \in]0,1[$, $n \in \mathbb{N}$ and denote by $\mathcal{W}$ the family of functions $v \in \overline{C}_{1+\alpha}(S_{n,T})$ such that

$$\overline{|v|}_{1+\alpha} \leq M, \tag{3.20}$$

$$|v(x,y,t)| \leq 2c_1\sqrt{1 + |(x,y)|^2} \qquad \text{in } S_{n,T}, \tag{3.21}$$

$$|v_{x_i}|_0 \leq 4c_1, \qquad i = 1, \dots, N, \tag{3.22}$$

$$|v_y|_0 \leq 4c_1, \tag{3.23}$$

where the positive constants M, T will be suitably chosen later. Clearly, $\mathcal{W}$ is a closed, convex subset of $\overline{C}_{1+\alpha}(S_{n,T})$. We define a transformation $u \equiv \mathcal{Z}v$ on $\mathcal{W}$ by choosing u as the unique classical solution of the linear Cauchy–Dirichlet problem (3.10)–(3.11). If we show that

(i) $\mathcal{Z}(\mathcal{W})$ is precompact in $\overline{C}_{1+\alpha}(S_{n,T})$;
(ii) $\mathcal{Z}$ is a continuous operator;
(iii) $\mathcal{Z}(\mathcal{W}) \subseteq \mathcal{W}$,

then we are done. The proof of (i) and (ii) is quite standard and relies on the following two estimates of u (see, for example, [10, Chap. 3, Thm. 6 and Chap. 7, Thm. 4]:

$$\overline{|u|}_{2+\alpha} \leq c\left(\overline{|g_n|}_{2+\alpha} + \overline{|f_n(\cdot, v)|}_\alpha\right) \leq \bar{c}\left(\overline{|g_n|}_{2+\alpha} + \overline{|v|}_\alpha\right) \tag{3.24}$$

for some constant $\bar{c} > 0$ dependent on $\varepsilon, n, M, \alpha$;

$$\overline{|u|}_{1+\delta} \le \widetilde{c}\left(|f_n|_0 + |L_v^\varepsilon g_n|_0 + \overline{|g_n|}_{1+\delta}\right), \qquad \delta \in]0,1[, \tag{3.25}$$

for some positive constant $\widetilde{c}$ dependent on ε, n, δ but not on M. Besides, (iii) is exactly the content of Lemma 3.2. Therefore, by Schauder's theorem, the operator $\mathcal{Z}$ has a fixed point u in $\mathcal{W}$.

Note that, by (3.6), a comparison principle in the space $\mathcal{W}$ does hold; therefore u is the unique classical solution of problem (3.18)–(3.19) verifying estimates (3.6) and (3.8). Moreover, by a standard bootstrap argument, $u \in C^\infty(S_{n,T})$.

Second step. We fix $\varepsilon > 0$ and denote by u^n the solution of the Cauchy–Dirichlet problem (3.18)–(3.19), whose existence has been proved in the previous step. We now want to obtain the solution of the Cauchy problem (3.3)–(3.4) letting n go to infinity.

Fixing $k \in \mathbb{N}$, we consider the sequence $(u^n \chi_{4k})_{n \ge 4k}$, where χ is the cut-off function introduced in (3.9). Then we have

$$\begin{aligned} L^\varepsilon_{u^n}(u^n\chi_{4k}) &= f_{4k}(\cdot, u_n) + 2\left(\langle \nabla_x u^n, \nabla_x \chi_{4k}\rangle + \varepsilon^2 \partial_y u^n \partial_y \chi_{4k}\right) + u^n L^\varepsilon_{u_n}\chi_{4k} \\ &\equiv F_{n,4k} \qquad \text{on } S_{4k,T}, \\ (u^n\chi_{4k})\,|_{\partial_p S_{4k,T}} &= g_{4k}. \end{aligned}$$

By classical Hölder estimates, we deduce

$$\overline{|u^n|}^{S_{2k,T}}_{\delta} \le \overline{|u^n\chi_{4k}|}^{S_{4k,T}}_{1+\delta} \le c\left(|F_{n,4k}|_0^{S_{4k,T}} + |L^\varepsilon_{u^n} g_{4k}|_0^{S_{4k,T}} + \overline{|g_{4k}|}^{S_{4k,T}}_{1+\delta}\right) \le \bar{c}$$

for every $n \ge 4k$ and $\delta \in]0,1[$, where $\bar{c} = \bar{c}(\delta, \varepsilon, c_1, k)$ does not depend on n. Moreover, since

$$\begin{aligned} L^\varepsilon_{u^n}(u^n\chi_{2k}) &= F_{n,2k} \qquad \text{on } S_{4k,T}, \\ (u^n\chi_{2k})\,|_{\partial_p S_{2k,T}} &= g_{2k}, \end{aligned}$$

we obtain

$$\overline{|u^n|}^{S_{k,T}}_{2+\delta} \le \overline{|u^n\chi_{2k}|}^{S_{2k,T}}_{2+\delta} \le c\left(\overline{|F_{n,2k}|}^{S_{2k,T}}_{\delta} + \overline{|g_{2k}|}^{S_{2k,T}}_{2+\delta}\right) \le \bar{\bar{c}} \qquad \forall n \ge 4k,$$

where $\bar{\bar{c}} = \bar{\bar{c}}(\delta, \varepsilon, c_1, k)$ does not depend on n.

Then, by the Ascoli–Arzelà theorem and Cantor's diagonal argument, we can extract from u^n a subsequence $\overline{|\ |}_{2+\alpha}$-convergent on compacts of S_T for every $\alpha \in]0,1[$ to the solution u^ε of (3.18)–(3.19) verifying estimates (3.6) and (3.8). The uniqueness of u^ε follows again from standard results.

Third step. We still assume $f, g, h \in C^\infty \cap \mathrm{Lip}$. We aim to prove estimate (3.7) for the solution u^ε found in the previous step. We fix $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}$ and set

$$w(x,y,t) = u^\varepsilon(x, \varepsilon y, t)\bar{\chi}(x, \varepsilon y), \qquad \varepsilon > 0,\ (x,y,t) \in S_T,$$

where $\bar{\chi}(x,y) = \chi(x - \bar{x}, y - \bar{y})$ and χ is the cut-off function in (3.9). We have

$$(\Delta_x + \partial_{yy} - \partial_t) w = \Psi^\varepsilon \qquad \text{on } S_T,$$

where

$$\begin{aligned} \Psi^\varepsilon(x,y,t) = \Big[&\bar{\chi}\left(f(\cdot, u^\varepsilon) - h(u^\varepsilon)u^\varepsilon_y\right) + u^\varepsilon\left(\Delta_x\bar{\chi} + \varepsilon^2\bar{\chi}_{yy}\right) \\ &+ 2\left(\langle \nabla_x u^\varepsilon, \nabla_x\bar{\chi}\rangle + \varepsilon^2 u^\varepsilon_y \bar{\chi}_y\right)\Big](x, \varepsilon y, t), \qquad (x,y,t) \in S_T. \end{aligned}$$

Denoting by $\Gamma_H(z;\zeta)$ the fundamental solution of the heat operator in $\mathbb{R}^{N+2}$ with pole at $\zeta=(\xi,\eta,\tau)$ and evaluated in $z=(x,y,t)$, we have the following representation of w:

$$\begin{aligned} w(z) &= \int_0^t \int_{\mathbb{R}^{N+1}} \Gamma_H(z;\zeta)\Psi^\varepsilon(\zeta)d(\xi,\eta)d\tau \\ &- \int_{\mathbb{R}^{N+1}} \Gamma_H(z;\xi,\eta,0)g\bar{\chi}(\xi,\varepsilon\eta)d(\xi,\eta) \equiv I_1(z)-I_2(z). \end{aligned} \tag{3.26}$$

In order to estimate I_1, it suffices to note that, by (3.5), (3.6), and (3.8), we have that

$$|\Psi^\varepsilon|_0 \le c\sqrt{1+|(\bar{x},\bar{y})|^2}, \tag{3.27}$$

with c dependent only on c_1. Hence, by an elementary argument, we get

$$|I_1(x,y,t+s)-I_1(x,y,t)| \le c\sqrt{1+|(\bar{x},\bar{y})|^2}\ |s|^{\frac{1}{2}} \qquad \forall (x,y,t)\in S_T,\ s\in[-t,T-t], \tag{3.28}$$

where c depends only on c_1.

To estimate I_2, we begin by noting that a simple change of variables gives

$$I_2(x,y,t)=\int_{\mathbb{R}^{N+1}} \Gamma_H(\xi,\eta,1;0)g\bar{\chi}\left(x-\xi\sqrt{t},\varepsilon(y-\eta\sqrt{t})\right)d\xi d\eta.$$

Then

$$\begin{aligned} &|I_2(x,y,t+s)-I_2(x,y,t)| \le \int_{\mathbb{R}^{N+1}} \Gamma_H(\xi,\eta,1;0) \\ &\cdot \left| g\bar{\chi}\left(x-\xi\sqrt{t+s},\varepsilon(y-\eta\sqrt{t+s})\right) - g\bar{\chi}\left(x-\xi\sqrt{t},\varepsilon(y-\eta\sqrt{t})\right)\right| d\xi d\eta \end{aligned}$$

(by the mean value theorem, for some constant $c=c(c_1)>0$)

$$\begin{aligned} &\le c\sqrt{1+|(\bar{x},\bar{y})|^2}\ \left|\sqrt{t+s}-\sqrt{t}\right| \int_{\mathbb{R}^{N+1}} \Gamma_H(\xi,\eta,1;0)\left(|\xi|+\varepsilon|\eta|\right)d\xi d\eta \\ &\le c\sqrt{1+|(\bar{x},\bar{y})|^2}\ \sqrt{2|s|} \qquad \forall (x,y,t)\in S_T,\ s\in[-t,T-t], \end{aligned} \tag{3.29}$$

where c depends only on c_1.

Then, by the definition of w and by (3.26), we obtain

$$u^\varepsilon(\bar{x},\bar{y},t)=I_1\left(\bar{x},\frac{\bar{y}}{\varepsilon},t\right)-I_2\left(\bar{x},\frac{\bar{y}}{\varepsilon},t\right),$$

and estimate (3.7) follows from (3.28), (3.29).

Fourth step. We finally consider the general case where f,g,h are only assumed to be globally Lipschitz continuous. We use the standard mollifiers to approximate f,g,h uniformly on compacts by some sequences (f_n), (g_n), (h_n) in $C^\infty\cap\mathrm{Lip}$ that verify the estimates (3.5). Since the interval $[0,T]$ of existence of the solution constructed in the second step does not depend on the regularity of the coefficients, we may employ the usual density argument to find a function u^ε which is the unique classical solution of (3.3)–(3.4). □

Proof of Theorem 2.1. By Theorem 3.1, there exists a sequence

$$u^{\varepsilon_n} \in C_{2+\alpha,\mathrm{loc}}\left(S_T\right) \cap C\left(\overline{S}_T\right),$$

with $\varepsilon_n \downarrow 0$, such that every function u^{ε_n} is a solution of (3.3)–(3.4) with $\varepsilon = \varepsilon_n$ and verifies (1.6) for a constant c_0 that does not depend on n, and (u^{ε_n}) converges uniformly on compact subsets of $\overline{S}_T$ to a function u.

Arguing as in [6, Lem. 2.4], we can prove the following a priori estimates of Caccioppoli type for the derivatives of the functions (u^{ε_n}): if $\varphi \in C_0^\infty\left(S_T\right)$, there exists a positive constant c which depends only on f, φ and on the constant c_0 in (1.6) such that

$$\sum_{j=1}^{N} \left(\|u^{\varepsilon_n}_{x_j x_j}\varphi\|_2 + \|u^{\varepsilon_n}_{x_j y}\varphi\|_2 \right) + \varepsilon_n \|u^{\varepsilon_n}_{yy}\varphi\|_2 + \|u^{\varepsilon_n}_{t}\varphi\|_2 \le c \tag{3.30}$$

for every n. Therefore, up to a subsequence, $\partial_{x_j,x_k} u^{\varepsilon_n}$, $\varepsilon_n^2 \partial_{yy} u^{\varepsilon_n}$, $\partial_y u^{\varepsilon_n}$, and $\partial_t u^{\varepsilon_n}$ weakly converge in $L^2_{\mathrm{loc}}(S_T)$ to $\partial_{x_j,x_k} u$, 0, $\partial_y u$, and $\partial_t u$, respectively. Hence $u \in H^1_{\mathrm{loc}}(S_T), \partial_{x_j x_k} u \in L^2_{\mathrm{loc}}(S_T)$ for $j,k = 1,\dots,N$, and (1.1) is satisfied a.e.

The uniqueness of the solution can be proved as in [2, Prop. 5.1]. Indeed, since (u^{ε_n}) converges uniformly on compact sets, it is standard to prove that the limit u is a viscosity solution of (1.1)–(1.2) satisfying (1.6). Then the uniqueness of u follows by the comparison principle for viscosity solutions. □

4. Global existence. The main purpose of this section is to prove Theorem 1.2 by a simple continuation argument which relies on a bound of the gradient of u.

Proof of Theorem 1.2. The local existence result stated in Theorem 3.1 and a standard argument ensure that there exist an interval $I = [0, \overline{T}[$, where $\overline{T} \in \mathbb{R}^+$ or $\overline{T} = +\infty$, and a solution $u \in C^2(\mathbb{R}^{N+1} \times I)$ to problem (1.1)–(1.2), which cannot be defined for $t \ge \overline{T}$. We claim that our assumptions on f, g, and h yield $\overline{T} = +\infty$. To this end, we consider the local solution $u \in C^2(\mathbb{R}^{N+1} \times [0,T])$, which has been constructed in Theorem 3.1, and we denote by c_T the spatial Lipschitz constant corresponding to the strip S_T:

$$c_T = \inf \Big\{ c > 0 : |u(x,y,t) - u(x',y',t)| \le c(|x-x'|^2 + |y-y'|^2)^{1/2} \\ \forall (x,y,t),(x',y',t) \in \mathbb{R}^{N+1} \times [0,T] \Big\}.$$

We explicitly note that if $\overline{T} \ne +\infty$, then $c_t \to +\infty$ as $t \to \overline{T}$; hence a bound of the form

$$c_t \le c e^{kt} \tag{4.1}$$

for some positive constants c, k will prove our claim.

In order to prove (4.1) we first observe that, as in the proof of Theorem 3.1, it is not restrictive to assume that f, g, and h are smooth and that u is the classical solution of the regularized equation (1.5). We next show that

$$0 \le -u_y \le \frac{c_1}{c_0} + 1, \tag{4.2}$$

$$|u_{x_j}| \le c_1 e^{kt} \quad \text{for} \quad j = 1,\dots,N \tag{4.3}$$

for every $(x,y,t) \in S_T$, where c_1 is the Lipschitz constant defined in (3.5) and $k > 0$ does not depend on ε. To prove the first inequality in (4.2) we set $w(x,y,t) = e^{-\lambda t} u_y(x,y,t)$ for some $\lambda > 0$, and we note that since u is smooth, w is a solution to

$$\begin{cases} L_\varepsilon w = e^{-\lambda t} f_y + (\lambda - h'(u) u_y + f_u) w & \text{in } S_T, \\ w(\,\cdot\,, 0) = g_y. \end{cases}$$

By our assumptions $f_y \geq 0, g_y \leq 0, c_0 \leq h' \leq c_1$ and also by Theorem 3.1, $h'(u)u_y + f_u$ is bounded in S_T. Then $\lambda - h'(u)u_y + f_u > 0$ for suitably large λ and, as a consequence, $w \leq 0$ by the maximum principle. This proves the first inequality in (4.2). To prove the second one we set $w(x,y,t) = \frac{1}{2}\big(u_y^2(x,y,t) - \lambda^2\big)$, $\lambda > 0$, and argue as in the proof of Theorem 3.1: w is a solution to

$$\begin{cases} L_\varepsilon w = |\nabla_x u_y|^2 + \varepsilon^2 u_{yy}^2 - u_y(h'(u)u_y^2 - u_y f_u - f_y) & \text{in } S_T, \\ w(\,\cdot\,, 0) = \frac{1}{2}\big(g_y^2 - \lambda^2\big). \end{cases}$$

Since $u_y \leq 0$, we may choose λ sufficiently large (for instance, $\lambda = \frac{c_1}{c_0} + 1$) so that the right-hand side of the above differential equation is positive when $w > 0$. Then the second inequality in (4.2) follows again from the maximum principle.

We finally consider the function $w(x,y,t) = e^{-2\lambda t}\frac{u_{x_j}^2}{2} - \frac{c_1^2}{2}$ for $j = 1, \ldots, N$. Clearly $w(x,y,0) \leq 0$ and

$$\begin{aligned} L_\varepsilon w &= e^{-2\lambda t}\left(u_{x_j}^2(\lambda - h'(u)u_y + f_u) + |\nabla_x u_{x_j}|^2 + \varepsilon^2 u_{x_j,y}^2 + u_{x_j} f_{x_j}\right) \\ &\geq e^{-2\lambda t}\left(u_{x_j}^2(\lambda + f_u) + u_{x_j} f_{x_j}\right) \end{aligned}$$

by (4.2). Since $w \geq 0$ implies $|u_{x_j} f_{x_j}| \leq u_{x_j}^2$ then, for suitably large λ, we find $L_\varepsilon w \leq 0$ for $w \geq 0$ and prove (4.3) as above. This gives (4.1) and concludes the proof of Theorem 1.2. □

Remark 4.1. Hypothesis (1.9) on h is related to the natural assumptions in the theory of conservation laws. A simple counterexample shows that we cannot drop this condition. Indeed if $h(u) = -u$, $f \equiv 0$ and $g(x,y) = x - y$, then $u(x,y,t) = \frac{x-y}{1-t}$.

REFERENCES

[1] F. Antonelli, E. Barucci, and M. E. Mancino, *A comparison result for FBSDE with applications to decisions theory*, Math. Methods Oper. Res., 54 (2001), pp. 407–423.

[2] F. Antonelli and A. Pascucci, *On the viscosity solutions of a stochastic differential utility problem*, J. Differential Equations, 186 (2002), pp. 69–87.

[3] L. A. Caffarelli and X. Cabré, *Fully Nonlinear Elliptic Equations*, Amer. Math. Soc. Colloq. Publ. 43, AMS, New York, 1995.

[4] G. Citti, *C^∞ regularity of solutions of a quasilinear equation related to the Levi operator*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 23 (1996), pp. 483–529.

[5] G. Citti, A. Pascucci, and S. Polidoro, *On the regularity of solutions to a nonlinear ultraparabolic equation arising in mathematical finance*, Differential Integral Equations, 14 (2001), pp. 701–738.

[6] G. Citti, A. Pascucci, and S. Polidoro, *Regularity properties of viscosity solutions of a non-Hörmander degenerate equation*, J. Math. Pures Appl., 80 (2001), pp. 901–918.

[7] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[8] M. Escobedo, J. L. Vazquez, and E. Zuazua, *Entropy solutions for diffusion-convection equations with partial diffusivity*, Trans. Amer. Math. Soc., 343 (1994), pp. 829–842.

[9] C. Fefferman and A. Sánchez-Calle, *Fundamental solutions for second order subelliptic operators*, Ann. of Math., 124 (1986), pp. 247–272.
[10] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
[11] S. N. Kruzhkov, *First order quasilinear equations in several independent variables*, Math. USSR-Sb., 10 (1970), pp. 217–243.
[12] L. P. Kuptsov, *Fundamental solutions of certain degenerate second-order parabolic equations*, Math. Notes, 31 (1982), pp. 283–289.
[13] S. Kusuoka and D. Stroock, *Applications of the Malliavin calculus.* III, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 34 (1987), pp. 391–442.
[14] E. Lanconelli and S. Polidoro, *On a class of hypoelliptic evolution operators*, Rend. Sem. Mat. Univ. Politec. Torino, 52 (1994), pp. 29–63.
[15] A. Nagel, E. M. Stein, and S. Wainger, *Balls and metrics defined by vector fields* I: *Basic properties*, Acta Math., 155 (1985), pp. 103–147.
[16] L. P. Rothschild and E. M. Stein, *Hypoelliptic differential operators on nilpotent groups*, Acta Math., 137 (1977), pp. 247–320.
[17] A. N. Shiryayev, *Selected Works of A.N. Kolmogorov. Vol.* II. *Probability Theory and Mathematical Statistics*, Kluwer Academic, Dordrecht, Boston, London, 1991.
[18] N. Th. Varopoulos, L. Saloff-Coste, and T. Coulhon, *Analysis and Geometry on Groups*, Cambridge Tracts in Math. 100, Cambridge University Press, Cambridge, UK, 1992.
[19] A. I. Vol'pert and S. I. Hudjaev, *Cauchy's problem for degenerate second order quasilinear parabolic equations*, Math. USSR-Sb., 7 (1970), pp. 365–387.
[20] L. Wang, *On the regularity theory of fully nonlinear parabolic equations.* I, Comm. Pure Appl. Math., 45 (1992), pp. 27–76.
[21] L. Wang, *On the regularity theory of fully nonlinear parabolic equations.* II, Comm. Pure Appl. Math., 45 (1992), pp. 141–178.

SIAM J. MATH. ANAL.
Vol. 35, No. 3, pp. 596–622

A GLOBAL STABILITY CRITERION FOR SCALAR FUNCTIONAL DIFFERENTIAL EQUATIONS*

EDUARDO LIZ†, VICTOR TKACHENKO‡, AND SERGEI TROFIMCHUK§

Abstract. We consider scalar delay differential equations $x'(t) = -\delta x(t) + f(t, x_t)$ $(*)$ with nonlinear f satisfying a sort of negative feedback condition combined with a boundedness condition. The well-known Mackey–Glass-type equations, equations satisfying the Yorke condition, and equations with maxima all fall within our considerations. Here, we establish a criterion for the global asymptotical stability of a unique steady state to $(*)$. As an example, we study Nicholson's blowflies equation, where our computations support the Smith conjecture about the equivalence between global and local asymptotical stabilities in this population model.

Key words. delay differential equations, global stability, Yorke condition, Schwarz derivative, Nicholson's blowflies equation

AMS subject classifications. 34K20, 92D25

DOI. 10.1137/S0036141001399222

1. Introduction.

We start by considering the simple autonomous linear equation

$$x'(t) = -\delta x(t) + ax(t-h), \tag{1.1}$$

governed by friction ($\delta \geq 0$) and delayed negative feedback ($a < 0$). Necessary and sufficient conditions for the asymptotic stability of (1.1) are well known [5]. For example, in the simplest case $\delta = 0$, (1.1) is asymptotically stable if and only if $-ah \in (0, \pi/2)$. If we allow for a variable delay in (1.1), we obtain the equation

$$x'(t) = -\delta x(t) + ax(t-h(t)),\ 0 \leq h(t) \leq h, \tag{1.2}$$

whose stability analysis is more complicated than that of the autonomous case. Nevertheless several sharp stability conditions were established for (1.2). The first of them is due to Myshkis (see [5, p. 164]) and it states that in the case $\delta = 0$ the inequality $-a \sup_{\mathbb{R}} h(t) < 3/2$ guarantees the asymptotic stability in (1.2). This condition is sharp (this fact was established by Myshkis himself). In particular, the upper bound $3/2$ cannot be increased to $\pi/2$. Later on, the result by Myshkis was improved by different authors, the most celebrated extensions being those of Yorke [17] and Yoneyama [16] (both for $\delta = 0$). Finally, the Myshkis condition has been recently generalized [6] for $\delta > 0$: equation (1.2) is asymptotically stable if

$$-\frac{\delta}{a} \exp(-h\delta) > \ln \frac{a^2 - a\delta}{\delta^2 + a^2}. \tag{1.3}$$

*Received by the editors December 5, 2001; accepted for publication (in revised form) March 28, 2003; published electronically October 2, 2003. This research was supported by FONDECYT (Chile), project 8990013.

http://www.siam.org/journals/sima/35-3/39922.html

†Departamento de Matemática Aplicada II, E.T.S.I. Telecomunicación, Universidad de Vigo, Campus Marcosende, 36280 Vigo, Spain (eliz@dma.uvigo.es). This author was supported in part by M.C.T. (Spain) and FEDER under project BFM2001-3884-C02-02.

‡Institute of Mathematics, National Academy of Sciences of Ukraine, Tereshchenkivs'ka str. 3, Kiev, Ukraine (vitk@imath.kiev.ua). This author was supported in part by F.F.D. of Ukraine, project 01.07/00109.

§Departamento de Matemáticas, Facultad de Ciencias, Universidad de Chile, Casilla 653, Santiago, Chile (trofimch@uchile.cl).

We note that for every fixed a, δ, and $h > 0$, condition (1.3) is sharp, and in the limit case $\delta = 0$ it coincides with the Myshkis condition. Here the sharpness means that if a, δ, h do not satisfy (1.3), then the asymptotic stability of (1.2) can be destroyed by an appropriate choice of a periodic delay $h(t)$ (see [6, Theorem 4.1]). Returning to (1.1), we can observe that (1.3) approximates exceptionally well the exact stability domain for (1.1) given in [5]; see Figure 2.1, where the domains of local (dashed line) and global (solid line) stability are shown in coordinates $(-a/\delta, \exp(-\delta h))$. When $\delta = 0$, we obtain $3/2$ as an approximation for $\pi/2$.

It is a rather surprising fact that the sharp global stability condition (1.3) works not only for linear equations but also for a variety of nonlinear delay differential equations of the form

$$x'(t) = -\delta x(t) + f(t, x_t), \quad (x_t(s) \stackrel{\text{def}}{=} x(t+s),\ s \in [-h, 0]), \tag{1.4}$$

where $f : \mathbb{R} \times \mathcal{C} \to \mathbb{R}$, $\mathcal{C} \stackrel{\text{def}}{=} C[-h, 0]$, is a measurable functional satisfying the additional condition (H) given below. Due to the rather general form of (H), (1.4) incorporates, possibly after some transformations, some of the most celebrated delay equations, such as equations satisfying the Yorke condition [17], equations of Wright [5, 8], Lasota–Wazewska and Mackey–Glass [2, 7, 10], and equations with maxima [6, 11]. Solutions to some of these equations can exhibit chaotic behavior so that the analysis of their global stability is of great importance—at least during the first stage of the investigation (see [7, p. 148] for further discussion). As an example, in section 2 we consider Nicholson's blowflies equation, for which our computations support the conjecture of Smith posed in [14].

Let us explain briefly the nature of our further assumptions. In part, they are motivated by the sharp stability results for (1.4) obtained in [17] ($\delta = 0$) and [6] ($\delta > 0$) under the assumption that for some $a < 0$ and for all $\phi \in \mathcal{C}$, the following Yorke condition holds:

$$a\mathcal{M}(\phi) \le f(t, \phi) \le -a\mathcal{M}(-\phi). \tag{1.5}$$

Here $\mathcal{M} : \mathcal{C} \to \mathbb{R}$ is the monotone continuous functional (sometimes called the Yorke functional) defined by $\mathcal{M}(\phi) = \max\{0, \max_{s\in[-h,0]} \phi(s)\}$. In general, f satisfying (1.5) is nonlinear in ϕ. On the other hand, in some sense it has a "quasi-linear" form (for example, $f(\phi) = \max_{s\in[-h,0]} \phi(s)$ can be written as $f(\phi) = \phi(-s_\phi)$). In particular, f is sublinear in ϕ, which makes impossible the application of the results from [6, 17] to the strongly nonlinear cases such as the celebrated Wright equation

$$x'(t) = a(1 - \exp(-x(t-h))),\ a < 0, \tag{1.6}$$

which is also globally asymptotically stable if $-ah \in (0, 3/2)$. Roughly speaking, the Yorke $3/2$-stability condition does not imply the Wright $3/2$-stability result. Our recent studies [8] of (1.6) revealed the following interesting fact: the essential feature of the function $f(x) = a(1 - \exp(-x))$ in (1.6) allowing the extension of the Wright $3/2$-stability result to some other nonlinearities is the position of the graph of f with respect to the graph of the rational function $r(x) = ax/(1 + bx)$ which coincides with f, f', and f'' at $x = 0$. This suggests the idea to consider a "rational in $\mathcal{M}$" version of the "linear in $\mathcal{M}$" condition (1.5) to manage the strongly nonlinear cases of (1.4). Therefore, we will assume the following conditions (H):

(H1) $f : \mathbb{R} \times \mathcal{C} \to \mathbb{R}$ satisfies the Carathéodory condition (see [5, p. 58]). Moreover, for every $q \in \mathbb{R}$ there exists $\vartheta(q) \ge 0$ such that $f(t, \phi) \le \vartheta(q)$ almost everywhere on $\mathbb{R}$ for every $\phi \in C$ satisfying the inequality $\phi(s) \ge q, s \in [-h, 0]$.

(H2) There are $b \geq 0,\ a < 0$ such that

$$(1.7)\quad f(t,\phi) \geq \frac{a\mathcal{M}(\phi)}{1+b\mathcal{M}(\phi)} \text{ for all } \phi \in \mathcal{C};$$

$$(1.8)\quad f(t,\phi) \leq \frac{-a\mathcal{M}(-\phi)}{1-b\mathcal{M}(-\phi)} \text{ for all } \phi \in \mathcal{C} \text{ such that } \min_{s\in[-h,0]} \phi(s) > -b^{-1} \in [-\infty, 0).$$

(H) is a kind of negative feedback condition combined with a boundedness condition; they will cause solutions to remain bounded and to tend to oscillate about zero. Furthermore, (H) implies that $x = 0$ is the unique steady state solution for (1.4) with $\delta > 0$. On the other hand, (H) does not imply that the initial value problems for (1.4) have a unique solution. In any case, the question of uniqueness is not relevant for our purposes. Notice finally that if (H2) holds with $b = 0$ (which is precisely (1.5)), then (H1) is satisfied automatically with $\vartheta(q) = -aM(-q)$.

Now we are ready to state the main result of this work.

THEOREM 1.1. *Assume that* (H) *holds and let* $x : [\alpha, \omega) \to \mathbb{R}$ *be a solution of* (1.4) *defined on the maximal interval of existence. Then* $\omega = +\infty$ *and* x *is bounded on* $[\alpha, +\infty)$. *If, additionally, condition* (1.3) *holds, then* $\lim_{t\to+\infty} x(t) = 0$. *Furthermore, condition* (1.3) *is sharp within the class of equations satisfying* (H)*: for every triple* $a < 0,\ \delta > 0,\ h > 0$ *which does not meet* (1.3), *there is a nonlinearity* f *satisfying* (H) *and such that the equilibrium* $x(t) = 0$ *of the corresponding equation* (1.4) *is not asymptotically stable.*

It should be noticed that in this paper we do not consider the limit cases when $b = 0$ and/or $\delta = 0$. When $b = 0, \delta > 0$, Theorem 1.1 was proved in [6, Theorem 2.9]. The limit case $\delta = 0,\ b \geq 0$ can be addressed by adapting the proofs in [8]. Here, due to the elimination of the friction term $-\delta x$, an additional condition is necessary (see [9] for details). In this latter case, (1.3) takes the limit form $-ah \in (0, 3/2)$.

REMARK 1.1. *The set of four parameters* ($h > 0, \delta > 0, a < 0, b > 0$) *can be reduced. Indeed, the change of time* $\tau = \delta t$ *transforms* (1.4) *into the same form but with* $\delta = 1$. *Finally, since* $\mathcal{M}$ *is a positively homogeneous functional* ($\mathcal{M}(k\phi) = k\mathcal{M}(\phi)$ *for every* $k \geq 0$, $\phi \in \mathcal{C}$), *and since the global attractivity property of the trivial solution of* (1.4) *is preserved under the simple scaling* $x = b^{-1}y$, *the exact value of* $b > 0$ *is not important and we can assume that* $b = 1$. *Also, the change of variables* $x = -y$ *transforms* (1.4) *into* $y'(t) = -\delta y(t) + [-f(t, -y_t)]$ *so that it suffices that at least one of the two functionals* $f(t,\phi), -f(t,-\phi)$ *satisfies* (1.7) *and* (1.8).

To prove Theorem 1.1, in sections 3 and 4 we will construct and study several one-dimensional maps which inherit the stability properties of (1.4). The form of these maps depends strongly on the parameters: in fact, we will split the domain of all admissible parameters given by (1.3) into several disjoint parts, and each one-dimensional map will be associated to a part. Some of the maps are rather simple, and an elementary analysis is sufficient to study their stability properties. Some other maps are more complicated: for example, the proof of Lemma 3.6 involves the concept of the Schwarz derivative, whose definition and several properties are recalled below. Unfortunately, several important one-dimensional maps appear in an implicit form, and though this form may be simple, its analysis requires considerable effort. For the convenience of the reader, the hardest and most technical parts of our estimations are placed in an appendix (section 6). In section 2, we will show the significance of the hypotheses (H) again by applying Theorem 1.1 to the well-known Nicholson blowflies equation.

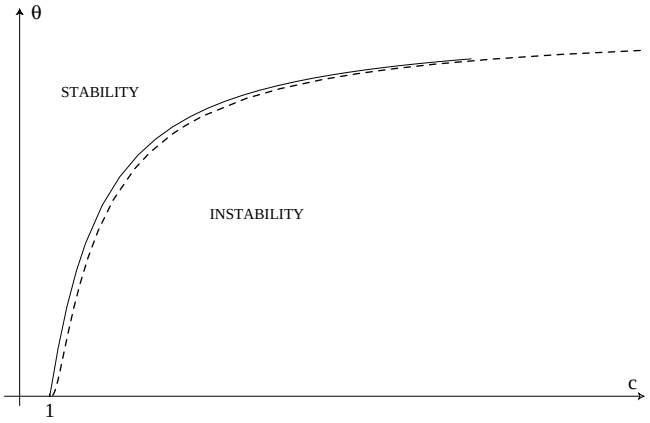


FIG. 2.1. *Domains of global and local stability in coordinates* (c, θ), $c = -a/\delta$, $\theta = \exp(-\delta h)$.

2. On the Smith conjecture and equations with nonpositive Schwarzian.

2.1. A global stability condition. In this section we will apply our results to the delay differential equation

$$N'(u) = -\delta N(u) + pN(u-h)e^{-\gamma N(u-h)},\ h > 0, \tag{2.1}$$

used by W. S. C. Gurney, S. P. Blythe, and R. M. Nisbet in 1980 to describe the dynamics of Nicholson's blowflies (see, for example, [1, section 3.6] or [14, p. 112]). Here p is the maximum per capita daily egg production rate, $1/\gamma$ is the size at which the population reproduces at its maximum rate, δ is the per capita daily adult death rate, h is the generation time, and $N(u)$ is the size of population at time u. In view of the biological interpretation, we consider only positive solutions of (2.1). If $p \le \delta$, then (2.1) has only one constant solution $x \equiv 0$. For $p > \delta$, the equation has an unstable constant solution $x \equiv 0$ and a unique positive equilibrium $N^* = \gamma^{-1}\ln(p/\delta)$. Global stability in (2.1) (when all positive solutions tend to the equilibrium N^*) has been studied by various authors by using different methods (see [2, 3, 14] for more references). Nevertheless, the exact global stability condition was not found. In this aspect, the work [14], where the conjecture about the equivalence between local and global asymptotic stabilities for (2.1) was posed (see [14, p. 116]), is of special interest for us. Indeed, an application of our main result to (2.1) strongly supports this conjecture, showing a surprising proximity between the boundaries of local and global stability domains; see Figure 2.1 and the following theorem.

THEOREM 2.1. *The positive equilibrium N^* of Nicholson's blowflies equation* (2.1) *is globally asymptotically stable if either $c \in (-1, 0]$ or*

$$\theta > c\ln[(c^2+c)/(c^2+1)]\,,\ c > 0, \tag{2.2}$$

where $\theta = \exp(-\delta h)$, $c = \ln(p/\delta) - 1$.

It follows from the observation given below (1.3) that condition (2.2) is sharp within the class of equations $N'(u) = -\delta N(u) + pN(u-\rho(u))e^{-\gamma N(u-\rho(u))}$ with variable delay $\rho : \mathbb{R} \to [0, h]$.

As can be seen from (2.2), not all parameters are independent in (2.1). Indeed, if we set $\tau = h\delta, u = t/\delta, q = p/\delta, x(t) = \gamma N(u)$, then (2.1) takes the form

$$x'(t) = -x(t) + g(x(t-\tau)), \tag{2.3}$$

where $g(x) = qx\exp(-x)$. For every $q > 1$, it has a unique positive equilibrium $x(t) \equiv \ln q$, which is globally asymptotically stable if $\ln(q) \le 2$ (see [3]). The change of variables $x(t) = \ln q + y(t)$ reduces (2.3) to the equation $y'(t) = -y(t) + w(y(t-\tau))$, where $w(y) = (y + \ln q)e^{-y} - \ln q$. In section 5, we will show that the nonlinearity $w(y)$ satisfies the following conditions (W) within some domain which attracts all nonnegative solutions of (2.3):

(W1) $w \in C^3(\mathbb{R}, \mathbb{R})$, $xw(x) < 0$ for $x \neq 0$, and $w'(0) < 0$.

(W2) w is bounded below and has at most one critical point $x^* \in \mathbb{R}$ which is a local extremum.

(W3) The Schwarz derivative $(Sw)(x) = w'''(x)(w'(x))^{-1} - (3/2)\left(w''(x)(w'(x))^{-1}\right)^2$ of w is nonnegative: $(Sw)(x) \le 0$ for all $x \neq x^*$.

Since $w'(0) = \ln(e\delta/p) < 0$ and $w''(0) > 0$ if $\ln q > 2$, Theorem 2.1 is a consequence of the following results.

LEMMA 2.2 (see [8]). *Let w meet conditions* (W) *and $w''(0) > 0$. Then the functional $f(t, \phi) = w(\phi(-h))$ satisfies hypotheses* (H) *with $a = w'(0)$ and $b = -w''(0)/(2w'(0))$.*

COROLLARY 2.3. *Suppose that w satisfies* (W) *and $w''(0) > 0$. If* (1.3) *holds with $a = w'(0)$, then the trivial steady state attracts all solutions of the delay differential equation*

$$x'(t) = -\delta x(t) + w(x(t-h)),\ \delta > 0. \tag{2.4}$$

Corollary 2.3 can be applied in a similar way to obtain global stability conditions for the positive equilibrium of other delay differential equations arising in biological models. For example, we can mention the celebrated Mackey–Glass equation proposed in 1977 to model blood cell populations (see, e.g., [10]), which is of the form (2.4) with $w(x) = b/(1 + x^n)$, $b > 0$, $n > 1$. Another important model that can be considered within our approach is the Wazewska-Czyzewska and Lasota equation describing the erythropoietic (red blood cell) system. In this case $w(x) = b_1\exp(-b_2 x)$, $b_i > 0$.

As proved in [8], the conclusion of Corollary 2.3 also holds for $\delta = 0$ by replacing (1.3) with its limit form $-ah \le 3/2$. In the particular case of the Wright equation, this result coincides with the 3/2-stability theorem by Wright (see [8] for more details).

2.2. The Smith and Wright conjectures revisited. Let us look again at Figure 2.1, which shows the boundaries of the domains of local and global asymptotic stability for the Nicholson equation; this observation (as well as Proposition 3.3 stated below) suggests the following.

CONJECTURE 2.1. *Under conditions* (W), *the trivial solution of* (2.4) *is globally attracting if it is locally asymptotically stable.*

An interesting particularity of Conjecture 2.1 is that it coincides with the celebrated Wright conjecture if we take $\delta = 0$, $w(x) = a(1 - \exp(-x))$, and it coincides with the Smith conjecture if we take Nicholson's blowflies equation.

Now, the following result was obtained in [15] as a simple consequence of an elegant approach toward stable periodic orbits for (2.4) with Lipschitz nonlinearities.

PROPOSITION 2.4 (see [15]). *For every $\alpha \ge 0$ there exists a smooth strictly decreasing function w satisfying* (W1), (W2), *$-w'(0) = \alpha$, and such that* (2.4) *has a nontrivial periodic solution which is hyperbolic, stable, and exponentially attracting with asymptotic phase (so therefore* (2.4) *is not globally stable).*

Proposition 2.4 shows clearly that the strong dependence between local (at zero) and global asymptotical stabilities of (2.4) cannot be explained only with the concepts

presented in (W1), (W2). We notice here that the condition of negative Schwarz derivative in (2.4) appears naturally also in some other contexts of the theory of delay differential equations; see, e.g., [10, sections 6–9], where it is explicitly used, and [5, Theorem 7.2, p. 388], where the condition $Sw < 0$ is implicitly required.

3. Preliminary stability analysis of (1.4). Throughout the paper, in view of Remark 1.1, we assume that $\delta = 1$ in (1.4) and $b = 1$ in (1.7), (1.8). Hence, with $\theta \stackrel{\text{def}}{=} \exp(-h)$, (1.3), (1.4), (1.7), and (1.8) take, respectively, the forms

$$-\theta/a > \ln \frac{a^2 - a}{a^2 + 1}; \tag{3.1}$$

$$x'(t) = -x(t) + f(t, x_t); \tag{3.2}$$

$$f(t, \phi) \geq r(\mathcal{M}(\phi)) \text{ for all } \phi \in \mathcal{C}; \tag{3.3}$$

$$f(t, \phi) \leq r(-\mathcal{M}(-\phi)) \text{ for all } \phi \in \mathcal{C} \text{ such that } \min_{s \in [-h, 0]} \phi(s) > -1, \tag{3.4}$$

where the rational function $r(x) = ax/(1+x)$ will play a key role in our constructions. In this section, we establish that the "linear" approximation to (3.1) of the form

$$-\theta/a > -(a+1)/(a^2+1) \tag{3.5}$$

implies the global stability of (3.2) (note here that $\ln(1 + x) < x$ is true for $x > 0$).

In what follows we will use some properties of the Schwarz derivative. The following lemma can be checked by direct computation.

LEMMA 3.1. *If g and f are functions which are at least C^3, then $S(f \circ g)(x) = (g'(x))^2 (Sf)(g(x)) + (Sg)(x)$. As a consequence, the inverse f^{-1} of a smooth diffeomorphism f with $Sf > 0$ has negative Schwarzian: $Sf^{-1} < 0$.*

We will also need the following lemma from [13].

LEMMA 3.2 (see [13, Lemma 2.6]). *Let $q : [\alpha, \beta] \to [\alpha, \beta]$ be a C^3 map with $(Sq)(x) < 0$ for all x. If $\alpha < \gamma < \beta$ are consecutive fixed points of some iteration $g = q^N$ of q, $N \geq 1$, and $[\alpha, \beta]$ contains no critical point of g, then $g'(\gamma) > 1$.*

This lemma allows us to prove the following proposition, which plays a central role in our analysis.

PROPOSITION 3.3. *Let $q : [\alpha, \beta] \to [\alpha, \beta]$ be a C^3 map with a unique fixed point γ and with at most one critical point x^* (maximum). If γ is locally asymptotically stable and the Schwarzian derivative $(Sq)(x) < 0$ for all $x \neq x^*$, then γ is the global attractor of q.*

Proof. Let W be the connected component of the open set $S = \{x \in [\alpha, \beta] : \lim_{k \to +\infty} q^k(x) = \gamma\}$ which contains γ. Clearly, $g(W) \subset W$. If $W \neq [\alpha, \beta]$, then we have three possibilities: $W = [\alpha, r)$, $W = (l, \beta]$, or $W = (l, r)$, $\alpha < l < r < \beta$.

If $W = [\alpha, r)$, then $q(r) = \lim_{\epsilon \to 0+} q(r - \epsilon) \leq r \leq q(r)$, a contradiction with the fact that q does not have fixed points in $[\alpha, \beta]$ different from γ. The case $W = (l, \beta]$ is completely analogous.

In the case $W = (l, r)$, by the same arguments, it should hold that $q(l) = r$, $q(r) = l$. Thus $l < \gamma < r$ are consecutive fixed points of $g = q^2$ and $g'(\gamma) = (q'(\gamma))^2 \leq 1$. By Lemma 3.2, $x^* \in (l, r)$, and therefore $q(x^*) < r$. Since q has a maximum at x^*, $r > q(x^*) \geq q(l)$, a contradiction.

Hence $W = [\alpha, \beta]$, and therefore $\{\gamma\}$ attracts each point of $[\alpha, \beta]$. This implies that γ is the global attractor of q (see [4, Chapter 2]). ∎

Now we are in a position to begin the stability analysis of (3.2).

LEMMA 3.4. *Suppose that* (H) *holds and let* $x : [\alpha - h, \omega) \to \mathbb{R}$ *be a solution of* (3.2) *defined on the maximal interval of existence. Then* $\omega = +\infty$ *and* $M = \limsup_{t\to\infty} x(t)$, $m = \liminf_{t\to\infty} x(t)$ *are finite. Moreover, if* $m \geq 0$ *or* $M \leq 0$, *then* $M = m = \lim_{t\to\infty} x(t) = 0$.

Proof. Note that (3.3) implies that $f(t, \phi) \geq a$ for all $t \in \mathbb{R}$ and $\phi \in \mathcal{C}$. Next, if $q \leq x_\alpha(s) \leq Q$, $s \in [-h, 0]$, then for all $t \in [\alpha, \omega)$, we have

$$x(t) = \exp(-(t-\alpha))x(\alpha) + \int_\alpha^t \exp(-(t-s))f(s, x_s)ds \geq a + (\min\{q, a\} - a)\exp(-(t-\alpha)) \geq \min\{q, a\}. \tag{3.6}$$

Next, (H1) implies that $f(s, x_s) \leq \vartheta(\min\{q, a\})$ for all $s \geq \alpha$, so that

$$x(t) \leq \max\{Q, 0\} + \vartheta(\min\{q, a\}),\ t \in [\alpha, \omega).$$

Hence $x(t)$ is bounded on the maximal interval of existence, which implies the boundedness of the right-hand side of (3.2) along $x(t)$. Thus $\omega = +\infty$ due to the corresponding continuation theorem (see [5, Chapter 2]).

Next, suppose, for example, that $M = \limsup_{t\to\infty} x(t) \leq 0$. Thus we have $\lim_{t\to\infty} \mathcal{M}(x_t) = 0$ so that, by virtue of (3.3), $f(t, x_t) \geq \inf_{s\geq t} a\mathcal{M}(x_s)/(1+\mathcal{M}(x_s)) \stackrel{\text{def}}{=} a(t)$, where $a : [\alpha, +\infty) \to (-\infty, 0]$ is nondecreasing and continuous, with $\lim_{t\to+\infty} a(t) = 0$. Thus, by (3.6), $x(t) \geq \exp(-(t-\beta))x(\beta) + a(\beta)$ for all $t \geq \beta > \alpha$. This implies that $m = \liminf_{t\to\infty} x(t) = 0$ so that $M = 0$. □

LEMMA 3.5. *Suppose that* (H) *holds and let* $x : [\alpha - h, \infty) \to \mathbb{R}$ *be a solution of* (3.2). *If* x *has a negative local minimum at some point* $s > \alpha$, *then* $\mathcal{M}(x_s) > 0$. *Analogously, if* x *has a positive local maximum at* $t > \alpha$, *then* $\mathcal{M}(-x_t) > 0$.

Proof. If $x(u) \leq 0$ for all $u \in [s - h, s]$, then $x'(s) \geq -x(s) + r(\mathcal{M}(x_s)) > 0$, a contradiction. The other case is similar. □

LEMMA 3.6. *Suppose that* (H) *holds and let* $x : [\alpha - h, \infty) \to \mathbb{R}$ *be a solution of* (3.2). *If* (3.5) *holds, then* $\lim_{t\to\infty} x(t) = 0$.

Proof. Let $M = \limsup_{t\to\infty} x(t)$, $m = \liminf_{t\to\infty} x(t)$. In view of Lemma 3.4, we only have to consider the case $m < 0 < M$, since otherwise ($m \geq 0$ or $M \leq 0$) we have a nonoscillatory solution to (3.2), which tends to zero as $t \to +\infty$. Thus in what follows we will consider only the oscillating solutions $x(t)$. In this case there are two sequences of points t_j, s_j of local maxima and local minima, respectively, such that $x(t_j) = M_j \to M, x(s_j) = m_j \to m$, and $s_j, t_j \to +\infty$ as $j \to \infty$.

First we prove that $M = m = 0$ if $a(1-\theta) > -1$. Indeed, for each s_j we can find $\varepsilon_j \to 0+$ such that $0 < \mathcal{M}(x_s) < M + \varepsilon_j$ for all $s \in [s_j - h, s_j]$. Next, by Lemma 3.5, there exists $h_j \in [0, h]$ such that $x(s_j - h_j) = 0$. Therefore, by the variation of constants formula,

$$m_j = \int_{s_j-h_j}^{s_j} e^{s-s_j} f(s, x_s)ds \geq \int_{s_j-h_j}^{s_j} e^{s-s_j} r(\mathcal{M}(x_s))ds \geq r(M + \varepsilon_j)(1-\theta).$$

As a limit form of this inequality, we get $m \geq r(M)(1-\theta)$. Hence $m > -1$ and we can use (3.4) for $\phi = x_t$ with sufficiently large t. Thus, in a similar way, we obtain that

$$M_j = \int_{t_j-h_j^*}^{t_j} e^{s-t_j} f(s, x_s)ds \leq \int_{t_j-h_j^*}^{t_j} e^{s-t_j} r(-\mathcal{M}(-x_s))ds \leq r(m - \varepsilon_j^*)(1-\theta)$$

for some sequences $\varepsilon_j^* \to 0+$ and $h_j^* \in [t_j - h, t_j]$. Hence we obtain $M \le r(m)(1-\theta) \le r(r(M)(1-\theta))(1-\theta)$. This gives $M^2 \le M(a(1-\theta)-1)$, which is only possible when $M = 0$.

Now, assume that $a(1-\theta) \le -1$. Since $m \ge r(M)(1-\theta) > a$ (see the first part of the proof), we conclude that $r^{-1}(m) = m/(a-m) > 0$ is well defined. Next, for s_j we can find a sequence of positive $\epsilon_j \to 0$ such that $m_j < m + \epsilon_j < 0$. We claim that $x(s_j - h_j) \ge r^{-1}(m + \epsilon_j)$ for some $h_j \in [0, h]$. Indeed, in the opposite case, $x(s) < r^{-1}(m + \epsilon_j)$ for all s in some open neighborhood of $[s_j - h, s_j]$. Thus $f(s, x_s) \ge r(\mathcal{M}(x_s)) > m + \epsilon_j$ for all s close to s_j. Finally, $x'(s) > -x(s) + m + \epsilon_j > 0$ almost everywhere in some neighborhood of s_j, contradicting the choice of s_j.

Next, there exists a sequence of positive $\epsilon_j^* \to 0$ such that $x(s) < M + \epsilon_j^*$ for all $s \in [s_j - h, s_j]$. Therefore, by the variation of constants formula,

$$\begin{aligned} m_j = x(s_j) &= e^{-h_j} x(s_j - h_j) + \int_{s_j - h_j}^{s_j} e^{s - s_j} f(s, x_s) ds \\ &\ge e^{-h_j} r^{-1}(m + \epsilon_j) + \int_{s_j - h_j}^{s_j} e^{s - s_j} r(\mathcal{M}(x_s)) ds \\ &\ge e^{-h_j} r^{-1}(m + \epsilon_j) + r(M + \epsilon_j^*)(1 - e^{-h_j}) \ge \theta r^{-1}(m + \epsilon_j) + r(M + \epsilon_j^*)(1 - \theta), \end{aligned}$$

so that $m - \theta r^{-1}(m) \ge r(M)(1-\theta) \ge a(1-\theta)$. This implies that $m \ge a(1-\theta)(a-m)/(a-m-\theta) > -1$, where the last inequality is evident when $1 + a(1-\theta) = 0$ and follows from the relations $m < 0 \le (a^2(1-\theta) + a - \theta)/(1 + a(1-\theta))$ otherwise. Since $m > -1$ we can use (3.4) for $\phi = x_t$ with sufficiently large t. Thus, in a similar way, we obtain that

$$\begin{aligned} M_j = x(t_j) = e^{-h_j^\#} x(t_j - h_j^\#) + \int_{t_j - h_j^\#}^{t_j} e^{s - t_j} f(s, x_s) ds \le e^{-h_j^\#} r^{-1}(M - \varepsilon_j) \\ + \int_{t_j - h_j^\#}^{t_j} e^{s - t_j} r(-\mathcal{M}(-x_s)) ds \le \theta r^{-1}(M - \varepsilon_j) + r(m - \varepsilon_j^*)(1 - \theta) \end{aligned}$$

for some sequences $\varepsilon_j, \varepsilon_j^* \to 0+$ and $h_j^\# \in [t_j - h, t_j]$. Thus $\psi(M) \stackrel{\text{def}}{=} M - \theta r^{-1}(M) \le r(m)(1-\theta)$. Now, $\psi : (a, +\infty) \to \mathbb{R}$ is a strictly increasing bijection so that $\chi(x) = \psi^{-1}((1-\theta) r(x))$ is well defined and strictly decreases on $(-1, +\infty)$. A direct computation shows that $\chi(-1^-) = +\infty$ and that $\chi(+\infty) = \psi^{-1}((1-\theta)a) > -1$. Therefore $\chi : [\chi(+\infty), \chi^2(+\infty)] \to [\chi(+\infty), \chi^2(+\infty)]$. Moreover, since $M \le \chi(m),\ m \ge \chi(M)$, we conclude that $m, M \in [\chi(+\infty), \chi^2(+\infty)]$ and that $[m, M] \subset \chi([m, M])$. Next, for $x > a$ we obtain by direct computation that $(S\psi)(x) = -6\theta a(a^2 - 2xa + x^2 - \theta a)^{-2} > 0$.

Since $(Sr)(x) = 0$ for all $x > -1$, it follows from Lemma 3.1 that

$$(S\chi)(x) = ((1-\theta) r'(x))^2 (S\psi^{-1})((1-\theta) r(x)) < 0.$$

Finally, by (3.5), $\chi'(0) = (1-\theta)a^2/(a-\theta) \in (-1, 0)$ so that we apply Proposition 3.3 (where we set $q = \chi$, $[\alpha, \beta] = [\chi(+\infty), \chi^2(+\infty)]$, and $\gamma = 0$) to conclude that $\chi^k([\alpha, \beta]) \to 0$ as $k \to \infty$. Since $[m, M] \subseteq \chi^k([m, M]) \subseteq \chi^k([\alpha, \beta])$ for all integers $k \ge 1$, it is clear that $m = M = 0$. □

4. Proof of the main result. The analysis done in the previous section shows that the only case that remains to be considered is when

$$0 < \ln \frac{a^2 - a}{a^2 + 1} < -\theta/a \le -\frac{a+1}{a^2+1}.$$

This case will be studied in the present section: we start describing a finer decomposition of the above-indicated domain of parameters (denoted below as $\mathcal{D}$).

4.1. Notation and domains. In what follows, we will always assume that $h > 0$ and $a < -1$, and we will use the following notation:

$$\theta = \exp(-h); \quad \lambda = \exp(\theta/a); \quad a_* = a + \frac{\theta}{1-\theta}; \quad \mu = -\frac{1}{a};$$

$$\alpha(a,\theta) = (1-a)\exp(\theta/a) + a;$$

$$\beta(a,\theta) = -\frac{a^2 + \exp(\theta/a)(1-2a+2\theta(a-1)) - (1-a)^2\exp(2\theta/a)}{a^2 + (a-a^2)\exp(\theta/a)};$$

$$\gamma(a,\theta) = a^3\alpha(a,\theta)\frac{1-\theta+\ln\theta}{2-\theta+\ln\theta}; \quad \mathcal{R}(r) = \mathcal{R}(r,a,\theta) = \frac{\alpha(a,\theta)r}{1-\beta(a,\theta)r}.$$

Obviously, $\theta,\ \mu \in (0,1),\ a_* > a$, and $\gamma(a,\theta)$ is well defined for all $\theta \in (0.16, 1)$, where it can be checked that $2 - \theta + \ln\theta > 0$. Next, we will need the following four curves considered within the open square $(\theta,\mu) \in (0,1)^2$:

$$\Pi_1 = \left\{(\theta,\mu)\ :\ \theta = \Pi_1(\mu) \stackrel{\text{def}}{=} \frac{1-\mu}{1+\mu^2}\right\};\ \Pi_2 = \left\{(\theta,\mu)\ :\ \theta = \Pi_2(\mu) \stackrel{\text{def}}{=} \frac{1}{\mu}\ln\frac{1+\mu}{1+\mu^2}\right\};$$
$$\Pi_3 = \left\{(\theta,\mu)\ :\ \theta = \Pi_3(\mu) \stackrel{\text{def}}{=} \frac{95-108\mu}{5(19+5\mu)}\right\};\ \Pi_4 = \left\{(\theta,\mu)\ :\ \theta = \Pi_4(\mu) \stackrel{\text{def}}{=} 0.8\right\}.$$

The geometric relations existing between curves Π_1–Π_4 are shown schematically on Figure 4.1. Notice that all three curves $\Pi_j,\ j \neq 4$, have the following asymptotics at zero: $\Pi_j(\mu) = 1 - k_j\mu + o(\mu)$, where $k_1 = 1, k_2 = 1.5, k_3 = 1.4$. An elementary analysis shows that Π_3 does not intersect Π_1 and Π_2 when $\theta \in (0.8, 1)$. Next, to prove our main result, we will have to use different arguments for the different domains of parameters a, h. For this purpose, we introduce here the following three subsets $\mathcal{D}, \mathcal{D}^*, \mathcal{S}$ of $(0,1)^2$:

$$\mathcal{D} = \{(\theta,\mu) : \Pi_2(\mu) \le \theta \le \Pi_1(\mu)\};$$

$$\mathcal{D}^* = \mathcal{D} \setminus \mathcal{S},\ \text{where } \mathcal{S} = \{(\theta,\mu) \in \mathcal{D} :\ \theta \in [0.8, 1),\ \Pi_3(\mu) \le \theta \le \Pi_1(\mu)\}.$$

We can see that $\mathcal{D}$ is situated between Π_1 and Π_2, while the sector $\mathcal{S}$ is placed among Π_1, Π_3, and Π_4. Sometimes it will be more convenient for us to use the coordinates (a,θ) instead of (θ,μ); we will preserve the same symbols for the domains and curves considered both in (a,θ) and (θ,μ).

Let us end this subsection indicating several useful estimations which will be of great importance for the proof of our main result.

LEMMA 4.1. *We have $\alpha(a,\theta) > 0$, $\beta(a,\theta) > 0$, and $a\alpha(a,\theta)/(1-a\beta(a,\theta)) > -1$ for all $(a,\theta) \in \mathcal{D}$. Next, if $(a,\theta) \in \mathcal{S}$, then $\gamma(a,\theta) < 1$.*

The proof of the lemma is given in section 6 (Lemmas 6.1, 6.2, and 6.3).

4.2. One-dimensional map $F : \mathbf{I} \to \mathbb{R}$. Throughout this subsection, we will suppose that $(a,\theta) \in \mathcal{D}$. Therefore $a(\theta-1)/\theta - 1 > 0$ so that the interval $I = (-1, a(\theta-1)/\theta - 1)$ is not empty. Furthermore, $t_1 = t_1(z) = -\ln(1 - z/r(z)) \in [-h, 0]$

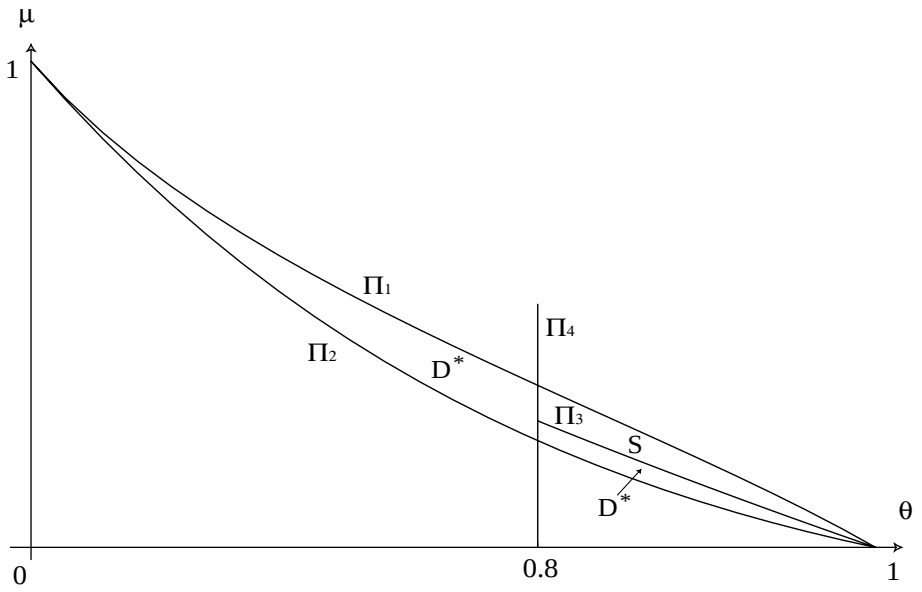


FIG. 4.1. *Domains of global stability in coordinates* (θ, μ).

for every $z \in I \setminus \{0\}$. Consider now the map $F : I \to \mathbb{R}$ defined in the following way:

$$F(z) = \begin{cases} 0 & \text{if } z = 0; \\ \min\limits_{t\in[0,h]} y(t,z) & \text{if } z \in I \text{ and } z > 0; \\ \max\limits_{t\in[0,h]} y(t,z) & \text{if } z \in (-1,0), \end{cases}$$

where $y(t,z)$ is the solution of the initial value problem $y(s,z) = z,\ s \in [t_1(z) - h, t_1(z)],\ z \in I$, for

$$y'(t) = -y(t) + r(y(t-h)). \tag{4.1}$$

Observe that $y(0,z) = 0$ for all $z \in I$ since $y(t,z) = r(z)(1 - \exp(-t))$ for all $t \in [t_1(z), t_1(z) + h]$. The following lemma explains why we have introduced such F (moreover, condition (3.1) says precisely that $F'(0) > -1$; see section 6.2).

LEMMA 4.2. *Let $x(t)$ be a solution of (3.2) and set $M = \limsup_{t\to\infty} x(t)$, $m = \liminf_{t\to\infty} x(t)$. If $m, M \in I$, then $m \geq F(M)$ and $M \leq F(m)$.*

Proof. Consider two sequences of extremal values $m_j = x(s_j) \to m$, $M_j = x(t_j) \to M$ such that $s_j \to +\infty, t_j \to +\infty$ as $j \to \infty$. Let $\varepsilon > 0$ be such that $(m - \varepsilon, M + \varepsilon) \subset I$. Then $m_j \geq m - \varepsilon$ and $M_j \leq M + \varepsilon$ for big j. We will prove that $m \geq F(M)$, the case $M \leq F(m)$ being completely analogous.

By Lemma 3.5, we can find $\tau_j \in [s_j - h, s_j]$ such that $x(\tau_j) = 0$ while $x(t) < 0$ for $t \in (\tau_j, s_j]$. Next, $v_j = \tau_j + t_1(M+\varepsilon) \geq \tau_j - h$ because of $M + \varepsilon \in I$. Thus the solution $y(t)$ of (4.1) with initial condition $y(s) = M + \varepsilon,\ s \in [v_j - h, v_j]$, satisfies $y(\tau_j) = 0$ while $M + \varepsilon \equiv y(t) \geq x(t)$ for all $t \in [v_j - h, v_j]$. Furthermore, for all $s \in [v_j, \tau_j]$, we have $\mathcal{M}(x_s) \leq M + \varepsilon$ so that, by (3.3), $f(s, x_s) \geq r(\mathcal{M}(x_s)) \geq r(M + \varepsilon)$, and

$$y(t) - x(t) = \int_{\tau_j}^{t} e^{-(t-s)}[r(M+\varepsilon) - f(s, x_s)]ds \geq 0,\ t \in [v_j, \tau_j],$$

proving that $y(t) \geq x(t)$ for all $t \in [\tau_j - h, \tau_j]$. Now, for $t \in (\tau_j, s_j]$,

$$y(t) - x(t) = \int_{\tau_j}^{t} \exp\{-(t-s)\}(r(y(s-h)) - f(s, x_s))ds \leq 0, \tag{4.2}$$

since $f(s, x_s) \geq r(\mathcal{M}(x_s)) \geq r(\mathcal{M}(y_s)) = r(y(s-h))$. Hence, by (4.2), $m_j = x(s_j) \geq y(s_j) \geq F(M+\varepsilon)$, which proves that $m \geq F(M)$. □

To study the properties of F, we use its more explicit form given below.

LEMMA 4.3. *Set $r^{-1}(u) = u/(a-u)$. For $z \in I$, $(F(z) - r(z))z \geq 0$ and*

$$\theta = \int_{r(z)}^{F(z)} \frac{du}{r^{-1}(u) - r(z)}. \tag{4.3}$$

Proof. Let us consider $z > 0$, the case $z < 0$ being similar. Consider the solution $y(t, z)$ of (4.1), and recall that $y(t, z) = z$ for $t \in [t_1 - h, t_1]$, where $t_1 = -\ln(1 - z/r(z)) \in [-h, 0]$. Next, $y(t, z) = r(z)(1 - \exp(-t))$, $t \in [t_1, t_1 + h]$, so that $y(0, z) = 0$ and $y'(h, z) = -y(h, z)$. Therefore $F(z) = y(t_*, z)$ at some point $t_* \in (t_1 + h, h)$, where also $y'(t_*, z) = 0$.

Since $t_* \in [t_1 + h, h]$, by the variation of constants formula we have

$$y(t_*, z) = F(z) = e^{-(t_* - h)} \left[y(t_1 + h, z)e^{t_1} + \int_{t_1}^{t_* - h} e^v r(y(v, z))dv \right]. \tag{4.4}$$

On the other hand, $y'(t_*, z) = 0 = -y(t_*, z) + r(y(t_* - h, z))$, so that $F(z) = y(t_*, z) = r(y(t_* - h, z)) \geq r(z)$ and $r^{-1}(F(z)) = y(t_* - h, z) = r(z)[1 - \exp\{-(t_* - h)\}]$. Thus

$$t_* - h = \ln(r(z)/[r(z) - r^{-1}(F(z))]).$$

Now let $y(v, z) = w$ (so that $\exp(v) = r(z)/(r(z) - w)$); then

$$\int_{t_1}^{t_* - h} e^v r(y(v, z))dv = \int_{z}^{r^{-1}(F(z))} r(w) d\frac{r(z)}{r(z) - w}$$

$$= r(z) \left[\frac{r(z)}{z - r(z)} - \frac{F(z)}{r^{-1}(F(z)) - r(z)} + \int_{z}^{r^{-1}(F(z))} \frac{dr(w)}{w - r(z)} \right].$$

Now putting the last expression and the values of t_1, $t_* - h$ in (4.4), we get (4.3). □

Finally, we state an important technical lemma whose proof can be found in Lemmas 6.5 and 6.6 of the appendix.

LEMMA 4.4. *Assume that $(a, \theta) \in \mathcal{D}$. Then $F(z) < \mathcal{R}(r(z))$ if $z \in ((a\beta - 1)^{-1}, 0)$, and $F(z) > \mathcal{R}(r(z))$ if $z \in (0, a(\theta - 1)/\theta - 1)$, where $\mathcal{R}$ is defined in subsection 4.1.*

We will also consider $\mathcal{F} : (a_*, +\infty) \to \mathbb{R}$ defined by $\mathcal{F}(x) = F(x/(a - x))$. It can be easily seen that $\mathcal{F}(r(z)) = F(z)$ for all $z \in I$.

4.3. One-dimensional map $F_1 : [0, +\infty) \to (a, 0]$. By definition, for $z \geq 0$, $F_1(z) = \min_{t \in [0,h]} y(t)$, where $y(t, z)$ satisfies (4.1) and has the initial value $y(s, z) = (1 - e^{-s})r(z)$, $s \in [-h, 0]$. We will need the following lemma.

LEMMA 4.5. *Let $x(t)$ be a solution of (3.2) and set $M = \limsup_{t \to \infty} x(t)$, $m = \liminf_{t \to \infty} x(t)$. If $(a, \theta) \in \mathcal{D}$, then $m \geq F_1(M)$.*

Proof. Take $\varepsilon, s_j, t_j, m_j, M_j, \tau_j$ as in the first two paragraphs of the proof of Lemma 4.2. Then, for $t \in [\tau_j - h, \tau_j]$, we have

$$x(t) = \int_{\tau_j}^{t} e^{-(t-u)} f(u, x_u) du \le \int_{\tau_j}^{t} e^{-(t-u)} r(M+\varepsilon) du = y(t - \tau_j, M + \varepsilon).$$

Thus, if $u \in [s_j - h, s_j]$, then $\mathcal{M}(x_u(s)) \le \mathcal{M}(y_u(s, M+\varepsilon))$ so that

$$f(u, x_u) \ge r(\mathcal{M}(y_u(s, M+\varepsilon))) = r(r(M)(1 - e^{-(u-h-\tau_j)})).$$

This implies that

$$\begin{aligned} m_j = x(s_j) &\ge \int_{\tau_j}^{s_j} e^{-(s_j-u)} r(r(M)(1 - e^{-(u-h-\tau_j)})) du \\ &= \int_{\tau_j}^{s_j} e^{-(s_j-u)} r(y(u-h, M+\varepsilon)) du = y(s_j - \tau_j, M+\varepsilon) \ge F_1(M+\varepsilon). \end{aligned}$$

Since $\varepsilon > 0$ and $m_j \to m$ are arbitrary, the lemma is proved. ∎

LEMMA 4.6. *Set $r_1(z) = r(r(z)(1 - e^h))$. For $z > 0$ we have that $F_1(z) > a$ and*

$$\frac{r_1(z)\theta}{r(z)} = \int_{r_1(z)}^{F_1(z)} \frac{du}{r^{-1}(u) - r(z)}. \tag{4.5}$$

Proof. Take $t_* \in (0, h)$ such that

$$F_1(z) = y(t_*, z) = \int_0^{t_*} e^{-(t_*-u)} r(r(z)(1 - e^{-(u-h)})) du. \tag{4.6}$$

Since $y'(h) > 0$, we have that $y'(t_*) = 0$, and therefore $F_1(z) = y(t_*) = -y'(t_*) + r(y(t_* - h)) = r(y(t_* - h)) > a$. This implies that $r^{-1}(F_1(z)) = y(t_* - h) = r(z)(1 - \exp\{-(t_* - h)\})$, from which

$$t_* - h = \ln(r(z)/[r(z) - r^{-1}(F_1(z))]). \tag{4.7}$$

Now, using (4.7) and setting $\xi = r(z)(1 - e^{-(u-h)})$ in (4.6), we obtain

$$\begin{aligned} F_1(z) &= -(r(z) - r^{-1}(F_1(z))) \int_{r(z)(1-e^h)}^{r^{-1}(F_1(z))} r(\xi) d\frac{1}{\xi - r(z)} \\ &= -(r(z) - r^{-1}(F_1(z))) \left(\frac{r(\xi)}{\xi - r(z)} \Bigg|_{r(z)(1-e^h)}^{r^{-1}(F_1(z))} + \int_{r(z)(1-e^h)}^{r^{-1}(F_1(z))} \frac{dr(\xi)}{\xi - r(z)} \right). \end{aligned}$$

Simplifying this relation, we obtain (4.5). ∎

We conclude this section by stating two lemmas which compare F_1 and the associated function $\mathcal{F}_1(r) \stackrel{\text{def}}{=} F_1(r/(a-r))$ with rational functions. The proofs of these statements are based on rather careful estimations of identity (4.5) and are given in Lemmas 6.7, 6.10, and 6.11 in the appendix. (It should be noted that $\mathcal{R}$ approximates $\mathcal{F}_1$ extremely well so that a very meticulous analysis of (4.5) is needed.)

LEMMA 4.7. *If $(a, \theta) \in \mathcal{D}^*$ and $z \ge a(\theta - 1)/\theta - 1$, then $F_1(z) > \mathcal{R}(r(z))$.*

LEMMA 4.8. *If $(a, \theta) \in \mathcal{S}$ and $z > 0$, then*

$$\mathcal{F}_1(r(z)) > \frac{1 + \ln\theta - \theta}{2 + \ln\theta - \theta} \frac{ar(z)}{1 + r(z)\frac{1+\ln\theta-\theta}{1-\theta}} = \mathcal{R}_2(r(z)). \tag{4.8}$$

Furthermore, $\mathcal{R}_2(a) > -1$ and $r(\mathcal{R}_2(a)) < 1/\beta$.

4.4. Proof of Theorem 1.1. Let $x : [\alpha - h, \infty) \to \mathbb{R}$ be a solution of (3.2) and set $M = \limsup_{t\to\infty} x(t)$, $m = \liminf_{t\to\infty} x(t)$. We will reach a contradiction if we assume that $m < 0 < M$. (Note that the cases $M \le 0$ and $m \ge 0$ were already considered in Lemma 3.4.)

First suppose that $(a, \theta) \in \mathcal{S}$. By Lemmas 4.5 and 4.8, we obtain that

$$m \ \ge \ F_1(M) = \mathcal{F}_1(r(M)) \ > \ \mathcal{R}_2(r(M)) > -1. \tag{4.9}$$

Now take an arbitrary $z \ge 0$. Since $r(z) \in (a, 0]$ and $\mathcal{R}_2(z)$ is increasing on $(a, 0]$, we get $r(\mathcal{R}_2(r(z))) < 1/\beta$ due to Lemma 4.8. Therefore, the rational function $\lambda \stackrel{\text{def}}{=} \mathcal{R} \circ r \circ \mathcal{R}_2 \circ r : [0, \infty) \to [0, \infty)$ is well defined. By Lemmas 4.2 and 4.4, we obtain

$$M \le \mathcal{F}(r(m)) < \mathcal{R}(r(\mathcal{R}_2(r(M)))) = \lambda(M).$$

On the other hand, due to the inequality $\lambda'(0) = \gamma(a, \theta) < 1$ (see Lemma 4.1), we obtain that $\lambda(z) < z$ for all $z > 0$, a contradiction.

Now let $(a, \theta) \in \mathcal{D}^*$ and define the rational function $R : [0, +\infty) \to (-\infty, 0]$ as $R = \mathcal{R} \circ r$. We note that (3.1) implies $R'(0) = a\alpha(a, \theta) \in (-1, 0)$. Next,

$$m > R(M) > \mathcal{R}(a) > -1. \tag{4.10}$$

Indeed, if $M \le a(\theta - 1)/\theta - 1$, then Lemmas 4.2 and 4.4 imply that $m \ge F(M) > \mathcal{R}(r(M)) = R(M)$. If $M \ge a(\theta - 1)/\theta - 1$, then Lemmas 4.5 and 4.7 give that $m \ge F_1(M) > \mathcal{R}(r(M)) = R(M)$. The last inequality in (4.10) follows from Lemma 4.1. Finally, applying Lemmas 4.2 and 4.4, and using (4.10) and the inequality $R \circ R(x) < x$, $x > 0$, which holds since $(R \circ R)'(0) = (R'(0))^2 < 1$, we obtain that

$$M \le F(m) < \mathcal{R}(r(m)) < \mathcal{R}(r(R(M))) = R(R(M)) < M,$$

a contradiction.

To prove the second part of Theorem 1.1 take $a < 0$ and $h > 0$ such that (3.1) is not satisfied. Then, by Theorem 2.9 from [6] there is a continuous functional f satisfying (1.5) and such that the equilibrium $x(t) = 0$ in (3.2) is not locally asymptotically stable.

5. Some estimations of the global attractor for (2.3). To complete the proof of Theorem 2.1, we need to estimate the bounds of the global attractor to (2.3). We start by stating a result from [3].

LEMMA 5.1 (see [3]). *Let $q > 1$. Then there exist finite positive limits*

$$M = \limsup_{t\to\infty} x(t)\,, \ \ m = \liminf_{t\to\infty} x(t)$$

for every nonnegative solution $x(t) \not\equiv 0$ of (2.3). Moreover, $[m, M] \subseteq g([m, M])$ and $[m, M] \subseteq g_1([m, M])$, where $g_1 = \theta \ln q + (1 - \theta)g$, $\theta = \exp(-\tau)$.

Since the global stability of (2.3) for $\ln q \in (0, 2]$ was already proved in [3], we can suppose that $\ln q > 2$. In this case the minimal root x_1 of equation $g(x_1) = \ln q$ belongs to the interval $(0, 1)$. Note that $x = 1 < \ln q$ is the point of absolute maximum for g and $g_1 = (1 - \theta)g + \theta \ln q$, so that $g(1) > \ln q$ and $g_1(1) > \ln q$. We will use the information about the values of g and g_1 at $x_1 < 1 < \ln q$ in the subsequent analysis.

Now, let us consider an arbitrary solution $x(t)$ of (2.3) and its bounds m, M defined in Lemma 5.1. It is clear that if we prove the existence of $m_* = m_*(q)$ such

that $m \geq m_*(q) > x_1$ and $m_*(q)$ does not depend on $x(t)$, then the change of variables $y = x - \ln q$ transforms (2.3) into an equation satisfying (W) within the domain of attraction, and therefore Theorem 1.1 can be applied.

Since $[m, M] \subseteq g([m, M])$, we obtain immediately that either $m = M = \ln q$ or $m < \ln q < M$. In the first case the theorem is proved, so we will consider the second possibility. Next, since $z < g(z)$ for $z \in (0, \ln q)$, we have that $g(m) > m$ and

$$[m, M] \subseteq g([m, M]) \subseteq [\min\{g(m), g(M)\}, g(1)] = [g(M), g(1)].$$

Hence, $[m, M] \subseteq g([g(M), g(1)]) \subseteq [\min\{g^2(1), g^2(M)\}, g(1)]$. On the other hand, since $g(M) < \ln q$ we get analogously that $g^2(M) > g(M)$. Next, since g is decreasing on $[1, +\infty)$ and $g(1) \geq M$ we find that $g^2(1) \leq g(M)$. Thus $g^2(1) \leq g(M) < g^2(M)$ so that $\min\{g^2(1), g^2(M)\} = g^2(1)$ and $[m, M] \subseteq [g^2(1), g(1)]$. Therefore $m \geq g(g(1))$. Since the inequality $m \geq g_1(g_1(1))$ can be proved analogously, the proof of theorem will be completed if we establish that $m_*(q) = \max\{g^2(1), g_1^2(1)\} > x_1$. We have the following:

(i) $g^2(1) > x_1$ *for all* $\ln q \in [2, 2.833157]$. This is an obvious fact if $g^2(1) \geq 1$, so that we need only consider the case $x_1, g^2(1) \in (0, 1)$. Since g is increasing on $(0, 1)$, the inequality $g^2(1) > x_1$ is equivalent to $g^3(1) > g(x_1) = \ln q$ in this case. Finally, a direct computation shows that

$$g^3(1) - \ln q = q^3 e^{-1-q/e} \exp(-q^2 e^{-1-q/e}) - \ln q > 0$$

whenever $\ln q \in [2, 2.833157]$.

(ii) $g_1^2(1) > x_1$ *for all* $\ln q > 2.5$. First, let us note that $x_1 \leq \ln q + y_1$, where

$$y_1 = \left(2 - \ln q - \sqrt{(\ln q)^2 + 4 \ln q - 4}\right)/2$$

is the negative root of $\tilde{g}(y) = (y + \ln q)(1 - y + y^2/2) - \ln q$. Indeed, with $x = y + \ln q$ and $y \in (y_1, 0)$, we have that $g(x) - \ln q = qxe^{-x} - \ln q = (y + \ln q)e^{-y} - \ln q \geq \tilde{g}(y) > 0$.

Since $g_1^2(1) \geq g_1(+\infty) = \theta \ln q$, to finish the proof of (ii), it suffices to show that $\theta \ln q \geq \ln q + y_1$. Taking into account (2.2) and using the inequality $\ln(1 + x) \geq x/(1 + x)$, we obtain that

$$\begin{aligned}(\theta - 1)\ln q - y_1 &= (\theta - 1)(1 + c) - y_1 \geq (1 + c)(-1 + c \ln((c^2 + c)/(c^2 + 1))) - y_1 \\ &\geq -2 - \frac{2 - \ln q - \sqrt{(\ln q)^2 + 4 \ln q - 4}}{2} \geq 0 \quad \text{for} \quad \ln q \geq 5/2.\end{aligned}$$

6. Appendix.

6.1. Preliminary estimations.

LEMMA 6.1. *For all* $(a, \theta) \in \mathcal{D}$, *we have that* $\alpha(a, \theta) > 0$, $\beta(a, \theta) > 0$, *and*

$$T(a, \theta) \stackrel{\text{def}}{=} (a^2 - a)\beta(a, \theta)(1 - \theta) + \alpha(a, \theta) - (1 - \theta) \geq 0.$$

Proof. Since $a(\theta - 1) > 1$ for all $(a, \theta) \in \mathcal{D}$ and

$$\begin{aligned} &1 + x < \exp(x) < 1 + x + x^2/2 \quad \text{for } x < 0, \\ &\exp(x) > 1 + x + x^2/2 + x^3/6 \quad \text{for } x > 0, \end{aligned} \tag{6.1}$$

we have $\alpha(a,\theta) = (1-a)\exp(\theta/a) + a > (1-a)(1+\theta/a) + a = 1 - \theta + \theta/a > 0$. Analogously, $\beta(a,\theta) > 0$ for all $(a,\theta) \in \mathcal{D}$ because of the following chain of relations:

$$
\begin{aligned}
&- a\alpha(a,\theta)\beta(a,\theta)e^{-\theta/a} = a^2 e^{-\theta/a} - (1-a)^2 e^{\theta/a} + (1 - 2a + 2\theta(a-1)) \\
&\geq a^2\left(1 - \frac{\theta}{a} + \frac{\theta^2}{2a^2} - \frac{\theta^3}{6a^3}\right) - (1-a)^2\left(1 + \frac{\theta}{a} + \frac{\theta^2}{2a^2}\right) + (1 - 2a + 2\theta(a-1)) \\
&= \frac{\theta}{6a^2}(-3\theta - a(\theta^2 - 6\theta + 6)) > \frac{\theta(2\theta a - 2a - \theta)}{2a^2} > 0.
\end{aligned}
$$

To prove that $T(a,\theta) \geq 0$ for all $(a,\theta) \in \mathcal{D}$, we replace α, β with their values in αT:

$$
\begin{aligned}
\alpha(a,\theta)T(a,\theta) = {} & a(2a - a^2 - a\theta + a^2\theta - 1 + \theta) \\
& + 2\theta(a-1)(-2a + a\theta + 1 - \theta)\lambda - (1-a)^2(-a + a\theta - \theta)\lambda^2.
\end{aligned} \tag{6.2}
$$

It should be noticed that $-a + a\theta - \theta = a(\theta - 1) - \theta > 1 - \theta > 0$. Similarly, $2\theta(a-1)(-2a + a\theta + 1 - \theta) < 0$ so that $T(a,\theta) \geq 0$ if

$$
\theta(-a + a\theta - \theta)(4a^4)^{-1}[-\theta(-\theta + a\theta - 2a)^2 + 4(\theta - 1)a^3] > 0,
$$

where the last expression was obtained from (6.2) by replacing $\lambda = \exp(\theta/a)$ by $1 + \theta/a + \theta^2/2a^2 > \exp(\theta/a)$. Taking into account that $a(\theta-1) > 1$ and $\theta < 1$ for $(a,\theta) \in \mathcal{D}$, we end the proof of this lemma by noting that $4(\theta-1)a^3 - \theta(\theta - a\theta + 2a)^2 \geq 4a^2 - (\theta - a\theta + 2a)^2 = (-4a - \theta + a\theta)\theta(1-a) > 0$. □

LEMMA 6.2. *For all $(a,\theta) \in \mathcal{D}$ one has*

$$
\frac{a\alpha(a,\theta)}{1 - a\beta(a,\theta)} > -1. \tag{6.3}
$$

Proof. It follows directly from the definitions of $\alpha(a,\theta)$ and $\mathcal{D}$ that $a\alpha(a,\theta) > -1$ for all $(a,\theta) \in \mathcal{D}$. Now, (6.3) follows from the fact that $a\beta(a,\theta) < 0$ if $(a,\theta) \in \mathcal{D}$. □

LEMMA 6.3. *If $(a,\theta) \in \mathcal{S}$, then $\gamma(a,\theta) < 1$.*

Proof. Notice that $(a,\theta) \in \mathcal{S}$ implies that $\theta \in [0.8, 1)$ and $\theta > \Pi_3(-1/a)$ (or, equivalently, $a > \pi_3(\theta) \stackrel{\text{def}}{=} -1/\Pi_3^{-1}(\theta) = (108 + 25\theta)(95\theta - 95)^{-1}$). Here Π_3^{-1} is the inverse function of Π_3. Next we prove the inequality

$$
\frac{1 - \theta + \ln\theta}{2 - \theta + \ln\theta} + \frac{(\theta-1)^2(2-\theta)}{2} > 0, \quad \theta \in [0.8, 1), \tag{6.4}
$$

which is equivalent to the relation

$$
\Xi(q) \stackrel{\text{def}}{=} q(q-2)(q^2+1) + (2 + q^2 - q^3)\ln(q+1) > 0, \quad q \in [-0.2, 0),
$$

with $\theta = q + 1$ (note that $1 - q + \ln(q+1) > 0$ for $q \in [-0.2, 0)$). To do that, we will need the following approximation of $\ln(1+q)$ when $q \in [-0.2, 0)$:

$$
\ln(1+q) > q - 0.5q^2 + 0.4q^3, \quad q \in [-0.2, 0). \tag{6.5}
$$

(Indeed, function $y(x) = \ln(1+x) - (x - 0.5x^2 + 0.4x^3)$ has exactly one critical point $x = -1/6$ on $[-0.2, 0)$, and $y(-0.2) = 0.00005644\ldots > 0$, $y(0) = 0$.) Inequality (6.5) implies that $\Xi(q) \geq -0.1q^3(5q + 2 - 9q^2 + 4q^3)$. Now, since $(5q + 2 - 9q^2 + 4q^3) > 0$ for all $q \in [-0.2, 0)$, we have that $\Xi(q) > 0$, and thus (6.4) is proved.

Next, due to (6.1) and (6.4),

$$\gamma(a,\theta) \le a^3\left(1-\theta+\frac{\theta}{a}-\frac{\theta^2}{2a}+\frac{\theta^2}{2a^2}\right)\frac{(\theta-1)^2(\theta-2)}{2} = w(a,\theta),$$

so that $\gamma(\pi_3(\theta),\theta) \le w(\pi_3(\theta),\theta)$. Now, $w(\pi_3(\theta),\theta)$ is a fifth degree polynomial, and an elementary analysis shows that $w(\pi_3(\theta),\theta) < 1$ for all $\theta \in [0.8, 1)$. Since

$$\partial w(a,\theta)/\partial a = 0.25(2-\theta)(\theta-1)^2[a(\theta-1)(2a+2\theta)+a(4a(\theta-1)-2\theta)] < 0,$$

we conclude that $\gamma(a,\theta) \le w(a,\theta) < 1$ for $a > \pi_3(\theta)$. □

LEMMA 6.4. *Let* $(a,\theta) \in \mathcal{D}$ *and* $r > -1/4$. *Set* $\mathcal{J}(r) = \mathcal{I}(N(r))$, *where* $\mathcal{I}(N) = N\coth(\nu N/2)$, $\nu = -\theta/a$, $N(r) = \sqrt{1+4r}$. *Then* $\mathcal{J}'(0) > 0$ *and*

$$\mathcal{J}(r) \le \mathcal{J}(0) + \mathcal{J}'(0)r = \frac{1+\lambda}{1-\lambda} + \left(2\frac{1+\lambda}{1-\lambda} + \frac{4\theta\lambda}{a(1-\lambda)^2}\right)r. \tag{6.6}$$

Proof. Set $k(N) = e^{2\nu N} - 2\nu N e^{\nu N} - 1 > 0$; then $k(0) = 0$ and $k'(N) = 2\nu e^{\nu N}[e^{\nu N} - 1 - \nu N] > 0$ for all $N > 0$. Hence $\mathcal{I}'(N) = k(N)/(e^{\nu N}-1)^2 > 0$ and $\mathcal{J}'(0) = \mathcal{I}'(1)N'(0) > 0$.

Next, since $dN/dr = 2(1+4r)^{-1/2} = 2/N$, $d^2N/dr^2 = -4/N^3$, we obtain that

$$\begin{aligned}\mathcal{J}''(r) &= \frac{\partial^2\mathcal{I}(N(r))}{\partial r^2} = \frac{\partial^2\mathcal{I}(N)}{\partial N^2}\left(\frac{\partial N(r)}{\partial r}\right)^2 + \frac{\partial\mathcal{I}(N)}{\partial N}\left(\frac{\partial^2 N(r)}{\partial r^2}\right)\\ &= \frac{4[-e^{3\nu N} + e^{2\nu N}(2\nu^2N^2 - 2\nu N + 1) + e^{\nu N}(2\nu^2N^2 + 2\nu N + 1) - 1]}{N^3(e^{\nu N}-1)^3}\\ &= \frac{\sum_{j=0}^{+\infty} p_j(\nu N)^j}{N^3(e^{\nu N}-1)^3} < 0, \quad \text{since } (\nu N) > 0,\ p_j = 0,\ j = 0,\dots,5, \text{ and}\end{aligned}$$

$$p_j = \frac{4}{(j-2)!}\left(\frac{-3^j + 2^j + 1}{j(j-1)} + \frac{-2^j+2}{j-1} + 2^{j-1} + 2\right) < 0,\ j \ge 6.$$

Thus $\mathcal{J}(r) \le \mathcal{J}(0) + \mathcal{J}'(0)r$ and (6.6) is proved. □

6.2. Properties of function *F*. To study the properties of functions $F : I \to \mathbb{R}$ and $\mathcal{F} : J = (a_*, +\infty) \to \mathbb{R}$, defined in subsection 4.2, it will be more convenient to use the integral representation (4.3) instead of the original definition of F. It should be noted that conditions $xF(x) < 0$, $(F(x) - r(x))x > 0$, $x \in I \setminus \{0\}$, define F in a unique way: moreover F and $\mathcal{F}$ are continuous and smooth at 0 with $\mathcal{F}'(0) = \alpha(a,\theta)$, $\mathcal{F}''(0) = 2\alpha(a,\theta)\beta(a,\theta)$. We have taken into consideration these facts to define the rational functions R and $\mathcal{R}$ (see subsection 4.2); however, since we do not use these characteristics of F anywhere, their proof is omitted here.

LEMMA 6.5. *Assume that* $r \in [a_*, 0]$. *Then*

$$\mathcal{F}(r) > \alpha\, r/(1-\beta r) = \mathcal{R}(r). \tag{6.7}$$

Proof. 1. First, suppose that $4r + 1 > 0$ and $r \ne 0$. Since $0 > \mathcal{F}(r) > r$, we have, for every $z \in [r, \mathcal{F}(r)]$ and $a < 0$,

$$r^{-1}(z) = z/(a-z) = (z/a)(1 + (z/a) + (z/a)^2 + \cdots) > (z/a) + (z/a)^2. \tag{6.8}$$

Hence

$$\theta = \int_r^{\mathcal{F}(r)} \frac{dz}{z(a-z)^{-1} - r} < \int_r^{\mathcal{F}(r)} \frac{dz}{\frac{z}{a} + (\frac{z}{a})^2 - r} = a \int_{r/a}^{\mathcal{F}(r)/a} \frac{du}{u + u^2 - r}. \tag{6.9}$$

Now, since for $r < 0$ the roots $\alpha_1 = (-1 - \sqrt{1+4r})/2$, $\alpha_2 = (-1 + \sqrt{1+4r})/2$ of the equation $u^2 + u - r = 0$ are negative, we obtain

$$a \int_{r/a}^{\mathcal{F}(r)/a} \frac{du}{u + u^2 - r} = -\frac{a}{\sqrt{1+4r}} \ln \left(\frac{\mathcal{F}(r) - a\alpha_1}{\mathcal{F}(r) - a\alpha_2} \, \frac{r - a\alpha_2}{r - a\alpha_1} \right) > \theta. \tag{6.10}$$

The last inequality implies that

$$\frac{\mathcal{F}(r) - a\alpha_1}{\mathcal{F}(r) - a\alpha_2} \, \frac{r - a\alpha_2}{r - a\alpha_1} > \exp\left(\frac{\theta\sqrt{1+4r}}{-a} \right) \stackrel{\text{def}}{=} \omega(r) \geq 1. \tag{6.11}$$

Taking into account that $\mathcal{F}(r) - a\alpha_2 < 0$ and $r - a\alpha_1 < 0$, and replacing α_1, α_2 in (6.11) by their values, we obtain

$$\mathcal{F}(r) > \frac{r(2a^2 - a + a\frac{\omega(r)+1}{\omega(r)-1}\sqrt{1+4r})}{2r + a + a\frac{\omega(r)+1}{\omega(r)-1}\sqrt{1+4r}}.$$

Next, since $\mathcal{J}(r) = \frac{\omega(r)+1}{\omega(r)-1}\sqrt{1+4r}$, we can apply Lemma 6.4 to see that

$$\begin{aligned} \mathcal{F}(r) &> \frac{r(2a - 1 + \mathcal{J}(r))}{2r/a + 1 + \mathcal{J}(r)} \geq \frac{r(2a - 1 + \mathcal{J}(0) + \mathcal{J}'(0)r)}{2r/a + 1 + \mathcal{J}(0) + \mathcal{J}'(0)r} \\ &= \frac{r(\lambda + a(1-\lambda)) + \frac{1}{2}\mathcal{J}'(0)(1-\lambda)r^2}{1 + (\frac{1-\lambda}{a} + \frac{1}{2}\mathcal{J}'(0)(1-\lambda))r} \stackrel{\text{def}}{=} \mathcal{L}(r). \end{aligned} \tag{6.12}$$

Now, $\mathcal{L}(r) = (a_1 r + a_2 r^2)/(1 + a_3 r)$, with $a_1 > 0, a_2 > 0$. Moreover, since $0 < \mathcal{J}(r) \leq \mathcal{J}(0) + \mathcal{J}'(0)r$, all denominators in (6.12) are positive so that $1 + a_3 r > 0$. Next,

$$a_1 a_3 - a_2 = (\lambda + a(1-\lambda))\left(\frac{1-\lambda}{a} + \frac{1}{2}\mathcal{J}'(0)(1-\lambda) \right) - \frac{1}{2}\mathcal{J}'(0)(1-\lambda) \leq 0. \tag{6.13}$$

Indeed, the last inequality is equivalent to the obvious relation

$$\frac{\lambda + a(1-\lambda)}{a} < 0 \leq \frac{1}{2}(1-\lambda)(1-a)\mathcal{J}'(0).$$

(Notice that $\alpha = \lambda + a(1-\lambda) > 0$, while, by Lemma 6.4, $\mathcal{J}'(0) > 0$.)

Finally, since the inequality $(a_1 r + a_2 r^2)(1 + a_3 r)^{-1} \geq a_1 r(1 + (a_3 - a_2/a_1)r)^{-1}$ holds for $r < 0,\ a_1 > 0,\ a_2 > 0,\ 1 + a_3 r > 0,\ a_1 a_3 \leq a_2$, we obtain

$$\mathcal{F}(r) > \frac{a_1 r}{1 + (a_3 - \frac{a_2}{a_1})r} = \frac{r(\lambda + a(1-\lambda))}{1 + (\frac{1-\lambda}{a} + \mathcal{J}'(0)\frac{1-\lambda}{2} - \frac{\mathcal{J}'(0)(1-\lambda)}{2(\lambda + a(1-\lambda))})r} = \mathcal{R}(r).$$

Hence the statement of the lemma is proved for $r \in (-1/4, 0)$. As an important consequence of the first part of the proof, we get the relation

$$\lim_{r \to -1/4} \frac{r(2a^2 - a + a\frac{\omega(r)+1}{\omega(r)-1}\sqrt{1+4r})}{2r + a + a\frac{\omega(r)+1}{\omega(r)-1}\sqrt{1+4r}} = \frac{a^2(1-\theta) + \theta a/2}{\theta(2a-1) - 4a^2} \geq \mathcal{R}(-1/4),$$

which will be used in the next stage of proof.

2. The case $r = -1/4$. From (6.9), evaluated at $r = -1/4$, we get $\theta < (-2a^2)/(2\mathcal{F}(-1/4) + a) + (4a^2)/(2a-1)$, so that

$$\mathcal{F}(-1/4) > \frac{a^2(1-\theta) + \theta a/2}{\theta(2a-1) - 4a^2} \geq \mathcal{R}(-1/4).$$

3. Assume now that $4r + 1 < 0$. We have

$$a\int_{r/a}^{\mathcal{F}(r)/a} \frac{du}{u + u^2 - r} = \frac{2a}{\sqrt{-4r-1}}\left(\arctan\frac{2\mathcal{F}(r)+a}{a\sqrt{-4r-1}} - \arctan\frac{2r+a}{a\sqrt{-4r-1}}\right). \tag{6.14}$$

By (6.9) and (6.14), we obtain

$$2\mathcal{F}(r) + a > a\sqrt{-4r-1}\,\frac{\frac{2r+a}{a\sqrt{-4r-1}} + \tan\frac{\theta\sqrt{-4r-1}}{2a}}{1 - \frac{2r+a}{a\sqrt{-4r-1}}\tan\frac{\theta\sqrt{-4r-1}}{2a}}.$$

Now, since $\tan x \leq x + x^3/3$ for $x \in (-\pi/2, 0)$ and $a < 0$, we obtain

$$\mathcal{F}(r) > r\frac{a^2(1-\theta) + \theta a/2 + \theta^3(-r-\frac{1}{4})(\frac{1}{2a}-1)(3)^{-1}}{a^2 - \theta(r+\frac{a}{2}) - \theta^3(-r-\frac{1}{4})(\frac{a}{2}+r)(3a^2)^{-1}} = G(r). \tag{6.15}$$

Therefore it will be sufficient to establish that $G(r) \geq \mathcal{R}(r)$ for $r < -1/4$. First, note that by the second part of the proof

$$G(-1/4) = \frac{a^2(1-\theta) + \theta a/2}{\theta(2a-1) - 4a^2} \geq \mathcal{R}(-1/4).$$

Let us consider now the function $H(r) = G(r) - \mathcal{R}(r)$ for $r \leq 0$. Since $G(r) = G_1(r)/G_2(r)$, where G_j are polynomials in r of second degree, $H(r)$ can be written as

$$H(r) = \frac{G_1(r)(1-\beta r) - \alpha r G_2(r)}{G_2(r)(1-\beta r)} = \frac{H_1(r)}{H_2(r)}, \tag{6.16}$$

so that H is a quotient of two polynomials of third degree with $H_2(r) > 0$ for $r \leq 0$. We get $\lim_{r\to-\infty} G(r) = a^2(1 - 1/(2a)) > 0$, and therefore $H(-\infty) = \lim_{r\to-\infty} H(r) > 0$. Furthermore, $H(0) = 0$ and

$$H'(0) = \frac{1 - \theta(1-\frac{1}{2a}) + \frac{\theta^3}{12a^2}(1-\frac{1}{2a})}{1 - \frac{\theta}{2a} + \frac{\theta^3}{24a^3}} - (a + e^{\frac{\theta}{a}}(1-a)) = \frac{\sum_{k=5}^{+\infty} p_k\theta^k}{1 - \frac{\theta}{2a} + \frac{\theta^3}{24a^3}} > 0,$$

since the denominator of the last fraction is positive and $p_{2m+1} > 0$, $p_{2m} < 0$, $p_{2m+1} + p_{2m+2} > 0$ for $m \geq 2$. Here we use the formula

$$p_k = \frac{a-1}{a^k k!}\left(1 - \frac{k}{2} + \frac{k(k-1)(k-2)}{24}\right),\ k \geq 5.$$

Finally, since $H(-1/4) = G(-1/4) - \mathcal{R}(-1/4) \geq 0$, there exists at least one zero of $H(r)$ in the interval $[-1/4, 0)$. $H_1(r)$ is a polynomial of third degree in r, and therefore it cannot have more than three zeros. Hence, since $H(-\infty) > 0$ and $H(-1/4) \geq 0$, we obtain that $H(r) \geq 0$ if $r < -1/4$. □

LEMMA 6.6. *If $(a, \theta) \in \mathcal{D}$, then $\mathcal{F}(r) < \mathcal{R}(r)$ for all $r \in (0, 1/\beta)$.*

Proof. By definition of $\mathcal{F}$, we have that $z > 0$ if $r > 0$ and $z \in [\mathcal{F}(r), r]$. We begin the proof by assuming that $r \in (0, a^2 - a)$. Since $z > 0, a < 0$, we find that $z(a-z)^{-1} < z/a + z^2/a^2$. Therefore, (6.9) holds under our present conditions. Now, since $r > 0$, the roots of equation $u^2 + u - r = 0$ are $\alpha_1 = (-1 - \sqrt{1+4r})/2 < 0$, $\alpha_2 = (-1+\sqrt{1+4r})/2 > 0$. Next, since $\mathcal{F}(r) - a\alpha_1 < r - a\alpha_1 < 0$ for all $r \in (0, a^2-a)$, we obtain that the relations (6.10), (6.11) hold in the new situation, and therefore

$$\mathcal{F}(r) < \frac{a^2\alpha_1\alpha_2(\omega(r)-1) + ar(\alpha_1 - \omega(r)\alpha_2)}{-a\alpha_2 + r + \omega(r)(a\alpha_1 - r)} = \frac{r(2a^2 - a + a\frac{\omega(r)+1}{\omega(r)-1}\sqrt{1+4r})}{2r + a + a\frac{\omega(r)+1}{\omega(r)-1}\sqrt{1+4r}},$$

where the denominator is positive for every $r \in (0, a^2 - a)$. Now, recall that $\mathcal{J}(r) = \frac{\omega(r)+1}{\omega(r)-1}\sqrt{1+4r}$; applying Lemma 6.4, we obtain $\mathcal{J}(r) \le \mathcal{J}(0) + \mathcal{J}'(0)r$. Next, since for all $r \in (0, a^2 - a)$ we have that $2r/a + 1 + \mathcal{J}(r) = -2(a(\omega(r)-1))^{-1}(-a\alpha_2 + r + \omega(r)(a\alpha_1 - r)) > 0$, and the function $p(x) = (r(2a-1+x))(2r/a+1+x)^{-1}$ is increasing in x, we get $\mathcal{F}(r) < \mathcal{L}(r)$ (compare with (6.12)). Now, $(a_1 r + a_2 r^2)(1 + a_3 r)^{-1} \le a_1 r(1 + (a_3 - a_2/a_1)r)^{-1}$ if $a_1 a_3 - a_2 \le 0,\ r > 0,\ a_1 > 0,\ a_2 > 0$. Therefore, by (6.13), $\mathcal{L}(r) \le \mathcal{R}(r)$.

Now we assume that $r \ge a^2 - a$. Taking into account that $z(a-z)^{-1} < 0$ for $z > 0$, we obtain the inequality

$$\theta = \int_r^{\mathcal{F}(r)} \frac{dz}{z(a-z)^{-1} - r} < \int_r^{\mathcal{F}(r)} \frac{dz}{-r} = \frac{\mathcal{F}(r) - r}{-r},$$

so that $\mathcal{F}(r) < r(1-\theta)$. Finally, the inequality $r(1-\theta) \le \mathcal{R}(r) = \alpha r/(1-\beta r)$ is equivalent to $r \ge (1-\theta-\alpha)/((1-\theta)\beta)$, which holds for all $r \ge a^2 - a$ due to the relation $a^2 - a \ge (1-\theta-\alpha)/((1-\theta)\beta)$, established in Lemma 6.1. □

6.3. Properties of function F_1 in the domain $\mathcal{D}^*$. Suppose now that $(a, \theta) \in \mathcal{D}^*$. We study some properties of function F_1 and the associated function $\mathcal{F}_1 : (a, 0) \to \mathbb{R}$ defined as $\mathcal{F}_1(r(z)) = F_1(z)$, which, by Lemma 4.6, satisfies

$$\frac{r_1(r)\theta}{r} = \int_{r_1(r)}^{\mathcal{F}_1(r)} \frac{dz}{r^{-1}(z) - r}, \quad \text{where } r_1(r) = \frac{ar(\theta-1)}{\theta + r(\theta-1)}.$$

LEMMA 6.7. *Assume that $(a, \theta) \in \mathcal{D}^*$ and that the inequalities $a < r \le a_* = a + \theta/(1-\theta)$ hold. Then $\mathcal{F}_1(r) > \mathcal{R}(r)$.*

Proof. Since $r_1(r) < \mathcal{F}_1(r) < 0$ and $a_* < -1$ for $(a, \theta) \in \mathcal{D}^*$, using (6.8) we get

$$\frac{r_1(r)\theta}{r} < \int_{r_1(r)}^{\mathcal{F}_1(r)} \frac{dz}{\frac{z}{a} + (\frac{z}{a})^2 - r} = a\int_{r_1(r)/a}^{\mathcal{F}_1(r)/a} \frac{du}{u + u^2 - r}.$$

The last integral can be transformed as in (6.14) to obtain

$$\frac{r_1(r)\theta}{r} < \frac{2a}{\sqrt{-4r-1}}\left(\arctan\frac{2\mathcal{F}_1(r)+a}{a\sqrt{-4r-1}} - \arctan\frac{2r_1+a}{a\sqrt{-4r-1}}\right).$$

Therefore

$$\varsigma_1 \stackrel{\text{def}}{=} \arctan\frac{2\mathcal{F}_1(r)+a}{a\sqrt{-4r-1}} < \arctan\frac{2r_1+a}{a\sqrt{-4r-1}} + \frac{\theta r_1\sqrt{-4r-1}}{2ar} \stackrel{\text{def}}{=} \varsigma_2 + \varsigma_3,$$

and since $\varsigma_1, \varsigma_2 \in (0, \pi/2), \varsigma_3 < 0$, we obtain

$$2\mathcal{F}_1(r) + a > a\sqrt{-4r-1}\,\frac{\frac{2r_1+a}{a\sqrt{-4r-1}} + \tan\frac{\theta r_1\sqrt{-4r-1}}{2ar}}{1 - \frac{2r_1+a}{a\sqrt{-4r-1}}\tan\frac{\theta r_1\sqrt{-4r-1}}{2ar}}.$$

Since $\tan x < x + x^3/3$ for $x \in (-\pi/2, 0)$, we have

$$\mathcal{F}_1(r) > \frac{A_1(P)r + A_2(P)r^2}{B_0(P) + B_1(P)r + B_2(P)r^2} = G_1(r, P, a, \theta), \tag{6.17}$$

where

$$A_1(P) = (1-\theta)P + \frac{\theta}{2a}P^2 + \frac{\theta^3}{24a^3}(2a-P)P^3, \quad A_2(P) = \frac{\theta^3}{6a^3}(2a-P)P^3,$$
$$B_0(P) = 1 - \frac{\theta P}{2a} + \frac{\theta^3P^3}{24a^3}, \quad B_1(P) = -\frac{\theta P^2}{a^2} + \frac{\theta^3P^3}{6a^3} + \frac{\theta^3P^4}{12a^4}, \quad B_2(P) = \frac{\theta^3P^4}{3a^4},$$
$$P = P(r, a, \theta) = r_1/r = \frac{a(\theta-1)}{\theta + r(\theta-1)}.$$

After substitution of the value of P into (6.17), we get

$$\mathcal{F}_1(r) > G_1(r, P(r, a, \theta), a, \theta) \stackrel{\text{def}}{=} \mathcal{G}_1(r, a, \theta) = \frac{rM(r, a, \theta)}{N(r, a, \theta)},$$

where

$$\begin{aligned} M(r, a, \theta) &= 24(A_1(P) + A_2(P)r)(\theta + r(\theta-1))^4 \\ &= -(\theta-1)^2 a[13\theta^3 - \theta^5 - 2\theta^2(\theta-1)(\theta+3)(3\theta-8)r \\ &\quad -4\theta(2\theta^2-15)(\theta-1)^2r^2 + 24(\theta-1)^3r^3], \\ N(r, a, \theta) &= 24(B_0(P) + B_1(P)r + B_2(P)r^2)(\theta + r(\theta-1))^4 = 35\theta^4 - 9\theta^5 \\ &\quad +\theta^7 - 3\theta^6 + \theta^3(\theta-1)(7\theta^3 - 17\theta^2 - 47\theta + 153)r + 12\theta^2(\theta^3 - 2\theta^2 \\ &\quad -6\theta + 19)(\theta-1)^2r^2 - 12\theta(3\theta-11)(\theta-1)^3r^3 + 24(\theta-1)^4r^4. \end{aligned}$$

To prove our lemma, it suffices to check the inequality $\mathcal{G}_1(r, a, \theta) \geq \mathcal{R}(r)$ for $r \in [a, a_*]$. First, considering $N(r, a, \theta) = N(r, \theta)$ as a polynomial in r of the form $N(r, a, \theta) = \sum_{k=0}^{4} N_k(\theta)r^k$, we can check that $(-1)^k N_k(\theta) > 0$ for $\theta \in (0, 1)$, and therefore, for all $\theta \in (0, 1)$ and $r < 0$,

$$N(r, a, \theta) = 24(B_0(P) + B_1(P)r + B_2(P)r^2)(\theta + r(\theta-1))^4 > 0. \tag{6.18}$$

Since $N(r, a, \theta) > 0$, $1 - \beta r > 0$ (recall that $\beta(a, \theta) > 0$ in the domain $\mathcal{D}$), the inequality $rM(r, a, \theta)/N(r, a, \theta) \geq \alpha r/(1 - r\beta)$ is equivalent to

$$Q(r, a, \theta) \stackrel{\text{def}}{=} (1 - r\beta(a, \theta))M(r, a, \theta) - \alpha(a, \theta)N(r, a, \theta) \leq 0. \tag{6.19}$$

Now, an easy comparison of $\mathcal{G}_1(a_*, a, \theta) = G_1(a_*, 1, a, \theta)$ with $G(a_*)$ given in (6.15) shows that the inequality (6.19) is fulfilled for $r = a_*$. In the next two lemmas, we will prove that $\partial Q(r, a, \theta)/\partial r > 0$ for all $r \in [a, a_*]$. Therefore, since $Q(a_*, a, \theta) \leq 0$, we obtain $Q(r, a, \theta) \leq 0$ for $r \in [a, a_*]$, which proves that $\mathcal{F}_1(r) > \mathcal{R}(r)$. □

LEMMA 6.8. *$S(r,a,\theta)=\frac{\partial}{\partial r}Q(r,a,\theta)>0$ at the point $r=a_*$.*

Proof. Recall that we are interested in the case $r=a_*$ (when $P=1$). By (6.19) and the above definitions of $M(r,a,\theta), N(r,a,\theta)$,

$$Q(r,a,\theta)=24(\theta+r(\theta-1))^4((A_1(P)+A_2(P)r)(1-\beta r)-\alpha(B_0(P)+B_1(P)r+B_2(P)r^2)).$$

Next, setting $P'=\partial P(r,a,\theta)/\partial r\mid_{r=a_*}=-a^{-1}$, $A_j'=\partial A_j(a,\theta,P)/\partial P|_{P=1}$, $B_j'=\partial B_j(a,\theta,P)/\partial P|_{P=1}$, $A_j=A_j(a,\theta,1), B_j=B_j(a,\theta,1)$, we obtain that

$$\partial Q(r,a,\theta)/\partial r|_{r=a_*}=24(Q_1(r,a,\theta)+Q_2(r,a,\theta)), \tag{6.20}$$

where

$$\begin{aligned}Q_1&=4a^3(\theta-1)^4((A_1+A_2a_*)(1-\beta a_*)-\alpha(B_0+B_1a_*+B_2a_*^2))\\&\quad+a^4(\theta-1)^4((A_1'+A_2'a_*)(1-\beta a_*)-\alpha(B_0'+B_1'a_*+B_2'a_*^2))P';\\Q_2&=a^4(\theta-1)^4(A_2-\beta A_1-\alpha B_1-2a_*\beta A_2-2a_*\alpha B_2).\end{aligned}$$

Now, for the convenience of the reader, the following part of the proof will be divided into several steps.

Step (i): $Q_2(r,a,\theta)>0$. Indeed, consider the second degree polynomial

$$\chi_1(r)\stackrel{\text{def}}{=}(A_1+A_2r)(1-\beta r)-\alpha(B_0+B_1r+B_2r^2).$$

Notice that $\chi_1(r)=\frac{H_1(r)}{r}$, where H_1 is defined in (6.16). This implies that the unique critical point of χ_1 belongs to $(-1/4,+\infty)$ and that $\chi_1(+\infty)=-\infty$. Hence $\chi_1'(r)>0$ for all $r<-1/4$ so that $Q_2(\theta,a,r)=a^4(\theta-1)^4\chi_1'(r)>0$.

Step (ii). The following inequality holds:

$$(A_1'+A_2'a_*)(B_0+B_1a_*+B_2a_*^2)-(A_1+A_2a_*)(B_0'+B_1'a_*+B_2'a_*^2)>0. \tag{6.21}$$

Indeed, the left-hand side of (6.21) can be transformed into

$$\begin{aligned}&\frac{1}{576a^6(1-\theta)^3}(-\theta^6(3\theta+1)^3+12\theta^6(3\theta+1)^2(\theta-1)a\\&-24\theta^4(\theta-1)(3\theta+1)(2\theta^3-2\theta^2+3\theta-5)a^2+32\theta^3(2\theta^4-2\theta^3+18\theta^2\\&-39\theta-9)(\theta-1)^2a^3-48\theta^2(8\theta^3-41\theta^2+30\theta-9)(\theta-1)^2a^4\\&-576\theta(\theta^2-\theta+2)(\theta-1)^3a^5+576(\theta-1)^4a^6).\end{aligned} \tag{6.22}$$

Taking into account that $\eta\stackrel{\text{def}}{=}(\theta-1)a>1$, the sum of the first two terms in (6.22) is positive:

$$-\theta^6(3\theta+1)^3+12\theta^6(3\theta+1)^2(\theta-1)a=\theta^6(3\theta+1)^2(-(3\theta+1)+12(\theta-1)a)>0.$$

The other terms in (6.22) can be written as

$$\begin{aligned}&a^2(-24\theta^4(\theta-1)(3\theta+1)(2\theta^3-2\theta^2+3\theta-5)\\&+32\theta^3(2\theta^4-2\theta^3+18\theta^2-39\theta-9)(\theta-1)\eta\\&-48\theta^2(8\theta^3-41\theta^2+30\theta-9)\eta^2-576\theta(\theta^2-\theta+2)\eta^3+576\eta^4)\stackrel{\text{def}}{=}a^2\Upsilon(\theta,\eta).\end{aligned}$$

By the Taylor formula,

$$\Upsilon(\theta,\eta)=\Upsilon(\theta,1)+(\eta-1)\partial\Upsilon(\theta,1)/\partial\eta+0.5(\eta-1)^2\partial^2\Upsilon(\theta,\eta_1)/\partial\eta^2, \tag{6.23}$$

where $\eta_1 \in [1,2]$. It is easy to verify that

$$\begin{aligned}\Upsilon(\theta,1) &= -8(\theta-1)^2[\theta^4(18\theta^3-2\theta^2+27\theta-81)+(108\theta^3-54\theta^2-72)] > 0,\\ \partial\Upsilon(\theta,1)/\partial\eta &= 32(\theta-1)[(2\theta^6+18\theta^4+36)(\theta-1)-45\theta^2(\theta-1)^2-36] > 0,\\ \partial^2\Upsilon(\theta,\eta)/\partial\eta^2 &= 3456\eta(2\eta-\theta(\theta^2-\theta+2))-96\theta^2(8\theta^3-41\theta^2+30\theta-9) > 0.\end{aligned}$$

(Here we use the inequality $8\theta^3-41\theta^2+30\theta-9<0,\ \theta\in[0,1]$.)

Finally, by (6.23), $\Upsilon(\theta,\eta)>0$ for $\theta\in(0,1), \eta\in(1,2)$.

Step (iii). We have

$$\varrho \stackrel{\text{def}}{=} (B_0'+B_1'a_*+B_2'a_*^2)(B_0+B_1a_*+B_2a_*^2)^{-1} < 1. \tag{6.24}$$

Indeed, taking into account (6.18), the latter inequality is equivalent to

$$(\upsilon \stackrel{\text{def}}{=}) \qquad \frac{\theta^3}{12a^3}+a_*\left(-\frac{\theta}{a^2}+\frac{\theta^3}{3a^3}+\frac{\theta^3}{4a^4}\right)+\frac{\theta^3}{a^4}a_*^2<1.$$

Now, we know that $|a_*|\le|a|$ and $|a^{-1}|<1-\theta$ (so that $\theta/|a|<1/4$). Therefore $\upsilon<1/4+(1/3)(1/16)+1/16<1$.

Step (iv): $Q_1(r,a,\theta)>0$. First, using (6.18) and (6.21), we obtain that

$$\begin{aligned}&(A_1'+A_2'a_*)(1-\beta a_*)-\alpha(B_0'+B_1'a_*+B_2'a_*^2)\\ &\ge \varrho((A_1+A_2a_*)(1-\beta a_*)-\alpha(B_0+B_1a_*+B_2a_*^2)).\end{aligned}$$

Next, using inequality (6.19), which was proved at $r=a_*$, we find that

$$(A_1+A_2a_*)(1-\beta a_*)-\alpha(B_0+B_1a_*+B_2a_*^2)=\frac{Q(a_*,a,\theta)}{24(\theta+a_*(\theta-1))^4}<0.$$

Therefore,

$$Q_1(r,a,\theta)\ge a^3(\theta-1)^4((A_1+A_2a_*)(1-\beta a_*)-\alpha(B_0+B_1a_*+B_2a_*^2))(4-\varrho)>0.$$

Step (v). Recalling (6.20) and Steps (i) and (iv), we finish the proof of the lemma. □

LEMMA 6.9. *$S(r,a,\theta)>0$ for $r\in[a,a_*]$.*

Proof. Differentiating function Q given by (6.19), we obtain

$$\begin{aligned}S(r,a,\theta)=\sum_{i=0}^{3}S_i(\theta,a)r^i &= 96(\theta-1)^4(\beta a(\theta-1)-\alpha)r^3\\ &+(-12(\theta-1)^4a\theta(2\theta^2-15)\beta+36\theta(\theta-1)^3(3\theta-11)\alpha-72(\theta-1)^5a)r^2\\ &+(-4a\theta^2(\theta-1)^3(\theta+3)(3\theta-8)\beta-24\theta^2(\theta-1)^2(\theta^3-2\theta^2-6\theta+19)\alpha\\ &+8a\theta(\theta-1)^4(2\theta^2-15))r\\ &+a(\theta-1)^2\theta^3(13-\theta^2)\beta-\theta^3(\theta-1)(7\theta^3-17\theta^2-47\theta+153)\alpha\\ &+2a\theta^2(\theta-1)^3(\theta+3)(3\theta-8).\end{aligned}$$

Now, inequalities $S_i(a,\theta)r^i<0,\ i=3,2,1,0$, are equivalent to $a\theta(\theta-1)\beta>T_i(a,\theta)$, where

$$T_3(a,\theta)=\theta\alpha,\quad T_2(a,\theta)=\frac{-6a(\theta-1)^2+3\theta(3\theta-11)\alpha}{2\theta^2-15},$$

$$T_1(a,\theta) = \frac{2a(\theta-1)^2(2\theta^2-15) - 6\theta(\theta^3-2\theta^2-6\theta+19)\alpha}{3\theta^2+\theta-24},$$

$$T_0(a,\theta) = \frac{2a(\theta-1)^2(3\theta^2+\theta-24) - \theta(7\theta^3-17\theta^2-47\theta+153)\alpha}{\theta^2-13}.$$

Next, for $(a,\theta) \in \mathcal{D}^*$, the following inequalities hold:

$$T_3(a,\theta) > T_2(a,\theta), \tag{6.25}$$

$$T_2(a,\theta) > T_1(a,\theta), \tag{6.26}$$

$$T_1(a,\theta) > T_0(a,\theta). \tag{6.27}$$

Indeed, taking into account that $\alpha = (1-a)\exp(\theta/a)+a$, inequality (6.26) is equivalent to

$$\begin{aligned} &4\theta^6 - 8\theta^5 - 41\theta^4 + 108\theta^3 + 99\theta^2 - 312\theta - 306 \\ &+ 3\theta(4\theta^5 - 8\theta^4 - 45\theta^3 + 106\theta^2 + 97\theta - 306)e^{\theta/a}(1-a)a^{-1} < 0. \end{aligned} \tag{6.28}$$

Since $3\theta(4\theta^5-8\theta^4-45\theta^3+106\theta^2+97\theta-306) < 0$, it is sufficient to prove (6.28) for the maximum value in a of the function $\frac{1-a}{-a}e^{\frac{\theta}{a}}$. The derivative of this function is equal to $-e^{\theta/a}(a\theta-a-\theta)/a^3$, and it is positive if $a < \theta/(\theta-1)$. Hence, it is sufficient to verify (6.28) at $a = \pi_1(\theta) = -1/\Pi_1^{-1}(\theta) = (1+\sqrt{1+4\theta(1-\theta)})/(2(\theta-1))$ if $\theta \in (0, 0.8]$, and at $a = \pi_3(\theta) = -1/\Pi_3^{-1}(\theta) = (133+25(\theta-1))/(95(\theta-1))$ if $\theta \in [0.8, 1)$.

Using (6.1) and replacing the value $a = \pi_3(\theta)$ in (6.28), we get the following expression:

$$\begin{aligned} \frac{q^2}{2(133+25q)^3}[&-40522972 + 220135634q - 410248779q^2 + 204446752q^3 + 279016108q^4 \\ &+ 23396520q^5 - 209505145q^6 - 30804850q^7 + 35072100q^8 + 7581000q^9], \end{aligned}$$

which is negative for $\theta = q+1 \in [0.8, 1)$. Direct computations show that (6.28) holds if $a = \pi_1(\theta)$ and $\theta \in [0, 0.8]$.

Analogously, inequality (6.25) is equivalent to

$$-(6 - 3\theta^2 + 6\theta + 2\theta^3) - \theta(18 - 9\theta + 2\theta^2)e^{\theta/a}(1-a)a^{-1} < 0. \tag{6.29}$$

Using (6.1) and substituting the value $a = \pi_3(\theta)$ in (6.29), we get the expression

$$\begin{aligned} \frac{q^2}{2(133+25q)^3}[&-4465209 - 971090q - 12743680q^2 \\ &- 5731130q^3 - 2242475q^4 + 1103900q^5 - 1263500q^6], \end{aligned}$$

which is negative for $\theta = q+1 \in [0.8, 1)$. Direct computations show again that (6.29) is satisfied for $a = \pi_1(\theta)$ and $\theta \in [0, 0.8]$.

Finally, (6.27) is equal to

$$\begin{aligned} &(\theta^6 - 16\theta^5 + 2\theta^4 + 226\theta^3 + 63\theta^2 - 570\theta - 762) \\ &+ \theta(15\theta^5 - 32\theta^4 - 212\theta^3 + 550\theta^2 + 813\theta - 2190)e^{\theta/a}(1-a)a^{-1} < 0. \end{aligned} \tag{6.30}$$

Next, employing (6.1) and using the value $a = \pi_3(\theta)$ in (6.30), we get the expression

$$\frac{q^2}{2(133+25q)^3}[-14595952 + 471367808q - 1571124744q^2 - 18258802q^3 + 723267159q^4 + 399356020q^5 - 311046000q^6 - 73063100q^7 + 42576625q^8 + 9476250q^9],$$

which is negative for $\theta = q + 1 \in [0.8, 1)$. Direct computations also show in this case that (6.30) holds if $a = \pi_1(\theta)$ and $\theta \in [0, 0.8]$.

To finish the proof of this lemma, we take an arbitrary $r \in [a, a_*]$ (so that $r = a_* k, k \geq 1$) and write function $S(r, a, \theta)$ in the form

$$S(r, a, \theta) = \sum_{i=0}^{3} S_i(a, \theta) a_*^i k^i = k^3 \left(S_3 a_*^3 + \frac{1}{k} S_2 a_*^2 + \frac{1}{k^2} S_1 a_* + \frac{1}{k^3} S_0 \right).$$

First, note that $S_3 a_*^3 > 0$. Indeed, if $S_3 a_*^3 \leq 0$, then, in view of (6.25)–(6.27), $S_i a_*^i \leq 0$ for $i = 0, 1, 2$, and therefore $S(a_*, a, \theta) \leq 0$, contradicting Lemma 6.8. Next, the conclusion of Lemma 6.9 is obvious if $S_i a_*^i \geq 0$ for all $i = 0, 1, 2$. Finally, if $S_i a_*^i \leq 0$ and $S_{i+1} a_*^{i+1} > 0$ for some i, then using the above representation for $S(r, a, \theta)$ and relations (6.26)–(6.27), it is easy to see that $S(r, a, \theta) \geq S(a_*, a, \theta) > 0$ for $r \in [a, a_*]$. □

6.4. Properties of function F_1 in the domain $\mathcal{S}$.

LEMMA 6.10. *If $r \in (a, 0)$ and $h \leq 1$, then*

$$\mathcal{F}_1(r) > \frac{1 - h - e^{-h}}{2 - h - e^{-h}} \frac{ar}{1 + r\frac{1-h-e^{-h}}{1-e^{-h}}} = \mathcal{R}_2(r). \tag{6.31}$$

Proof. Take $z > 0$ and consider the point $t_* \in (0, h)$ defined in (4.6); by Lemma 4.6, $F_1(z) > a$. Since $r(r(z)(1 - e^{-(s-h)})) < 0$ for all $s \in (0, h)$, it follows from (4.6) that

$$\begin{aligned} \mathcal{F}_1(r(z)) = F_1(z) &> e^{-(t^*-h)} \int_0^h e^{s-h} r(r(z)(1 - e^{-(s-h)}))ds \\ &= \frac{r(z) - r^{-1}(F_1(z))}{r(z)} \int_{-h}^0 e^u r(r(z)(1 - e^{-u}))du \\ &= \phi(r(z)) - r^{-1}(F_1(z))\psi(r(z)), \end{aligned} \tag{6.32}$$

where

$$\psi(x) = \phi(x)/x, \ \phi(x) = \int_{-h}^0 e^u r(x(1 - e^{-u}))du.$$

Applying Jensen's inequality [12, p. 110] to the last integral, we obtain that

$$\begin{aligned} \phi(x) &= \int_{-h}^0 (1 - e^{-h}) r(x(1 - e^{-u})) d\left(e^u/(1 - e^{-h})\right) \\ &\geq (1 - e^{-h}) r\left(\frac{\int_{-h}^0 x(e^u - 1)du}{1 - e^{-h}} \right) = \frac{ax(1 - h - e^{-h})}{1 + x\frac{1-h-e^{-h}}{1-e^{-h}}} \stackrel{\text{def}}{=} x\mathcal{H}(x). \end{aligned} \tag{6.33}$$

Denote $\psi = \psi(r)$, $\phi = \phi(r)$, $\mathcal{F}_1 = \mathcal{F}_1(r)$. Now, for $r < 0$, (6.32) implies that $\mathcal{F}_1 > \phi - (\mathcal{F}_1\psi)/(a - \mathcal{F}_1)$. Since $a - \mathcal{F}_1 < 0$, we conclude that

$$\mathcal{F}_1^2 - \mathcal{F}_1(\phi + \psi + a) + a\phi > 0. \tag{6.34}$$

Next we prove that, under our assumptions,

$$(\psi+\phi+a)^2-4a\phi>(\psi+\phi-a-2\psi_0)^2\geq 0, \tag{6.35}$$

where $\psi_0=a(1-h-e^{-h})$. Indeed, (6.35) amounts to

$$\psi(\psi_0 r+a+\psi_0)>\psi_0(a+\psi_0).$$

Since $\psi_0 r+a+\psi_0<0$, the latter inequality is equivalent to

$$\psi<\frac{\psi_0(a+\psi_0)}{\psi_0 r+a+\psi_0}=\frac{a(1-h-e^{-h})}{1+r\frac{1-h-e^{-h}}{2-h-e^{-h}}}\stackrel{\text{def}}{=}\mathcal{G}(r),$$

which holds because for $a<0, r<0, h\leq 1$ we have $\mathcal{H}(r)<\mathcal{G}(r)$, and because, by (6.33), $\psi(r)\leq\mathcal{H}(r)$. Now, the inequalities $a\phi(r(z))>0$, (6.35) and the continuous dependence of $\phi(r),\psi(r),\mathcal{F}_1(r)$ on $r\in(a,0)$ imply that the quadratic polynomial $y(x)=x^2-x(\phi+\psi+a)+a\phi$ has two roots $x_1=x_1(r)<x_2=x_2(r)$ with the same sign and that this sign is the same for all $r\in(a,0)$. Similarly, by (6.34), we have that either $\mathcal{F}_1(r)<x_1(r)$ or $\mathcal{F}_1(r)>x_2(r)$ for all $r\in(a,0)$. Since $\mathcal{F}_1(0^-)=0>x_1(0^-)=\psi_0+a$, we conclude that $x_1(r),x_2(r)$ are negative for all $r\in(a,0)$, and $\mathcal{F}_1(r)>x_2(r)$. In other words,

$$\begin{aligned}\mathcal{F}_1&>\frac{1}{2}(\psi+\phi+a+\sqrt{(\psi+\phi+a)^2-4a\phi})\\&=\frac{2a\phi}{\psi+\phi+a-\sqrt{(\psi+\phi+a)^2-4a\phi}}\geq\frac{2a\phi}{2(a+\psi_0)},\end{aligned}\tag{6.36}$$

where the last inequality is due to the following consequence of (6.35):

$$\sqrt{(\psi+\phi+a)^2-4a\phi}\geq -a+\phi+\psi-2\psi_0.$$

Finally, combining (6.33) and (6.36), we obtain (6.31). □

LEMMA 6.11. *Assume that* $(a,\theta)\in\mathcal{S}$. *Then* $r(\mathcal{R}_2(a))<\beta^{-1}$.

Proof. Step (i). In the new variables $q=\theta-1, k=a(\theta-1)$, the expression for $\mathcal{R}_2(a)$ takes the form

$$\mathcal{R}_2(a)=\frac{-q+\ln(q+1)}{1-q+\ln(q+1)}\frac{k^2}{q^2-k(-q+\ln(q+1))}.$$

Next we prove that

$$\mathcal{R}_2(a)\geq\frac{6k^2(q-1)}{3kq^2-4kq+12+6k}\stackrel{\text{def}}{=}\bar{R}_2(q,k) \tag{6.37}$$

for all $q\in[-0.2,0), k\in[1,1.5]$. (Note that for $(a,\theta)\in\mathcal{D}$, the inequalities $1\leq a(\theta-1)\leq 1.5$ hold.) Indeed, we have

$$\mathcal{R}_2(a)-\bar{R}_2=\frac{-k^2C(q,k)}{(1-q+L)(-q^2-kq+kL)(3kq^2-4kq+12+6k)},$$

where $L=\ln(1+q)$, and

$$\begin{aligned}C(q,k)=&\,q(6q^3-12q^2+3kq^2+6q-8kq-12)\\&+(14kq+12-9kq^2+6q^2-6q^3)L+6k(q-1)L^2.\end{aligned}$$

Next, the following inequalities hold in an obvious way for $q \in [-0.2, 0)$ and $k \in [1, 1.5]$:

$$1 - q + \ln(1+q) \geq 1 - q + q/(1+q) > 0,$$
$$3kq^2 - 4kq + 12 + 6k > 0,$$
$$-q^2 - kq + k\ln(1+q) < -q^2 < 0.$$

Thus $C(q,k) > 0$ will imply that $\mathcal{R}_2(a) > \bar{R}_2$. Now, in view of (6.5) and the obvious inequalities $\min_{\{q\in[-0.2,0),k\in[1,1.5]\}}(14kq + 12 - 9kq^2 + 6q^2 - 6q^3) \geq 7.26 > 0$ and $6k(q-1) < 0$, we obtain that $C(q,k) > (q^3/50)(-168kq^3 + 48kq^4 - 120q^3 + 255kq^2 + 270q^2 - 150q - 110kq - 50k - 60) \geq -0.356176q^3 > 0$.

Step (ii). Using the new variables, we obtain the following expression for α:

$$\alpha(a,\theta) = \alpha(k/q, 1+q) = (1 - k/q)\exp(q(q+1)/k) + k/q.$$

We will prove that $\alpha > -q(24k^2 - 12k - 7q)/(24k^2) \stackrel{\text{def}}{=} \bar{\alpha} > 0$. Indeed, since $\exp(x) > 1 + x + x^2/2 + x^3/6$ for all $x = q(1+q)/k < 0$, we get

$$\alpha(q,k) - \bar{\alpha} > q^2(24k)^{-3}[4q^4 + (12-4k)q^3 + 12q^2 + (-12k^2 + 12k + 4)q + k],$$

where the right-hand side is positive for all $q \in [-0.2, 0), k \in [1, 1.5]$.

Step (iii). Set $E = \exp(\frac{q(q+1)}{k})$. Here we prove that

$$\alpha\beta = -k/q - (q/k)(-2q + 2k - 1)E + (k-q)^2(kq)^{-1}E^2$$
$$< \bar{\beta}_0 \stackrel{\text{def}}{=} q^2(q+1)[2k^2 - q(k + 2k^2) + q^2(9 - 11k + 2k^2)](6k^4)^{-1}.$$

Indeed, due to (6.1) and since $\frac{-q}{k}(-2q + 2k - 1) > 0$, $\frac{(k-q)^2}{kq} < 0$, we obtain

$$\bar{\beta}_0 - \alpha\beta > (1/6)(q/k)^4(q+1)(-8q^2 + 10kq - 16q + 1) > 0$$

if $q \in [-0.2, 0)$, $k \in [1, 1.5]$.

Step (iv). First, note that $\bar{R}_2(q,k) > -1$ for all $q \in [-0.2, 0), k \in [1, 1.5]$ so that $r(\bar{R}_2(q,k))$ and $r(\mathcal{R}_2(a))$ are well defined. Moreover, since r is strictly decreasing over $(-1, 0)$, by virtue of (6.37) we get

$$0 < r(\mathcal{R}_2(a)) < r(\bar{R}_2) = \frac{6k^3(q-1)}{q[3kq^2 + q(6k^2 - 4k) - 6k^2 + 6k + 12]}. \tag{6.38}$$

Step (v). The above steps imply that

$$r(\mathcal{R}_2(a))\beta = r(\mathcal{R}_2(a))\alpha\beta/\alpha < r(\bar{R}_2)\bar{\beta}_0/\bar{\alpha}.$$

Hence, Lemma 6.8 will be proved if we show that $r(\bar{R}_2)\bar{\beta}_0/\bar{\alpha} - 1 < 0$. We have

$$r(\bar{R}_2)\frac{\bar{\beta}_0}{\bar{\alpha}} - 1 = \frac{\sum_{i=0}^4 Z_i q^i}{(3kq^2 - 4qk + 12 + 6k - 6k^2 + 6qk^2)(-12k - 7q + 24k^2)}, \tag{6.39}$$

where $Z_0 = 24k(6k^3 - 7k^2 - 9k + 6)$, $Z_1 = -144k^4 + 120k^3 - 114k^2 + 42k + 84$, $Z_2 = -2k(36k^2 + 93k - 94)$, $Z_3 = 3k(16k^2 + 8k + 7)$, $Z_4 = -24k(k-1)(2k-9)$.

Now, in view of (6.38), we have that the denominator of the right-hand side of (6.39) is positive for all $q \in [-0.2, 0), k \in [1, 1.5]$. Therefore it suffices to prove that $\sum_{i=0}^4 Z_i q^i < 0$; we finish the proof by observing that, for $q \in [-0.2, 0)$, $k \in [1, 1.5]$,

$$Z_0 + Z_1 q \leq Z_0 - 0.2Z_1 = 0.3(2k-3)(288k^3 + 112k^2 - 154k - 5) - 21.3 < 0,$$
$$Z_2 q^2 = -2kq^2(36k^2 + 93k - 94) < 0,$$
$$Z_3 + Z_4 q \geq Z_3 - 0.2Z_4 = 57.6k^3 - 28.8k^2 + 64.2k > 0. \qquad \square$$

Acknowledgment. The authors are greatly indebted to an anonymous referee for his/her valuable suggestions, which helped them to improve the exposition of the results.

REFERENCES

[1] F. Brauer and C. Castillo-Chávez, *Mathematical Models in Population Biology and Epidemiology*, Springer-Verlag, New York, 2001.
[2] K. Cooke, P. van den Driessche, and X. Zou, *Interaction of maturation delay and nonlinear birth in population and epidemic models*, J. Math. Biol., 39 (1999), pp. 332–352.
[3] I. Györi and S. Trofimchuk, *Global attractivity in* $x'(t) = -\delta x(t) + pf(x(t-\tau))$, Dynam. Systems Appl., 8 (1999), pp. 197–210.
[4] J.K. Hale, *Asymptotic Behavior of Dissipative Systems*, Math. Surveys Monogr. 25, AMS, Providence, RI, 1988.
[5] J.K. Hale and S.M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Appl. Math. Sci., Springer-Verlag, New York, 1993.
[6] A. Ivanov, E. Liz, and S. Trofimchuk, *Halanay inequality, Yorke* 3/2 *stability criterion, and differential equations with maxima*, Tohoku Math. J. (2), 54 (2002), pp. 277–295.
[7] Y. Kuang, *Delay Differential Equations with Applications in Population Dynamics*, Academic Press, Boston, MA, 1993.
[8] E. Liz, M. Pinto, G. Robledo, V. Tkachenko, and S. Trofimchuk, *Wright type delay differential equations with negative Schwarzian*, Discrete Contin. Dynam. Systems, 9 (2003), pp. 309–321.
[9] E. Liz, V. Tkachenko, and S. Trofimchuk, *Yorke and Wright* 3/2*-stability theorems from a unified point of view*, Discrete Contin. Dynam. Systems, expanded volume (2003), pp. 580–589.
[10] J. Mallet-Paret and R. Nussbaum, *A differential-delay equation arising in optics and physiology*, SIAM J. Math. Anal., 20 (1989), pp. 249–292.
[11] M. Pinto and S. Trofimchuk, *Stability and existence of multiple periodic solutions for a quasilinear differential equation with maxima*, Proc. Roy. Soc. Edinburgh Sect. A, 130 (2000), pp. 1103–1118.
[12] H.L. Royden, *Real Analysis*, Macmillan, New York, 1969.
[13] D. Singer, *Stable orbits and bifurcation of maps of the interval*, SIAM J. Appl. Math., 35 (1978), pp. 260–267.
[14] H.L. Smith, *Monotone Dynamical Systems. An Introduction to the Theory of Competitive and Cooperative Systems*, AMS, Providence, RI, 1995.
[15] H.-O. Walther, *Contracting return maps for some delay differential equations*, in Functional Differential and Difference Equations, T. Faria and P. Freitas, eds., Fields Inst. Commun. 29, AMS, Providence, RI, 2001, pp. 349–360.
[16] T. Yoneyama, *On the* 3/2 *stability theorem for one-dimensional delay-differential equations*, J. Math. Anal. Appl., 125 (1987), pp. 161–173.
[17] J.A. Yorke, *Asymptotic stability for one dimensional differential-delay equations*, J. Differential Equations, 7 (1970), pp. 189–202.

SIAM J. MATH. ANAL.
Vol. 35, No. 3, pp. 623–638

ON THE OVERSAMPLING OF AFFINE WAVELET FRAMES*

BRODY DYLAN JOHNSON†

Abstract. The properties of oversampled affine frames are considered here with two main goals in mind. The first goal is to generalize the approach of Chui and Shi [*Proc. Amer. Math. Soc.*, 121 (1994), pp. 511–517], [*SIAM J. Math. Anal.*, 28 (1997), pp. 213–232] to the matrix oversampling setting for expanding, lattice-preserving dilations, whereby we obtain a new proof of the second oversampling theorem for affine frames. The second oversampling theorem, proven originally by Ron and Shen [*J. Funct. Anal.*, 148 (1997), pp. 408–447] via Gramian analysis, states that oversampling an affine frame with dilation M by a matrix P will result in a frame with the same bounds (after renormalization), provided that P and M satisfy a certain relative primality condition. In this case, the matrix P is said to be admissible for M. The second goal of this work is to examine the compatibility of admissible oversampling with the refinable affine frames arising from a certain class of scaling functions. In this setting we show that oversampling dual affine systems by an admissible P preserves the multiresolution structure and, from this fact, that the oversampled systems remain dual. We then show that the admissibility of P is also sufficient to endow the dual oversampled systems with a discrete wavelet transform. The novelty of this work lies both in our approach to the second oversampling theorem as well as our consideration of oversampling in the context of multiresolution analysis.

Key words. affine system, oversampling, wavelet, multiresolution analysis

AMS subject classifications. 42C15, 65T60

DOI. 10.1137/S0036141002406758

1. Introduction.

Unless otherwise stated, M will denote a fixed $n \times n$ dilation matrix with integer entries such that each eigenvalue λ of M satisfies $|\lambda| > 1$. We will refer to $M \in GL_n(\mathbb{R})$ as *expanding* if its eigenvalues satisfy this latter condition. Thus M is a $\mathbb{Z}^n$-lattice preserving, expanding dilation. The unitary dilation operator on $L^2(\mathbb{R}^n)$ induced by M will be denoted D and is defined by $Df(x) := |\det M|^{\frac{1}{2}} f(Mx)$ for $f \in L^2(\mathbb{R}^n)$. For $u \in \mathbb{R}^n$, let T_u denote the usual translation operator, i.e., $T_u f(x) := f(x-u)$. With these basic ingredients we may now recall the definition of an affine system.

DEFINITION 1.1. *Let $\Psi = \{\psi_1, \dots, \psi_L\} \subset L^2(\mathbb{R}^n)$. The affine system generated by Ψ, denoted $X(\Psi)$, is the collection*

$$X(\Psi) = \{\psi_{\ell;j,k} : 1 \le \ell \le L, j \in \mathbb{Z}, k \in \mathbb{Z}^n\},$$

where $\psi_{\ell;j,k} := D^j T_k \psi_\ell$.

Our interest lies in those affine systems that constitute frames for $L^2(\mathbb{R}^n)$.

DEFINITION 1.2. *Let $\mathbb{H}$ be a Hilbert space. The collection $\{h_j\}_{j\in J} \subset \mathbb{H}$ is a frame for $\mathbb{H}$ if there exist constants $A, B > 0$ such that for all $f \in \mathbb{H}$*

$$A\|f\|_{\mathbb{H}}^2 \le \sum_{j\in J} |\langle f, h_j\rangle_{\mathbb{H}}|^2 \le B\|f\|_{\mathbb{H}}^2. \tag{1.1}$$

The constants A and B are referred to as the lower and upper frame bounds, respectively. In the case that $A = B$ the frame is said to be *tight*. If only the right

*Received by the editors May 1, 2002; accepted for publication (in revised form) March 14, 2003; published electronically October 2, 2003.
http://www.siam.org/journals/sima/35-3/40675.html

†Department of Mathematics, Washington University, One Brookings Drive, St. Louis, MO 63103. Current address: Department of Mathematics and Mathematical Computer Science, Saint Louis University, St. Louis, MO 63103 (brody@slu.edu).

inequality holds, the system is called a *Bessel system*, and in this case B is referred to as the Bessel bound. We say two frames for $\mathbb{H}$, $\{h_j\}_{j\in J}$ and $\{\tilde{h}_j\}_{j\in J}$, are *dual frames* if for each $f \in \mathbb{H}$ we have

$$f = \sum_{j\in J} \langle f, \tilde{h}_j \rangle h_j. \tag{1.2}$$

Let $GL_n(\mathbb{Z})$ denote the set of all $n \times n$ matrices with integer entries having nonzero determinant. Given $P \in GL_n(\mathbb{Z})$, we now define the oversampled affine system generated by a family Ψ relative to P.

DEFINITION 1.3. *Let $\Psi = \{\psi_1, \ldots, \psi_L\} \subset L^2(\mathbb{R}^n)$. The oversampled affine system generated by Ψ relative to $P \in GL_n(\mathbb{Z})$, denoted $X^P(\Psi)$, is the collection*

$$X^P(\Psi) := \{\psi^P_{\ell;j,k} : 1 \le \ell \le L, j \in \mathbb{Z}, k \in \mathbb{Z}^n\},$$

where $\psi^P_{\ell;j,k} := \frac{1}{\sqrt{p}} D^j T_{P^{-1}k} \psi_\ell$ and $p := |\det P|$.

The factor $\frac{1}{\sqrt{p}}$ in Definition 1.3 compensates for the increase in the density of the lattice of translations caused by the oversampling. This allows us to compare the frame bounds of the oversampled and nonoversampled systems.

This notion of oversampling was introduced by Chui and Shi in [2] when they proved that oversampling a dyadic affine frame ($M = 2$) in one dimension by p odd preserves the frame bounds. In [3], Chui and Shi later extended this result to the multivariate case in which the dilation $M \in GL_n(\mathbb{Z})$ is expanding and $P = pI$ with $\gcd(p, |\det M|) = 1$. The result is referred to there as the second oversampling theorem. Since the one-dimensional result appeared, several other researchers have investigated the problem of bound-preserving oversampling for affine frames. In the case that $M, P \in GL_n(\mathbb{Z})$ with M expanding, Ron and Shen have used their Gramian analysis to show that a relative primality condition on the lattices $M^T\mathbb{Z}^n$ and $P^T\mathbb{Z}^n$ is sufficient for bound-preserving oversampling [7]. More recently, the work of Laugesen [6] provides another approach to the second oversampling theorem, which employs the concept of almost periodicity. In [6] it is observed that the conditions on M and P for bound-preserving oversampling described in [7] and [6] are equivalent. We should also mention that Chui, Czaja, Maggioni, and Weiss have developed a notion of tightness-preserving oversampling based on the characterization of affine tight-frames [1]. In their work the dilation matrix M is not required to have integer entries; however, the result applies only to tight-frames.

During the revision of this paper the author became aware of an interesting work by Hernández et al. [4] in which the various embodiments of oversampling are unified into a single theory, including quasi-affine systems as well as oversampled affine systems.

The techniques used by Ron and Shen in [7] and Laugesen in [6] are quite different from those used originally by Chui and Shi in [2]. In each case a notion of relative primality between the dilation matrix and the oversampling matrix has proven essential in the proof of the second oversampling theorem. One of the goals of our work is to extend the original ideas of Chui and Shi to the matrix oversampling case with a careful development of the relative primality conditions. These conditions are introduced in section 2, where we define a notion of admissible oversampling and develop related elementary results. In section 3 we present our version of the second oversampling theorem.

The second goal of this work is to explore the compatibility of admissible oversampling with multiresolution analysis. In section 4 we restrict our attention to dual

refinable affine systems associated with a certain class of scaling functions. We introduce multiresolution operators for the oversampled systems and show that they behave much like those associated with the original affine systems. This allows us to prove that the duality of refinable affine frames is preserved under admissible oversampling. Finally, we show that admissible oversampling endows the dual oversampled systems with a discrete wavelet transform (DWT).

To close the section let us note that we will adopt the following definition for the Fourier transform, $\hat{f}$, of $f \in L^2(\mathbb{R}^n)$:

$$\hat{f}(\xi) = \int_{\mathbb{R}^n} f(x) e^{-i\langle \xi, x\rangle} dx.$$

2. Admissible oversampling matrices. Given a candidate oversampling matrix, $P \in GL_n(\mathbb{Z})$, we are concerned with the quotient group $P^{-1}\mathbb{Z}^n/\mathbb{Z}^n$. Let $\{\theta_r\}_{r=0}^{p-1}$ be a complete set of distinct coset representatives of the quotient group $P^{-1}\mathbb{Z}^n/\mathbb{Z}^n$, where again $p = |\det P|$. In the following proposition we consider conditions on P such that the action of M on $P^{-1}\mathbb{Z}^n/\mathbb{Z}^n$ is nice.

PROPOSITION 2.1. *Suppose $M, P \in GL_n(\mathbb{Z})$ with $m := |\det M|$ and $p := |\det P|$. Let $\{\theta_r\}_{r=0}^{p-1}$ be a complete set of distinct coset representatives of $P^{-1}\mathbb{Z}^n/\mathbb{Z}^n$ with $\theta_0 = 0$. Suppose $PMP^{-1} \in GL_n(\mathbb{Z})$. Then $\{M\theta_r\}_{r=0}^{p-1}$ is a complete set of representatives of $P^{-1}\mathbb{Z}^n/\mathbb{Z}^n$ if and only if M and P satisfy $M^{-1}\mathbb{Z}^n \bigcap P^{-1}\mathbb{Z}^n = \mathbb{Z}^n$.*

Proof. The statement is trivial if $p = 1$. We proceed to prove the result in the case that $p \geq 2$.

($\Rightarrow$) By way of contradiction assume that $M^{-1}\mathbb{Z}^n \bigcap P^{-1}\mathbb{Z}^n \supsetneq \mathbb{Z}^n$. Then there exists θ_{r_0}, $1 \leq r_0 \leq p-1$, such that $\theta_{r_0} \in \left(P^{-1}\mathbb{Z}^n \bigcap M^{-1}\mathbb{Z}^n\right) \setminus \mathbb{Z}^n$. This implies that $M\theta_{r_0} \equiv 0 \pmod{\mathbb{Z}^n}$, which means $\{M\theta_r\}_{r=0}^{p-1}$ cannot be a complete set of representatives of $P^{-1}\mathbb{Z}^n/\mathbb{Z}^n$. This is a contradiction.

($\Leftarrow$) The condition $PMP^{-1} \in GL_n(\mathbb{Z})$ implies that $Mx \in P^{-1}\mathbb{Z}^n$ if $x \in P^{-1}\mathbb{Z}^n$; i.e., the multiplication map induced by M maps $P^{-1}\mathbb{Z}^n$ into itself. Suppose $\{M\theta_r\}_{r=0}^{p-1}$ is not a complete set of coset representatives of $P^{-1}\mathbb{Z}^n/\mathbb{Z}^n$. Then there is some θ_{r_0}, $1 \leq r_0 \leq p-1$, such that $M\theta_{r_0} \in \mathbb{Z}^n$, which implies that $\theta_{r_0} \in M^{-1}\mathbb{Z}^n$. Since $r_0 \neq 0$, $\theta_{r_0} \notin \mathbb{Z}^n$, implying that $\theta_{r_0} \in \left(P^{-1}\mathbb{Z}^n \bigcap M^{-1}\mathbb{Z}^n\right) \setminus \mathbb{Z}^n$, and we have $M^{-1}\mathbb{Z}^n \bigcap P^{-1}\mathbb{Z}^n \supsetneq \mathbb{Z}^n$, a contradiction. ∎

DEFINITION 2.2. *Let $P \in GL_n(\mathbb{Z})$. P is an admissible oversampling matrix for M if $PMP^{-1} \in GL_n(\mathbb{Z})$ and $M^{-1}\mathbb{Z}^n \bigcap P^{-1}\mathbb{Z}^n = \mathbb{Z}^n$. If the matrix M is clear from the context we will simply say that P is admissible.*

In the terminology of the preceding proposition, notice that if $\gcd(m, p) = 1$, then $M^{-1}\mathbb{Z}^n \bigcap P^{-1}\mathbb{Z}^n = \mathbb{Z}^n$. Indeed, suppose $\theta \in \left(M^{-1}\mathbb{Z}^n \bigcap P^{-1}\mathbb{Z}^n\right)$. Since the order of θ divides both m and p we conclude that $\theta \in \mathbb{Z}^n$.

Example 1. Consider the following examples of admissible oversampling matrices.

(a) Let $M = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$, the Quincunx dilation matrix, and let $P = \begin{pmatrix} 3 & 1 \\ -2 & 1 \end{pmatrix}$. It is easy to check that PMP^{-1} has integer entries and, in light of the previous remark, that P is admissible.

(b) Let $M = mI_n$, where $m \geq 2$ is an integer and I_n is the $n \times n$ identity matrix. Clearly, $PMP^{-1} \in GL_n(\mathbb{Z})$ for all $P \in GL_n(\mathbb{Z})$, which means a sufficient condition for P to be admissible is that $\gcd(m, |\det P|) = 1$.

Given that P is admissible, Proposition 2.1 tells us that the mapping $\theta_r \mapsto M\theta_r$, $0 \leq r \leq p-1$, acts to permute the coset representatives of $P^{-1}\mathbb{Z}^n/\mathbb{Z}^n$. Let σ be the permutation of $\{0, \ldots, p-1\}$ such that $\theta_{\sigma(r)} \equiv M\theta_r \pmod{\mathbb{Z}^n}$ in $P^{-1}\mathbb{Z}^n$. Let σ^{-1} be the associated inverse permutation. The following result, which replaces Lemma 2 of [2] in this setting, describes a basic property of the permutation σ.

LEMMA 2.3. *Let $j_0 \in \mathbb{Z}$. If $P \in GL_n(\mathbb{Z})$ is admissible, then for $j \geq j_0$ and $0 \leq r \leq p-1$, $\theta_{\sigma^j(r)} \equiv M^{j-j_0}\theta_{\sigma^{j_0}(r)}$ $(mod\, \mathbb{Z}^n)$ in $P^{-1}\mathbb{Z}^n$.*

Proof. The statement holds trivially for $j = j_0$. By induction, assume the formula holds for j, and we will derive the formula for $j+1$. Proposition 2.1 and the definition of σ imply that

$$\theta_{\sigma^{j+1}(r)} \equiv M\theta_{\sigma^j(r)} \equiv M^{j+1-j_0}\theta_{\sigma^{j_0}(r)} \pmod{\mathbb{Z}^n}. \qquad \square$$

COROLLARY 2.4. *Let $j, j_0 \in \mathbb{Z}$ with $j \geq j_0$, and suppose $P \in GL_n(\mathbb{Z})$ is admissible. For $0 \leq r \leq p-1$,*

$$\{D^j T_{\theta_{\sigma^j(r)}+k}\psi_\ell : 1 \leq \ell \leq L, k \in \mathbb{Z}^n\} = \{T_{M^{-j_0}\theta_{\sigma^{j_0}(r)}} D^j T_k \psi_\ell : 1 \leq \ell \leq L, k \in \mathbb{Z}^n\}.$$

3. The second oversampling theorem. We now seek to describe our version of the second oversampling theorem, generalizing the approach of Chui and Shi introduced in [2]. We begin by demonstrating the preservation of Bessel bounds for admissible oversampling, which follows essentially from Proposition 2.1.

LEMMA 3.1. *Suppose $X(\Psi)$ is a Bessel system with bound $B > 0$ relative to an expanding dilation matrix $M \in GL_n(\mathbb{Z})$. If $P \in GL_n(\mathbb{Z})$ is an admissible oversampling matrix, then $X^P(\Psi)$ is a Bessel system with the same bound.*

Proof. The fact that $X(\Psi)$ is Bessel with bound $B > 0$ implies for each $f \in L^2(\mathbb{R}^n)$ that

$$\sum_{\ell=1}^{L}\sum_{j\geq 0}\sum_{k\in\mathbb{Z}^n}\sum_{r=0}^{p-1}\frac{1}{p}\left|\langle T_{-\theta_r} f, \psi_{\ell;j,k}\rangle\right|^2 \leq B\|f\|^2.$$

We now relate this sum to the inner products of the oversampled system, $X^P(\Psi)$:

$$\begin{aligned}
\sum_{\ell=1}^{L}\sum_{j\geq 0}\sum_{k\in\mathbb{Z}^n}\sum_{r=0}^{p-1}\frac{1}{p}\left|\langle T_{-\theta_r} f, \psi_{\ell;j,k}\rangle\right|^2 &= \sum_{\ell=1}^{L}\sum_{j\geq 0}\sum_{k\in\mathbb{Z}^n}\sum_{r=0}^{p-1}\frac{1}{p}\left|\langle f, D^j T_{M^j\theta_r+k}\psi_\ell\rangle\right|^2 \\
\text{(by Proposition 2.1)} \quad &= \sum_{\ell=1}^{L}\sum_{j\geq 0}\sum_{k\in\mathbb{Z}^n}\frac{1}{p}\left|\langle f, D^j T_{P^{-1}k}\psi_\ell\rangle\right|^2 \\
&= \sum_{\ell=1}^{L}\sum_{j\geq 0}\sum_{k\in\mathbb{Z}^n}\left|\langle f, \psi^P_{\ell;j,k}\rangle\right|^2.
\end{aligned}$$

Letting $J \geq 0$ and observing that $\|D^J f\|^2 = \|f\|^2$ it is easily shown that for each $f \in L^2(\mathbb{R}^n)$

$$\sum_{\ell=1}^{L}\sum_{j\geq -J}\sum_{k\in\mathbb{Z}^n}\left|\langle f, \psi^P_{\ell;j,k}\rangle\right|^2 \leq B\|f\|^2.$$

Letting $J \to \infty$ we see that $X^P(\Psi)$ is a Bessel system with upper bound B. $\square$

Given an admissible oversampling matrix we can use the permutation σ guaranteed by Proposition 2.1 to rewrite the oversampled system as the union of appropriate

affine-like systems, one for each coset of $P^{-1}\mathbb{Z}^n/\mathbb{Z}^n$. Namely,

$$X^P(\Psi) = \bigcup_{r=0}^{p-1} \frac{1}{\sqrt{p}} S_r \quad \text{(disjointly)},$$

where $S_r := \{D^j T_{\theta_{\sigma^j(r)}+k}\psi_\ell : 1 \le \ell \le L, j \in \mathbb{Z}, k \in \mathbb{Z}^n\}$. S_0 is precisely $X(\Psi)$, while the remaining collections S_r, $1 \le r \le p-1$, are slightly more complicated. This decomposition plays a key role in our proof of the second oversampling theorem.

THEOREM 3.2 (second oversampling theorem). *Suppose $X(\Psi)$ is a frame with lower and upper bounds $A, B > 0$, respectively, relative to an expanding dilation matrix $M \in GL_n(\mathbb{Z})$. If $P \in GL_n(\mathbb{Z})$ is an admissible oversampling matrix, then $X^P(\Psi)$ is a frame with the same bounds.*

Remark. In [6], Laugesen shows that if $X(\Psi)$ and $X(\tilde{\Psi})$ are dual frames and P is admissible, then $X^P(\Psi)$ and $X^P(\tilde{\Psi})$ are also dual frames. This statement will be proven in the next section (see Theorem 4.4 below) for a certain class of refinable functions.

Proof. The preservation of the upper bound was discussed above; hence, it suffices to demonstrate the lower bound. Since $S_0 = X(\Psi)$, S_0 is a frame with lower bound A. It is, therefore, sufficient to prove that each of the collections S_r, $1 \le r \le p-1$, is a frame with lower bound A.

Fix r, $1 \le r \le p-1$, and let $f \in L^\infty_c(\mathbb{R}^n)$, the dense subset of $L^2(\mathbb{R}^n)$ consisting of essentially bounded functions of compact support. It is sufficient to demonstrate the lower bound for such an f. Suppose that $\operatorname{supp} f \subset K$, where K is a compact subset of $\mathbb{R}^n$ containing 0. Let $R := \operatorname{diam} K$. Lastly, let $\lambda_- > 1$ and λ_+ be the strict lower and upper bounds, respectively, for the moduli of the eigenvalues of M.

1. For $j_0 \in \mathbb{Z}$, let $f^r_{j_0} = T_{-M^{-j_0}\theta_{\sigma^{j_0}(r)}} f$. By defining $K^r_{j_0} := K - M^{-j_0}\theta_{\sigma^{j_0}(r)}$ we see that $\operatorname{supp} f^r_{j_0} \subset K^r_{j_0}$. Observe that by Corollary 2.4 we have

$$\begin{aligned}
\sum_{g\in S_r} |\langle f, g\rangle|^2 &= \sum_{\ell=1}^{L}\sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}^n} \left|\langle f, D^j T_{\theta_{\sigma^j(r)}+k}\psi_\ell\rangle\right|^2 \\
&= \sum_{\ell=1}^{L}\sum_{j\ge j_0}\sum_{k\in\mathbb{Z}^n} \left|\langle f^r_{j_0}, D^j T_k \psi_\ell\rangle\right|^2 + \sum_{\ell=1}^{L}\sum_{j<j_0}\sum_{k\in\mathbb{Z}^n} \left|\langle f, D^j T_{\theta_{\sigma^j(r)}+k}\psi_\ell\rangle\right|^2 \\
&\ge \sum_{\ell=1}^{L}\sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}^n} \left|\langle f^r_{j_0}, \psi_{\ell;j,k}\rangle\right|^2 - \sum_{\ell=1}^{L}\sum_{j<j_0}\sum_{k\in\mathbb{Z}^n} \left|\langle f^r_{j_0}, \psi_{\ell;j,k}\rangle\right|^2 \\
&\ge A\|f\|^2 - \sum_{\ell=1}^{L}\sum_{j<j_0}\sum_{k\in\mathbb{Z}^n} \left|\langle f^r_{j_0}, \psi_{\ell;j,k}\rangle\right|^2.
\end{aligned}$$

   We are left to prove that the latter sum tends to 0 as $j_0 \to -\infty$.

2. Let $\varepsilon > 0$. Let us adopt the notation

$$S^r_{j_0}(f) = \sum_{\ell=1}^{L}\sum_{j<j_0}\sum_{k\in\mathbb{Z}^n} \left|\langle f^r_{j_0}, \psi_{\ell;j,k}\rangle\right|^2.$$

   Estimating the inner product of $f^r_{j_0}$ with $\psi_{\ell;j,k}$ by

$$\left|\langle f^r_{j_0}, \psi_{\ell;j,k}\rangle\right|^2 \le \|f^r_{j_0}\|^2 \, \|\psi_{\ell;j,k}\chi_{K^r_{j_0}}\|^2 \le \|f\|^2 \int_{M^j K^r_{j_0}} |\psi_\ell(x-k)|^2 dx,$$

we obtain a bound on $S^r_{j_0}(f)$:

$$S^r_{j_0}(f) \leq \|f\|^2 \sum_{\ell=1}^{L} \sum_{j<j_0} \sum_{k\in\mathbb{Z}^n} \int_{M^j K^r_{j_0}-k} |\psi_\ell(x)|^2 dx.$$

We will break up the sum over j into two pieces, corresponding to $j < j_0 - J$ and $j_0 - J \leq j < j_0$, where $J > 0$ will be fixed below (independent of j_0). Since M is expanding it is well known that there exists $\beta \geq 1$ such that for $x \in \mathbb{R}^n$ and $j > 0$ we have the estimates

$$\beta^{-1}\lambda_-^j \|x\| \leq \|M^j x\| \leq \beta\lambda_+^j \|x\|$$

and

$$\beta^{-1}\lambda_+^{-j} \|x\| \leq \|M^{-j} x\| \leq \beta\lambda_-^{-j} \|x\|,$$

where $\beta \geq 1$ depends on λ_-, λ_+, and M.
We now make a pair of technical assumptions that will be used below, each of which relies on the expanding property of M.

(a) We may assume j_0 is negative and sufficiently less than 0 such that

$$R < \frac{1}{2}\|M^{-j_0}\theta_{\sigma^{j_0}(r)}\|.$$

(b) We will assume $J > 0$ is such that $\lambda_-^J > 2\beta^{-1}(R + \|\theta_{\sigma^{j_0}(r)}\|)$.

3. Let us first handle the terms for which $j < j_0 - J$. Consider the set

$$E := \bigcup_{j<j_0-J} M^j K^r_{j_0} = \bigcup_{j<j_0-J} \big(M^j(K - M^{-j_0}\theta_{\sigma^{j_0}(r)})\big).$$

Our first estimate involves replacing the sum over j by integration over E, which requires the that the sets $\{M^j K^r_{j_0}\}_{j<j_0-J}$ have finite overlaps which can be bounded independently of j_0. Supposing for the moment that this is the case, we have

$$I_1 := \|f\|^2 \sum_{\ell,k} \sum_{j<j_0-J} \int_{M^j K^r_{j_0}-k} |\psi_\ell(x)|^2 dx \leq C\|f\|^2 \sum_{\ell,k} \int_{E-k} |\psi_\ell(x)|^2 dx.$$

We now investigate the disjointness of $\{M^j K^r_{j_0}\}_{j<j_0-J}$. Suppose that

$$M^{j_1} K^r_{j_0} \bigcap M^{j_2} K^r_{j_0} \neq \emptyset,$$

with $j_1 > j_2$; then $M^{j_1-j_2} K^r_{j_0} \bigcap K^r_{j_0} \neq \emptyset$. Thus it suffices to prove that there exists $j_1 > 0$ (independent of j_0) such that $M^j K^r_{j_0} \bigcap K^r_{j_0} = \emptyset$ for all $j \geq j_1$. For $x \in K^r_{j_0}$ we have by assumption (a) that

$$\|x\| \leq R + \|M^{-j_0}\theta_{\sigma^{j_0}(r)}\| \leq \frac{3}{2}\|M^{-j_0}\theta_{\sigma^{j_0}(r)}\|$$

and

$$\|x\| \geq \|M^{-j_0}\theta_{\sigma^{j_0}(r)}\| - R \geq \frac{1}{2}\|M^{-j_0}\theta_{\sigma^{j_0}(r)}\|.$$

Again using the expanding property of M, we have for $x \in K_{j_0}^r$ and $j > 0$

$$\|M^j x\| \geq \frac{1}{\beta}\lambda_-^j \frac{1}{2}\|M^{-j_0}\theta_{\sigma^{j_0}(r)}\|.$$

Hence, a sufficient condition for the disjointness of the sets $M^j K_{j_0}^r$ and $K_{j_0}^r$ is

$$\frac{1}{\beta}\lambda_-^j \frac{1}{2}\|M^{-j_0}\theta_{\sigma^{j_0}(r)}\| \geq \frac{3}{2}\|M^{-j_0}\theta_{\sigma^{j_0}(r)}\|,$$

or, equivalently,

$$\lambda_-^j \geq 3\beta.$$

Again, since M is expanding we have $\lambda_- > 1$, so we may choose $j_1 > 0$ to be the smallest j for which this last inequality holds.

4. Returning to the estimate from step 3 above, we will next fix J large enough to control the term I_1. Let us examine a typical j in the sum defining I_1, which is of the form $j = j_0 - J - j_1$ with $j_1 \geq 1$. If $x \in M^j K_{j_0}^r$, then

$$\|x\| \leq \beta\lambda_-^{j_0-J-j_1}R + \beta\lambda_-^{-J-j_1}\|\theta_{\sigma^{j_0}(r)}\| \leq \beta\lambda_-^{-J-j_1}\big(R + \|\theta_{\sigma^{j_0}(r)}\|\big),$$

where we have used the assumption that $j_0 < 0$ and β is as above. We conclude that if $x \in E$, then $\|x\| \leq \lambda_-^{-J-1}\big(R + \|\theta_{\sigma^{j_0}(r)}\|\big) \leq \frac{1}{2}$ by the assumption (b) above regarding J. This means that $\operatorname{diam} E \leq 1$, implying that any distinct integer translates of E are disjoint. While the definition of E above does depend on both j_0 and J, we just showed that the measure of E can be made arbitrarily small independent of j_0, by choosing J sufficiently large. Since each $\psi_\ell \in L^2(\mathbb{R}^n)$ the dominated convergence theorem allows us to fix $J > 0$ so large that $I_1 < \varepsilon$ independent of j_0.

5. We now estimate the terms in $S_{j_0}^r(f)$ for $j_0 - J \leq j < j_0$ with J fixed as in the last step. It suffices to consider an arbitrary term of this sort, which can be written as $j = j_0 - j_1$ with $1 \leq j_1 \leq J$. By definition, $M^j K_{j_0}^r = M^{j_0-j_1}K - M^{-j_1}\theta_{\sigma^{j_0}(r)}$, where $M^{-j_1}\theta_{\sigma^{j_0}(r)}$ is one of $p-1$ constants depending on j_0. Hence,

$$I_{2,j} := \sum_{\ell,k}\int_{M^j K_{j_0}^r - k} |\psi_\ell(x)|^2 dx = \sum_{\ell,k}\int_{M^{j_0-j_1}K - M^{-j_1}\theta_{\sigma^{j_0}(r)} - k} |\psi_\ell(x)|^2 dx.$$

Applying the dominated convergence theorem we may choose j_0 sufficiently less than zero such that $I_{2,j} < \varepsilon$.

By definition we have

$$S_{j_0}^r = I_1 + \sum_{j=j_0-J}^{j_0-1} I_{2,j},$$

so, combining all the estimates, we have shown that $S_{j_0}^r \to 0$ as $j_0 \to -\infty$. Thus, S_r is a frame with lower bound A for each r, $0 \leq r \leq p-1$, completing the proof. ∎

4. Oversampling and multiresolution analysis. Consider two families of generating functions, $\Psi := \{\psi_1, \dots, \psi_L\}$ and $\tilde{\Psi} := \{\tilde{\psi}_1, \dots, \tilde{\psi}_L\} \subset L^2(\mathbb{R}^n)$. Let us assume that the families are produced by refinement with scaling functions $\varphi, \tilde{\varphi} \in \mathbb{E}$, respectively, where $\mathbb{E} := \{f \in L^2(\mathbb{R}^n) : [\hat{f}, \hat{f}] \in L^\infty(\mathbb{T}^n)\}$. Recall that $[f, g]$, the *bracket product of* $f, g \in L^2(\mathbb{R}^n)$, is defined by

$$[f, g] = \sum_{k \in \mathbb{Z}^n} T_{2\pi k} f \overline{T_{2\pi k} g}.$$

Adopting the convention that $\psi_0 := \varphi$ and $\tilde{\psi}_0 := \tilde{\varphi}$ for notational convenience, we have the refinement identities

$$\hat{\psi}_\ell(M^T \xi) = m_\ell(\xi)\hat{\varphi}(\xi) \text{ and } \hat{\tilde{\psi}}_\ell(M^T \xi) = \tilde{m}_\ell(\xi)\hat{\tilde{\varphi}}(\xi) \tag{4.1}$$

for $0 \le \ell \le L$ and a.e. $\xi \in \mathbb{R}^n$, where $m_\ell, \tilde{m}_\ell \in L^\infty(\mathbb{T}^n)$ for $0 \le \ell \le L$. We assume here that $M \in GL_n(\mathbb{Z})$ is expansive. Finally, we assume that the filters satisfy the generalized Smith–Barnwell equations for the dilation M; namely, for $0 \le s \le m-1$ we have

$$\sum_{\ell=0}^{L} \overline{m_\ell(\xi)} \tilde{m}_\ell(\xi + 2\pi (M^T)^{-1} \vartheta_s) = \delta_{0,s} \quad \text{a.e. } \xi \in \mathbb{T}^n, \tag{4.2}$$

where $\{\vartheta_s\}_{p=0}^{m-1}$ is a complete set of distinct coset representatives of $\mathbb{Z}^n / M^T \mathbb{Z}^n$, $m := |\det M|$, and $\delta_{0,s}$ is the Kronecker delta. Implicitly assumed here is the fact that $\vartheta_0 = 0$. Together, the scaling functions and filters specify the generating families Ψ and $\tilde{\Psi}$ that define the affine systems $X(\Psi)$ and $X(\tilde{\Psi})$. We will rely on some basic properties of this class of refinable affine systems as found in [5].

Given that these two systems are dual frames for $L^2(\mathbb{R}^n)$ we are interested in two properties of the resulting oversampled systems. First, we will investigate when the oversampled affine systems $X^P(\Psi)$ and $X^P(\tilde{\Psi})$ relative to a matrix $P \in GL_n(\mathbb{Z})$ are again dual frames. Second, we will examine the scaling equations associated with the oversampled system and determine conditions on the oversampling matrix that endow the oversampled system with a bonafide DWT. We conclude the section by reconciling the conditions required for these two properties.

4.1. Multiresolution operators and duality. Our analysis will involve multiresolution operators that arise naturally as generalizations of the orthogonal projections found in the orthonormal MRA case. The affine approximation and detail operators at the scale $j \in \mathbb{Z}$, $\mathcal{P}_j$ and $\mathcal{Q}_j$, respectively, act on $f \in L^2(\mathbb{R}^n)$ by

$$\mathcal{P}_j f := \sum_{k \in \mathbb{Z}^n} \langle f, \tilde{\varphi}_{j,k} \rangle \varphi_{j,k} \text{ and } \mathcal{Q}_j f := \sum_{\ell=1}^{L} \sum_{k \in \mathbb{Z}^n} \langle f, \tilde{\psi}_{\ell;j,k} \rangle \psi_{\ell;j,k}, \tag{4.3}$$

whereas the oversampled affine approximation and detail operators at the scale j, $\mathcal{P}_j^P$ and $\mathcal{Q}_j^P$, respectively, are defined similarly by

$$\mathcal{P}_j^P f := \sum_{k \in \mathbb{Z}^n} \langle f, \tilde{\varphi}_{j,k}^P \rangle \varphi_{j,k}^P \text{ and } \mathcal{Q}_j^P f := \sum_{\ell=1}^{L} \sum_{k \in \mathbb{Z}^n} \langle f, \tilde{\psi}_{\ell;j,k}^P \rangle \psi_{\ell;j,k}^P. \tag{4.4}$$

We have the following basic properties for $\mathcal{P}_j$ and $\mathcal{Q}_j$.

LEMMA 4.1 (see [5]). *Suppose* $\varphi, \tilde{\varphi} \in \mathbb{E}$ *and* $\Psi, \tilde{\Psi} \subset L^2(\mathbb{R}^n)$ *are such that* (4.1) *and* (4.2) *hold.*

(a) $\mathcal{P}_j$ *and* $\mathcal{Q}_j$ *are bounded operators on* $L^2(\mathbb{R}^n)$ *for each* $j \in \mathbb{Z}$.
(b) $\mathcal{P}_j + \mathcal{Q}_j = \mathcal{P}_{j+1}$ *for each* $j \in \mathbb{Z}$.
(c) *For each* $f \in L^2(\mathbb{R}^n)$, $\lim_{j\to-\infty} \|\mathcal{P}_j f\| = 0$.
(d) *If* $X(\Psi)$ *and* $X(\tilde{\Psi})$ *are dual frames for* $L^2(\mathbb{R}^n)$, *then for each* $f \in L^2(\mathbb{R}^n)$ *we have*

$$f = \lim_{j\to\infty} \mathcal{P}_j f = \sum_{j\in\mathbb{Z}} \mathcal{Q}_j f. \tag{4.5}$$

Our objective is to establish similar properties for the oversampled multiresolution operators armed with this information. We begin by expressing the oversampled multiresolution operators in terms of the original affine counterparts.

PROPOSITION 4.2. *Let* $\varphi, \tilde{\varphi} \in \mathbb{E}$. *Let* $P \in GL_n(\mathbb{Z})$. *For each* $j \in \mathbb{Z}$, $\mathcal{P}_j^P$ *and* $\mathcal{Q}_j^P$ *are bounded operators on* $L^2(\mathbb{R}^n)$, *and we have*

(a) $\mathcal{P}_j^P = \frac{1}{p}\sum_{r=0}^{p-1} T_{M^{-j}\theta_r} \mathcal{P}_j T_{-M^{-j}\theta_r}$,
(b) $\mathcal{Q}_j^P = \frac{1}{p}\sum_{r=0}^{p-1} T_{M^{-j}\theta_r} \mathcal{Q}_j T_{-M^{-j}\theta_r}$.

Proof. It is sufficient to derive (a). We have for each $f \in L^2(\mathbb{R}^n)$

$$\begin{aligned}
\mathcal{P}_j^P f &= \sum_{k\in\mathbb{Z}^n} \langle f, \tilde{\varphi}_{j,k}^P \rangle \varphi_{j,k}^P \\
&= \frac{1}{p}\sum_{r=0}^{p-1} \sum_{k\in\mathbb{Z}^n} \langle f, D^j T_{\theta_r+k}\tilde{\varphi}\rangle D^j T_{\theta_r+k}\varphi \\
&= \frac{1}{p}\sum_{r=0}^{p-1} \sum_{k\in\mathbb{Z}^n} \langle T_{-M^{-j}\theta_r} f, D^j T_k \tilde{\varphi}\rangle T_{M^j\theta_r} D^j T_k \varphi \\
&= \frac{1}{p}\sum_{r=0}^{p-1} T_{M^{-j}\theta_r} \mathcal{P}_j T_{-M^{-j}\theta_r} f. \qquad \square
\end{aligned}$$

It is important to realize that the representations of $\mathcal{P}_j^P$ and $\mathcal{Q}_j^P$ above are independent of the choice of coset representatives. This is because $\mathcal{P}_j$ and $\mathcal{Q}_j$ are invariant under conjugation by translation operators over $M^{-j}\mathbb{Z}^n$. Indeed, for each $f \in L^2(\mathbb{R}^n)$ and $k_0 \in \mathbb{Z}^n$ we have

$$\begin{aligned}
T_{M^{-j}k_0} \mathcal{P}_j T_{-M^{-j}k_0} f &= \sum_{k\in\mathbb{Z}^n} \langle T_{-M^{-j}k_0} f, D^j T_k \tilde{\varphi}\rangle T_{M^{-j}k_0} D^j T_k \varphi \\
&= \sum_{k\in\mathbb{Z}^n} \langle f, T_{M^{-j}k_0} D^j T_k \tilde{\varphi}\rangle T_{M^{-j}k_0} D^j T_k \varphi \\
&= \sum_{k\in\mathbb{Z}^n} \langle f, D^j T_{k+k_0} \tilde{\varphi}\rangle D^j T_{k+k_0} \varphi \\
&= \mathcal{P}_j f.
\end{aligned}$$

Since any two representatives of the same coset in $P^{-1}\mathbb{Z}^n/\mathbb{Z}^n$ differ by an element of $\mathbb{Z}^n$ the claim follows. This independence is particularly important for the following proposition.

PROPOSITION 4.3. *Let* $\varphi, \tilde{\varphi} \in \mathbb{E}$. *If* $P \in GL_n(\mathbb{Z})$ *is an admissible oversampling matrix, then for each* $j \in \mathbb{Z}$

$$\mathcal{P}_j^P + \mathcal{Q}_j^P = \mathcal{P}_{j+1}^P.$$

Proof. The condition on the oversampling matrix P guarantees that $\{M\theta_r\}_{r=0}^{p-1}$ is a complete set of coset representatives for $P^{-1}\mathbb{Z}^n/\mathbb{Z}^n$. This fact, together with the previous proposition, implies that for $f \in L^2(\mathbb{R}^n)$

$$\begin{aligned}
\mathcal{P}_j^P f + \mathcal{Q}_j^P f &= \frac{1}{p}\sum_{r=0}^{p-1}\left(T_{M^{-j}\theta_r}\mathcal{P}_j T_{-M^{-j}\theta_r} f + T_{M^{-j}\theta_r}\mathcal{Q}_j T_{-M^{-j}\theta_r} f\right)\\
&= \frac{1}{p}\sum_{r=0}^{p-1} T_{M^{-j}\theta_r}\mathcal{P}_{j+1} T_{-M^{-j}\theta_r} f\\
&= \frac{1}{p}\sum_{r=0}^{p-1} T_{M^{-(j+1)}M\theta_r}\mathcal{P}_{j+1} T_{-M^{-(j+1)}M\theta_r} f\\
&= \frac{1}{p}\sum_{r=0}^{p-1} T_{M^{-(j+1)}\theta_r}\mathcal{P}_{j+1} T_{-M^{-(j+1)}\theta_r} f\\
&= \mathcal{P}_{j+1}^P f. \qquad \square
\end{aligned}$$

We are now in the position to examine the duality of the oversampled systems.

THEOREM 4.4. *Suppose $\varphi, \tilde{\varphi} \in \mathbb{E}$ and $\Psi, \tilde{\Psi} \subset L^2(\mathbb{R}^n)$ are such that* (4.1) *and* (4.2) *hold. If $X(\Psi)$ and $X(\tilde{\Psi})$ are dual frames and $P \in GL_n(\mathbb{Z})$ is an admissible oversampling matrix, then $X^P(\Psi)$ and $X^P(\tilde{\Psi})$ are dual frames with the same bounds as $X(\Psi)$ and $X(\tilde{\Psi})$, respectively. Moreover, for each $f \in L^2(\mathbb{R}^n)$ we have*

$$f = \lim_{j\to\infty} \mathcal{P}_j^P f = \sum_{j\in\mathbb{Z}} \mathcal{Q}_j^P f = f \tag{4.6}$$

and

$$\lim_{j\to-\infty} \|\mathcal{P}_j^P f\| = 0. \tag{4.7}$$

Proof. For this class of scaling functions we use the corresponding properties of $\mathcal{P}_j$ and $\mathcal{Q}_j$ contained in Lemma 4.1. Let $f \in L^2(\mathbb{R}^n)$. Since $\mathcal{P}_j^P$ is the finite sum of translated versions of $\mathcal{P}_j$ we conclude (4.7) directly from Lemma 4.1(c) and Proposition 4.2. By Lemma 4.1(d) we also have that $\mathcal{P}_j f \to f$ in $L^2(\mathbb{R}^n)$ as $j \to \infty$, from which we will obtain the first equality of (4.6) by approximation. Indeed, for each $u \in \mathbb{R}^n$ we have the estimate

$$\begin{aligned}
\|T_{M^{-j}u}\mathcal{P}_j T_{-M^{-j}u} f - f\| &\le \|T_{M^{-j}u} f - f\| + \|T_{M^{-j}u}\mathcal{P}_j T_{-M^{-j}u} f - T_{M^{-j}u} f\|\\
&\le \|T_{M^{-j}u} f - f\| + \|\mathcal{P}_j T_{-M^{-j}u} f - f\|\\
&\le \|T_{M^{-j}u} f - f\| + \|\mathcal{P}_j f - f\| + \|\mathcal{P}_j T_{-M^{-j}u} f - \mathcal{P}_j f\|\\
&\le \|T_{M^{-j}u} f - f\| + \|\mathcal{P}_j f - f\| + C\|T_{-M^{-j}u} f - f\|.
\end{aligned}$$

Each of the three terms in this estimate tend to zero as $j \to \infty$, and thus the first equality of (4.6) follows by summing the above as $u = \theta_r$, $0 \le r \le p-1$.

By Theorem 3.2, we have that $X^P(\Psi)$ and $X^P(\Psi)$ are frames with the same bounds as their respective affine counterparts. Lastly, the second equality of (4.6) follows from a telescoping argument using Proposition 4.3 and the fact that the oversampled systems are Bessel, thereby implying that $X^P(\Psi)$ and $X^P(\tilde{\Psi})$ are indeed dual. $\square$

Remark. We should note that in many cases the assumption in Theorem 4.4 that the original affine systems are actually dual frames may be avoided. For example, in [8] Ron and Shen have derived sufficient conditions (assuming a weak smoothness condition on the refinable family) involving identities of the form (4.2) under which a pair of refinable affine Bessel systems will constitute dual frames.

4.2. Discrete wavelet transform. Throughout this section we assume that $P \in GL_n(\mathbb{Z})$ is an admissible oversampling matrix for the dilation M. Recall that the refinement equations (4.1) can be written in the space domain as

$$\psi_{\ell;j,k} = m^{\frac{1}{2}} \sum_{r\in\mathbb{Z}^n} \alpha_{\ell;r}\varphi_{j+1,r+Mk}$$

for $0 \le \ell \le L$, $j \in \mathbb{Z}$, and $k \in \mathbb{Z}^n$, where $m_\ell(\xi) = \sum_{r\in\mathbb{Z}^n} \alpha_{\ell;r}e^{-i\langle\xi,r\rangle}$. We omit the analogous formulas for the dual functions and filters. We can obtain a similar formula for the oversampled system by observing that

$$\begin{aligned}\psi^P_{\ell;j,k} = \frac{1}{\sqrt{p}}\psi_{\ell;j,P^{-1}k} &= \sqrt{m}\sum_{r\in\mathbb{Z}^n} \alpha_{\ell;r}\frac{1}{\sqrt{p}}\varphi_{j+1,r+MP^{-1}k}\\ &= \sqrt{m}\sum_{r\in\mathbb{Z}^n} \alpha_{\ell;r}\frac{1}{\sqrt{p}}\varphi_{j+1,P^{-1}(Pr+\tilde{M}k)}\\ &= \sqrt{m}\sum_{r\in\mathbb{Z}^n} \alpha_{\ell;r}\varphi^P_{j+1,Pr+\tilde{M}k},\end{aligned}$$

where $\tilde{M} := PMP^{-1}$. Notice that because P is admissible $\tilde{M}$ has integer entries. Letting $\alpha^P_{\ell;r}$ be the coefficient sequence given by

$$\alpha^P_{\ell;r} := \begin{cases}\alpha_{\ell;s}, & r = Ps, s \in \mathbb{Z}^n,\\ 0 & \text{otherwise,}\end{cases}$$

we arrive at

$$\psi^P_{\ell;j,k} = \sqrt{m}\sum_{r\in\mathbb{Z}^n} \alpha^P_{\ell;r}\varphi^P_{j+1,r+\tilde{M}k}.$$

Thus, given $f \in L^2(\mathbb{R}^n)$ the sequence of inner products $\{\langle f, \psi^P_{\ell;j,k}\rangle\}_{k\in\mathbb{Z}^n}$ is given by

$$\langle f, \psi^P_{\ell;j,k}\rangle = \sqrt{m}\sum_{r\in\mathbb{Z}^n}\sum_{r\in\mathbb{Z}^n} \overline{\alpha^P_{\ell;r}}\langle f, \varphi^P_{j+1,r+\tilde{M}k}\rangle$$

for $0 \le \ell \le L$ and $j \in \mathbb{Z}$. To those familiar with subband coding theory, this is immediately recognizable as a convolution followed by a downsampling operation. Note that the downsampling is relative to $\tilde{M}$ rather than M, as for the original affine system. We would like this decomposition to be reversible, meaning that the sequence $\{\langle f, \varphi_{j+1,k}\rangle\}_{k\in\mathbb{Z}^n}$ should be recoverable from the sequences $\{\langle f, \psi_{\ell;j,k}\rangle\}_{k\in\mathbb{Z}^n}$, $0 \le \ell \le L$, by first upsampling each sequence by $\tilde{M}$ and then summing the respective convolutions with the dual filter coefficient sequences $\tilde{\alpha}^P_{\ell;r}$. It is well known that this is equivalent to the coefficient sequences $\alpha^P_{\ell;r}$ and $\tilde{\alpha}^P_{\ell;r}$ satisfying the filter equations (4.2) with $\tilde{M}$ instead of M. Letting m^P_ℓ be defined by

$$m^P_\ell(\xi) = \sum_{r\in\mathbb{Z}^n} \alpha^P_{\ell;r}e^{-i\langle\xi,r\rangle}$$

for $0 \le \ell \le L$, we have $m_\ell^P(\xi) = m_\ell(P^T\xi)$ by the definition of $\alpha_{\ell;r}^p$. The necessary generalized Smith–Barnwell equations are thus

$$\sum_{\ell=0}^{L} \overline{m_\ell^P(\xi)}\tilde{m}_\ell^P(\xi + 2\pi(\tilde{M}^T)^{-1}\tilde{\vartheta}_s) = \delta_{0,s} \tag{4.8}$$

for a.e. $\xi \in \mathbb{T}^n$ and $0 \le s \le m-1$, where $\{\tilde{\vartheta}_s\}_{s=0}^{m-1}$ is a complete set of coset representatives of $\mathbb{Z}^n/\tilde{M}^T\mathbb{Z}^n$ with $\tilde{\vartheta}_s = 0$. In terms of the original filters, (4.8) is equivalent to

$$\sum_{\ell=0}^{L} \overline{m_\ell(\xi)}\tilde{m}_\ell(\xi + 2\pi P^T(\tilde{M}^T)^{-1}\tilde{\vartheta}_s) = \delta_{0,s} \tag{4.9}$$

for $0 \le s \le m-1$, because $m_\ell^P(\xi) = m_\ell(P^T\xi)$. With the following theorem we describe a condition on the oversampling matrix that reduces (4.9) to (4.2), giving a condition under which the dual oversampled affine systems have an associated DWT.

THEOREM 4.5. *Let $P \in GL_n(\mathbb{Z})$ be an admissible oversampling matrix, and assume that P also satisfies $(P^T)^{-1}\mathbb{Z}^n \bigcap (\tilde{M}^T)^{-1}\mathbb{Z}^n = \mathbb{Z}^n$. Then* (4.2) *and* (4.9) *are equivalent.*

Proof. It is sufficient to prove that $\{M^TP^T(\tilde{M}^T)^{-1}\tilde{\vartheta}_s\}_{s=0}^{m-1}$ is a complete set of representatives for $\mathbb{Z}^n/M^T\mathbb{Z}^n$. In other words we need to show only that

$$(M^T)^{-1}\mathbb{Z}^n = \bigcup_{s=0}^{m-1} \left(P^T(\tilde{M}^T)^{-1}\tilde{\vartheta}_s + \mathbb{Z}^n\right).$$

Observing that $\{(\tilde{M}^T)^{-1}\tilde{\vartheta}_s\}_{s=0}^{m-1}$ is a complete set of representatives of $(\tilde{M}^T)^{-1}\mathbb{Z}^n/\mathbb{Z}^n$ our problem is equivalent to showing that if $\{\gamma_s\}_{s=0}^{m-1}$ is a complete set of distinct coset representatives for $(\tilde{M}^T)^{-1}\mathbb{Z}^n/\mathbb{Z}^n$, then $\{P^T\gamma_s\}_{s=0}^{m-1}$ is a compete set of coset representatives for $(M^T)^{-1}\mathbb{Z}^n/\mathbb{Z}^n$.

The first bit of business is to establish that $P^T\gamma_s \in (M^T)^{-1}\mathbb{Z}^n$. This requires for each $x \in \mathbb{Z}^n$ a corresponding $y \in \mathbb{Z}^n$ such that $P^T(\tilde{M}^T)^{-1}x = (M^T)^{-1}y$ or, equivalently, that $M^TP^T(\tilde{M}^T)^{-1}$ has integer entries. Computing this we see

$$M^TP^T(\tilde{M}^T)^{-1} = M^TP^T(P^T)^{-1}(M^T)^{-1}P^T = P^T,$$

which clearly has integer entries.

We now proceed by way of contradiction. Suppose that $\{P^T\gamma_s\}_{s=0}^{m-1}$ is not a complete set of coset representatives. Then $P^T\gamma_{s_0} \in \mathbb{Z}^n$ for some s_0, $1 \le s_0 \le m-1$. Since $\gamma_{s_0} \notin \mathbb{Z}$ this implies that $(P^T)^{-1}\mathbb{Z}^n \bigcap (\tilde{M}^T)^{-1}\mathbb{Z}^n \supsetneq \mathbb{Z}^n$, a contradiction. ∎

4.3. Reconciliation of the hypotheses. In the last subsection we found a relationship between the dilation matrix, M, and the oversampling matrix, P, that is sufficient for the existence of a DWT for the oversampled system. This condition essentially ensures that the perfect reconstruction filter equations for the matrix $\tilde{M} := PMP^{-1}$ are equivalent to those associated with M. It turns out that this condition is automatically satisfied for all admissible oversampling matrices.

THEOREM 4.6. *Let $M, P \in GL_n(\mathbb{Z})$ such that $\tilde{M} := PMP^{-1} \in GL_n(\mathbb{Z})$. Then*

$$P^{-1}\mathbb{Z}^n \bigcap M^{-1}\mathbb{Z}^n = \mathbb{Z}^n \tag{4.10}$$

and

$$(P^T)^{-1}\mathbb{Z}^n \bigcap (\tilde{M}^T)^{-1}\mathbb{Z}^n = \mathbb{Z}^n \tag{4.11}$$

are equivalent.

Proof. We prove the result step by step.

1. By symmetry, it is sufficient to prove (4.10) implies (4.11). Indeed, letting $M' = \tilde{M}^T$ and $P' = P^T$ we have $(\tilde{M}')^T = M$ and $(P')^T = P$.
2. Consider (4.11). If $x \in (\tilde{M}^T)^{-1}\mathbb{Z}^n \bigcap (P^T)^{-1}\mathbb{Z}^n$, then $x = (\tilde{M}^T)^{-1}r = (P^T)^{-1}s$ for some $r, s \in \mathbb{Z}^n$. This allows us to write

$$(\tilde{M}^T)^{-1}\mathbb{Z}^n \bigcap (P^T)^{-1}\mathbb{Z}^n = \left\{(P^T)^{-1}s : s \in \mathbb{Z}^n \text{ and } (P^T)^{-1}M^T s \in \mathbb{Z}^n\right\}.$$

   Letting $S = \{s \in \mathbb{Z}^n : (P^T)^{-1}M^T s \in \mathbb{Z}^n\}$, we have $(\tilde{M}^T)^{-1}\mathbb{Z}^n \bigcap (P^T)^{-1}\mathbb{Z}^n = (P^T)^{-1}S$. Thus, (4.11) is equivalent to $S = P^T\mathbb{Z}^n$.
3. $P^T\mathbb{Z}^n \subseteq S$. *Proof.* Let $s \in P^T\mathbb{Z}^n$, and write $s = P^T x$, $x \in \mathbb{Z}^n$. Then $(P^T)^{-1}M^T P^T x = \tilde{M}^T x \in \mathbb{Z}^n$ because $\tilde{M}$ has integer entries. Hence, $s \in S$.
4. We now provide an unusual characterization of S. It is easy to see that $x \in \mathbb{Z}^n$ if and only if $\langle x, y\rangle \in \mathbb{Z}$ for all $y \in \mathbb{Z}^n$. Thus, $s \in S$ if and only if $\langle (P^T)^{-1}M^T s, y\rangle \in \mathbb{Z}$ for all $y \in \mathbb{Z}^n$. This, in turn, is equivalent to $s \in S$ if and only if $\langle s, My\rangle \in \mathbb{Z}$ for all $y \in P^{-1}\mathbb{Z}^n$.
5. Recall from Proposition 2.1 that (4.10) is equivalent to $\{M\theta_r\}_{r=0}^{p-1}$ being a complete set of coset representatives for $P^{-1}\mathbb{Z}^n/\mathbb{Z}^n$. In other words, (4.10) allows us to write $u \in P^{-1}\mathbb{Z}^n$ as $u = y + Mv$, where $y \in \mathbb{Z}^n$ and $v \in P^{-1}\mathbb{Z}^n$. Notice that if $u \notin \mathbb{Z}^n$, then $v \notin \mathbb{Z}^n$. This will be used below.
6. $S \subseteq P^T\mathbb{Z}^n$. *Proof.* Let $s \in \mathbb{Z}^n$, and suppose that $s \notin P^T\mathbb{Z}^n$. Then there exists $u \in P^{-1}\mathbb{Z}^n \setminus \mathbb{Z}^n$ such that $\langle s, u\rangle \notin \mathbb{Z}$. As explained above, (4.10) allows us to write $u = y + Mv$, where $y \in \mathbb{Z}^n$ and $v \in P^{-1}\mathbb{Z}^n$. Since $u \notin \mathbb{Z}^n$, we must have $v \notin \mathbb{Z}^n$. Then $\langle s, u\rangle = \langle s, y\rangle + \langle s, Mv\rangle \notin \mathbb{Z}$, and since $\langle s, y\rangle \in \mathbb{Z}$ we conclude that $\langle s, Mv\rangle \notin \mathbb{Z}$. Hence, $s \notin S$. ∎

Theorem 4.6 shows that the additional assumption of (4.11) in Theorem 4.5 is redundant and that the dual oversampled affine systems always have an associated DWT if P is admissible.

COROLLARY 4.7. *If $P \in GL_n(\mathbb{Z})$ is an admissible oversampling matrix, then* (4.2) *and* (4.9) *are equivalent.*

5. Discussion of related work. With the number of variations on this theme of bound-preserving oversampling, some comparisons are in order. In particular, we will discuss in detail how our work relates to that of Chui and Shi, Ron and Shen, and Laugesen.

As mentioned in the introduction, the problem of identifying sufficient conditions for the bound-preserving oversampling of affine frames started with Chui and Shi in the one-dimensional setting with dyadic wavelets frames and oversampling by an odd integer [2]. Our proof for Theorem 3.2 is an adaptation of that given in [2] to the n-dimensional case for expansive dilations $M \in GL_n(\mathbb{Z})$ and admissible oversampling matrices P. Chui and Shi improved on their one-dimensional result in [3], extending the second oversampling theorem to expanding dilations $M \in GL_n(\mathbb{Z})$ (i.e., λ an eigenvalue of M implies $|\lambda| > 1$) and $P = pI_n$, with $\gcd(p, \det M) = 1$. This version of the second oversampling theorem also allows for the replacement of the $\mathbb{Z}^n$-translations by $b\mathbb{Z}^n$-translations, where $b > 0$. We will see below, more generally, that this case

actually follows from the $\mathbb{Z}^n$-translation case, which means that Theorem 3.2 implies each of the results of Chui and Shi.

We turn now to the Gramian analysis of Ron and Shen. The version of the second oversampling theorem offered by Ron and Shen in [7] achieves the result of Theorem 3.2, provided that $M \in GL_n(\mathbb{Z})$ is expansive and $P \in GL_n(\mathbb{Z})$ satisfies

$$P^T\mathbb{Z}^n \bigcap (M^T)^j\mathbb{Z}^n = (M^T)^j P^T\mathbb{Z}^n \tag{5.1}$$

for each $j \geq 0$. We will see that this rather complicated expression is actually equivalent to our notion of admissibility. Let us begin by showing that (5.1) implies that P is admissible in terms of Definition 2.2. The $j = 1$ statement of (5.1) says that

$$P^T\mathbb{Z}^n \bigcap M^T\mathbb{Z}^n = M^T P^T\mathbb{Z}^n,$$

which is equivalent to

$$\mathbb{Z}^n \bigcap (P^T)^{-1}M^T\mathbb{Z}^n = (P^T)^{-1}M^T P^T\mathbb{Z}^n =: \tilde{M}^T\mathbb{Z}^n,$$

from which we conclude that $\tilde{M} := PMP^{-1} \in GL_n(\mathbb{Z})$. Moreover, we have

$$P^T\mathbb{Z}^n \bigcap M^T\mathbb{Z}^n = M^T P^T\mathbb{Z}^n \Leftrightarrow (\tilde{M}^T)^{-1}\mathbb{Z}^n \bigcap (P^T)^{-1}\mathbb{Z}^n = \mathbb{Z}^n.$$

But by Theorem 4.6 we have

$$(\tilde{M}^T)^{-1}\mathbb{Z}^n \bigcap (P^T)^{-1}\mathbb{Z}^n = \mathbb{Z}^n \Leftrightarrow M^{-1}\mathbb{Z}^n \bigcap P^{-1}\mathbb{Z}^n = \mathbb{Z}^n,$$

implying that P satisfies our admissibility condition.

For the reverse implication, assume P is admissible according to Definition 2.2. Using the notation of Proposition 2.1 we have that $\{M^j\theta_r\}_{r=0}^{p-1}$ is a complete set of coset representatives for $P^{-1}\mathbb{Z}^n/\mathbb{Z}^n$ for each $j \geq 0$. This is achieved by successively applying the proposition to the collection $\{M^{j-1}\theta_r\}_{r=0}^{p-1}$. Note that since $PMP^{-1} \in GL_n(\mathbb{Z})$, it follows that $PM^jP^{-1} = \tilde{M}^j \in GL_n(\mathbb{Z})$. Thus, Proposition 2.1 implies that

$$M^{-j}\mathbb{Z}^n \bigcap P^{-1}\mathbb{Z}^n = \mathbb{Z}^n.$$

Theorem 4.6 now guarantees

$$M^{-j}\mathbb{Z}^n \bigcap P^{-1}\mathbb{Z}^n = \mathbb{Z}^n \Leftrightarrow (\tilde{M}^T)^{-j}\mathbb{Z}^n \bigcap (P^T)^{-1}\mathbb{Z}^n = \mathbb{Z}^n,$$

but we also have

$$(\tilde{M}^T)^{-j}\mathbb{Z}^n \bigcap (P^T)^{-1}\mathbb{Z}^n = \mathbb{Z}^n \Leftrightarrow (M^T)^j\mathbb{Z}^n \bigcap P^T\mathbb{Z}^n = (M^T)^j P^T\mathbb{Z}^n.$$

This string of equivalences shows that our notion of admissibility is equivalent to (5.1).

Finally, we compare our version of the second oversampling theorem to the one of Laugesen [6]. Laugesen handles two kinds of dilation matrices $M \in GL_n(\mathbb{Z})$: the *expanding* and *amplifying* dilations. The expanding dilations include, but are not limited to, the expansive dilations considered in this work, while the class of amplifying matrices includes such dilations as $\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$. Thus, Laugesen's result applies to a larger class of dilations than Theorem 3.2.

Laugesen also considers translations over the lattice $b\mathbb{Z}^n$, where $b \in GL_n(\mathbb{R})$ commutes with M. This generalization is not essential, as we will now describe using notation as in Theorem 3.2. Let us denote the affine system generated by $\Psi \subset L^2(\mathbb{R}^n)$ relative to translations over $b\mathbb{Z}^n$ by $X_b(\Psi)$ and, similarly, the associated oversampled system $\{D^j T_{bP^{-1}k}\psi_\ell : 1 \le \ell \le L, j \in \mathbb{Z}, k \in \mathbb{Z}^n\}$ by $X_b^P(\Psi)$. For $b \in GL_n(\mathbb{R})$, let D_b be the unitary dilation operator mapping $f \in L^2(\mathbb{R}^n)$ to $D_b f := |\det b|^{\frac{1}{2}} f(b\cdot)$. Suppose that $X_b(\Psi)$ is a frame for $L^2(\mathbb{R}^n)$ with lower bound A and upper bound B. We then have for each $f \in L^2(\mathbb{R}^n)$

$$A\|f\|^2 \le \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}^n} \left|\langle D_b^{-1} f, D^j T_{bk}\psi_\ell\rangle\right|^2 \le B\|f\|^2.$$

Using the fact that b and M commute we obtain for each $f \in L^2(\mathbb{R}^n)$

$$A\|f\|^2 \le \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}^n} \left|\langle f, D^j T_k D_b\psi_\ell\rangle\right|^2 \le B\|f\|^2.$$

This argument shows that $X_b(\Psi)$ is a frame with bounds A, B if and only if $X(D_b\Psi)$ is a frame with the same bounds. If P is admissible, then using Theorem 3.2 we conclude that $X^P(D_b\Psi)$ is a frame with bounds A, B. Thus, for each $f \in L^2(\mathbb{R}^n)$ we have

$$A\|f\|^2 \le \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}^n} \left|\langle D_b f, D^j T_{P^{-1}k} D_b\psi_\ell\rangle\right|^2 \le B\|f\|^2,$$

from which it follows that

$$A\|f\|^2 \le \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}^n} \left|\langle f, D^j T_{bP^{-1}k}\psi_\ell\rangle\right|^2 \le B\|f\|^2.$$

Hence, if M and P satisfy the hypotheses of Theorem 3.2 and $X_b(\Psi)$ is a frame with bounds A, B, then the preceding argument shows that Theorem 3.2 is sufficient to conclude that $X_b^P(\Psi)$ is a frame with the same bounds.

We will now discuss how the notion of admissibility for an oversampling matrix P in [6] is equivalent to ours. Laugesen uses a notion of relative primality for $M, P \in GL_n(\mathbb{Z})$ in which M *is prime relative to* P if $M^T\mathbb{Z}^n \bigcap P^T\mathbb{Z}^n \subseteq M^T P^T\mathbb{Z}^n$. In [6], given a dilation M (which we will assume is expansive), an oversampling matrix $P \in GL_n(\mathbb{Z})$ is admissible if $PMP^{-1} \in GL_n(\mathbb{Z})$ and M is prime relative to P. Observe that

$$M^T\mathbb{Z}^n \bigcap P^T\mathbb{Z}^n \subseteq M^T P^T\mathbb{Z}^n \Leftrightarrow (P^T)^{-1}\mathbb{Z}^n \bigcap (\tilde{M}^T)^{-1}\mathbb{Z}^n \subseteq \mathbb{Z}^n, \tag{5.2}$$

where $\tilde{M} = PMP^{-1}$. On the other hand, since $P^T, \tilde{M}^T \in GL_n(\mathbb{Z})$ we have

$$\mathbb{Z}^n \subseteq (P^T)^{-1}\mathbb{Z}^n \bigcap (\tilde{M}^T)^{-1}\mathbb{Z}^n.$$

It follows that under the hypothesis that $PMP^{-1} \in GL_n(\mathbb{Z})$, M being prime relative to P requires equality rather than containment in (5.2). In light of Theorem 4.6 we see that the notion of admissibility in [6] is equivalent to ours.

Acknowledgments. The author would like to thank Eugenio Hernández, Guido Weiss, and Edward Wilson for their comments and suggestions in the preparation of this paper.

REFERENCES

[1] C. Chui, W. Czaja, M. Maggioni, and G. Weiss, *Characterization of tight-frame wavelets with arbitrary dilation and general tightness preserving oversampling*, J. Fourier Anal. Appl., 8 (2002), pp. 173–200.

[2] C. Chui and X. Shi, *$n\times$ oversampling preserves any tight affine frame for odd n*, Proc. Amer. Math. Soc., 121 (1994), pp. 511–517.

[3] C. K. Chui and X. Shi, *Inequalities on matrix-dilated Littlewood–Paley energy functions and oversampled affine operators*, SIAM J. Math. Anal., 28 (1997), pp. 213–232.

[4] E. Hernández, D. Labate, G. Weiss, and E. Wilson, *Oversampling, quasi-affine frames, and wave packets*, Appl. Comput. Harmon. Anal., to appear.

[5] B. Johnson, *On the relationship between quasi-affine systems and the à trous algorithm*, Collect. Math., 53 (2002), pp. 187–210.

[6] R. Laugesen, *Translational averaging for completeness, characterization, and oversampling of wavelets*, Collect. Math., 53 (2002), pp. 211–249.

[7] A. Ron and Z. Shen, *Affine systems in $L_2(\mathbb{R}^d)$: The analysis of the analysis operator*, J. Funct. Anal., 148 (1997), pp. 408–447.

[8] A. Ron and Z. Shen, *Affine systems in $L_2(\mathbb{R}^d)$: Dual systems*, J. Fourier. Anal. Appl., 3 (1997), pp. 617–637.

SIAM J. MATH. ANAL.
Vol. 35, No. 3, pp. 639–707

NONLINEAR STABILITY OF SEMIDISCRETE SHOCK WAVES*

S. BENZONI-GAVAGE†, P. HUOT‡, AND F. ROUSSET§

Abstract. The orbital stability of possibly large semidiscrete shock waves is considered. These waves are traveling wave solutions of discrete in space and continuous in time systems of conservation laws, which constitute a class of lattice dynamical systems (LDSs). The underlying lattice $\Delta x\mathbb{Z}$ is by nature not invariant by change of frame. Thus semidiscrete shock waves cannot really be transformed into stationary waves, unlike other kinds of approximate shock waves (e.g., viscous or relaxation shocks). This implies that the linearization of the LDS about a given semidiscrete shock wave yields a nonautonomous linear LDS, which cannot be tackled by means of Laplace transform in time. However, viewing the LDS as a finite-difference PDE and performing afterall the change of frame, the profile becomes a stationary solution of the transformed equation. Then, linearizing about the profile, we get an evolution finite-difference PDE in which the spatial operator L, a delayed and advanced differential operator, plays a crucial role in our stability analysis. In particular, we point out an integral formula relating the Green's function of the linearized LDS to the Green's function G_λ of $(\lambda - L)$. Specializing to the upwind scheme, we take advantage of the material introduced in an earlier work [S. Benzoni-Gavage, *J. Dynam. Differential Equations*, 14 (2002), pp. 613–674], in particular of an Evans function, to decompose the Green's function similarly as Zumbrun et al. did for other approximate shock waves. This decomposition relies on explicit representations of the projections involved in the exponential dichotomies and their extensions through the gap lemma. It enables us in turn to prove the orbital stability of the wave, provided that the Evans function D does not vanish in the right half-plane but on the discrete set $2i\pi\sigma\mathbb{Z}$ (σ being the speed of the wave) and that D has a simple root at 0. Additionally, we show that this spectral stability condition is satisfied at least for (extreme) weak shocks. Our spectral stability condition, and more specifically the $D'(0) \neq 0$ part, appears to be a relaxed version of the requirement of Chow, Mallet-Paret, and Shen [*J. Differential Equations*, 149 (1998), pp. 248–291], which we show to be too strong for traveling waves in conservative LDSs.

Key words. lattice dynamical system, traveling wave, Green's function, orbital stability

AMS subject classifications. 35L65, 65M06, 34K06, 37K60

DOI. 10.1137/S0036141002418054

1. Introduction.

Besides numerical analysis, which is concerned with the convergence of numerical schemes, the study of qualitative features of schemes on fixed meshes is of interest to apprehend computer simulations. When the scheme is discrete in space and continuous in time, it is also called a lattice dynamical system (LDS), and when it is fully discrete it is called a coupled map lattice (CML). LDSs and CMLs are also interesting from a purely analytical point of view. They are believed to exhibit richer structures than evolution PDEs, similar to discrete dynamical systems compared to ODEs. The existence and stability of special, traveling wave solutions is still a wide open problem for most LDSs and CMLs. In particular, no direct method is known to tackle the existence of traveling waves in CMLs. A very nice approach was proposed by Chow, Mallet-Paret, and Shen [6], who were able to deduce the existence of discrete traveling waves from the existence of stable enough semidiscrete traveling

*Received by the editors November 18, 2002; accepted for publication (in revised form) February 7, 2003; published electronically October 2, 2003.
http://www.siam.org/journals/sima/35-3/41805.html
†Maply, Université Claude Bernard Lyon I, 21, Avenue Claude Bernard, F-69622 Villeurbanne Cedex, France (benzoni@maply.univ-lyon1.fr).
‡Unite de Mathematiques Pures et Appliquees (CNRS UMR 128), Ecole Normale Superieure-Lyon, 46 Alle d'Italie, F-69364 Lyon cedex 07, France.
§Laboratoire Dieudonné, Universite de Nice – Sophia Antipolis, Parc Valrose, 06108 Nice cedex 02, France (frousset@math.unice.fr).

waves. However, their general theorems were designed for, and basically apply to, *scalar dissipative* equations (in connection with reaction-diffusion equations). Here we are interested in *conservative systems*, having in mind the schemes discretizing hyperbolic systems of conservation laws.

We proved in earlier papers the existence of semidiscrete shock waves of small enough amplitude by means of a center manifold argument. In [2], we dealt with the basic *upwind scheme*, valid when all characteristic speeds are of the same sign. The center manifold theorem used in [1] came from the standard theory of *delay differential equations*. In [3], we were able to generalize our existence result to more general schemes, by first proving a center manifold theorem for *functional differential equations of mixed type* (i.e., with both delay and advance). In the present paper, we are concerned with the time asymptotic stability of these semidiscrete waves. This problem was first addressed in [2], where we constructed an *Evans function*, presumably encoding the linearized stability of the wave, regardless of its amplitude. Our first purpose is to clarify this claim. In particular, we shall make the connection with the approach of Chow, Mallet-Paret, and Shen [6] and show that their stability condition is in fact too restrictive in our context. The stability condition we propose instead is, in terms of the Evans function, the same as the one derived by Zumbrun et al. for viscous shock waves [8, 24, 23] and also relaxation shocks [18]. The drawback of this approach is that, up to the present day, we have been able to construct the Evans function only for the upwind scheme. However, this construction is basically linked to *exponential dichotomies*. On this topic, some recent results obtained simultaneously and independently by Härterich, Sandstede, and Scheel [10] and Mallet-Paret and Verduyn Lunel [17] for delayed and advanced differential operators indicate that it might be possible to define and take advantage of an Evans function for more general schemes. This is why within this paper we consider general schemes whenever it is possible. The other important point is that we aim at dealing with waves of arbitrary amplitude. This makes at least two basic differences with the approach adopted by Bianchini in his nevertheless very interesting related work [5].

Going through the contents, section 2 provides all the background necessary for our analysis. We discuss the assumptions on the scheme, recall the existence result of [3], and prove a corollary giving uniform bounds on the profile and its derivatives. These bounds will in turn enable us to prove the stability of small amplitude profiles (Theorem 3.9) by means of an energy method in the manner of Goodman [9] (also see [13] for fully discrete shocks of rational speed), thus providing an example where our spectral stability condition holds. In section 3, we concentrate on the spectral stability of semidiscrete shocks, making the connection with Chow, Mallet-Paret, and Shen's approach [6]. In particular, we show spectral mapping properties between the bounded operator $\mathcal{R}$ considered in [6] (made out of the solution operator of the LDS linearized about the wave and the shift operator, in a rather natural way as we explain in section 3.1) and the delayed and advanced differential operator L obtained after a not natural (because it does not preserve the lattice) change of frame. The interplay between the two approaches eventually appears to be very useful to gain insight. This should be clear to the reader from the derivation of the Green's function in section 4.1. Before that, we build in section 3.7 all the material necessary to have suitable decompositions of that Green's function. In the meanwhile, we also recall the construction of the Evans function. That part of our work is for the moment restricted to the upwind scheme. However, it is to be noted that the argument given at the end of section 3.7 is valid for all schemes. It shows that Chow, Mallet-Paret, and Shen's

spectral condition fails for *conservative systems*, thus stressing the interest of our alternative approach. Also valid for general schemes is our Theorem 4.2, giving the Green's function of the linearized LDS in terms of a Cauchy integral for the Green's function G_λ of $(\lambda - L)$. From section 4.2, we specialize the upwind scheme, for which we have explicit representations of the projections involved in the decomposition of the Green's function G_λ (Proposition 4.4 and Theorem 4.7). Eventually, we arrive at a decomposition of G_λ (Theorem 4.8) that is analogous to those obtained by Zumbrun et al. in different contexts [24, 23, 18]. This enables us to prove suitable pointwise estimates (Theorem 4.11) on the Green's function of the linearized LDS by choosing appropriates shiftings of the contour. In fact, we make use of simple, pointwise straight contours, as in [23, 18]. Although not surprising, the interesting point is that we recover in those estimates some heat kernels with diffusion coefficients corresponding to the viscosity matrix of the scheme. This gives an a posteriori interpretation of the scheme's properties in terms of the modified (viscous) system of conservation laws. Finally, we deduce from Theorem 4.11 the orbital nonlinear stability in all discrete spaces[1]

$$\mathcal{L}^\alpha(\mathbb{Z};\mathbb{R}^N) \;=\; \left\{ v=(v_j)_{j\in\mathbb{Z}}\,;\sum_{j\in\mathbb{Z}} |v_j|^\alpha \;<\; +\infty \right\}, \quad \alpha \geq 1,$$

with respect to small enough summable perturbations (i.e., in $\mathcal{L}^1$), of semidiscrete shocks that satisfy our spectral assumption. This is our Theorem 5.1, in which the expected rates of decay are also obtained.

2. Semidiscrete conservative systems.

2.1. Background. We consider a conservative semidiscrete scheme

$$\frac{dv_j}{dt} \;+\; \frac{1}{\Delta x}\;\big(\, g(v_{j-p+1},\dots,v_{j+q}) - g(v_{j-p},\dots,v_{j+q-1})\,\big) = 0, \tag{2.1}$$

where p and q are given integers. For the most classical schemes p and q are in fact less than or equal to 1. We are interested in *systems*, and thus the unknowns v_j are vector valued, say, in $\mathbb{R}^N$.

We equivalently look at (2.1) as an LDS operating on sequences $v=(v_j)_{j\in\mathbb{Z}}$,

$$\frac{dv}{dt} \;=\; \mathcal{G}(v), \tag{2.2}$$

where by definition

$$\mathcal{G}(v)_j = G(v_{j-p},\dots,v_{j+q}) := -\tfrac{1}{\Delta x}\,\big(g(v_{j-p+1},\dots,v_{j+q}) - g(v_{j-p},\dots,v_{j+q-1})\big)\,. \tag{2.3}$$

The form of $\mathcal{G}$ in (2.3) actually characterizes finitely supported LDSs that are *conservative*, in the sense that constants are solutions of (2.2) and that solutions associated with summable initial data are *mass preserving* (i.e., $\sum v$ is independent of t). Conversely, it is well known and easy to show that a finitely supported mapping $\mathcal{G}$, vanishing on constants and such that $\sum \mathcal{G}(v) = 0$ for all summable v, is necessarily of the form (2.3).

[1] Our somewhat unusual notation is intended to reserve the notation ℓ^m for left eigenvectors of the system.

We are concerned with the stability of traveling wave solutions of (2.1), approximating shock waves of the underlying continuous system of conservation laws

$$\partial_t u + \partial_x f(u) = 0, \tag{2.4}$$

with

$$f(u) := g(u, \dots, u). \tag{2.5}$$

Those traveling waves are referred to as semidiscrete shock waves. Our motivation is not the convergence of the scheme (2.1) as $\Delta x \to 0$, which is still an open problem in general (see, however, the very interesting recent work by Bianchini [5], dealing with small total variations solutions). Rather, we address the long time behavior of (2.2) for initial data close to a possibly large reference semidiscrete shock. The question is of interest in itself, since very few results are known on the nonlinear stability of traveling waves in vector valued LDSs. Furthermore, part of our motivation was to investigate the possibility of obtaining fully discrete profiles from semidiscrete ones. Indeed, the existence of semidiscrete profiles has been shown in earlier works [1, 3]. But the existence of traveling wave solutions of the fully discrete (explicit) scheme

$$v^{n+1} = \mathcal{H}(v^n) := v^n + \Delta t\, \mathcal{G}(v^n) \tag{2.6}$$

remains an open problem in general. Discrete shock profiles must solve a "boundary value problem"

$$V(j - \eta) = H(V(j-p), \dots, V(j+q)), \qquad V(\pm\infty) = u_\pm, \tag{2.7}$$

where η is the *discrete wave speed* and H is just related to $\mathcal{H}$ by the formula

$$\mathcal{H}(v)_j =: H(v_{j-p}, \dots, v_{j+q}).$$

This problem has received much attention in the last few decades, but, up to now, the existence results are limited to η rational [15] and η *Diophantine* [14]. In other respects, the work of Chow, Mallet-Paret, and Shen [6] has *provided a unified treatment of* —discrete traveling—*waves irrespective of whether their speed is rational or irrational* (sic), showing in particular (Theorem B) that the existence of *spectrally stable* semidiscrete traveling waves implies the existence of fully discrete traveling waves for small enough time step Δt. In this paper, we show that the spectral requirement of Chow, Mallet-Paret, and Shen is in fact too strong for shock waves and propose a way to relax it in terms of an Evans function.

Our main purpose is to show the nonlinear stability, with a shift, of semidiscrete shocks. It is to be noted that this result would directly follow from Theorem A in [6] if their spectral requirement were satisfied.

We introduce below rather standard assumptions. Namely, we assume that g is at least of class $\mathcal{C}^2$. To avoid confusion with left eigenvectors appearing below, we denote by $\mathcal{L}^\alpha(\mathbb{Z};\mathbb{R}^N)$, or simply $\mathcal{L}^\alpha$, the Banach space of sequences with values in $\mathbb{R}^N$ of which the αth power is summable on $\mathbb{Z}$. Then $\mathcal{G}$ obviously maps $\mathcal{L}^\alpha$ into $\mathcal{L}^\alpha$ for all $\alpha \geq 1$ (for this we need only that g be Lipschitz). We point out that $\mathcal{G}$ is of class $\mathcal{C}^r$ if g is so—the "only if" part being also true when considering the smoothness of $\mathcal{G}$ on $\mathcal{L}^1$. Next, we assume that the exact flux f is *strictly hyperbolic* in some connected domain $\mathcal{U}$ of $\mathbb{R}^N$; that is, the Jacobian matrix $A(u) := df(u)$ has N distinct real eigenvalues denoted by

$$a^1(u) < \cdots < a^N(u). \tag{2.8}$$

As usual, we introduce a basis of eigenvectors, $\{r^1(u),\dots,r^N(u)\}$, and the dual basis $\{\ell^1(u),\dots,\ell^N(u)\}$. This means that

$$L(u)\,A(u)\,R(u) \;=\; \operatorname{diag}\,(a^1(u),\dots,a^N(u)) \;=:\; a(u),$$

where

$$L(u) := \begin{pmatrix} -\ell^1(u)- \\ \vdots \\ -\ell^N(u)- \end{pmatrix} \quad \text{and} \quad R(u) := \begin{pmatrix} | & & | \\ r^1(u) & \cdots & r^N(u) \\ | & & | \end{pmatrix}.$$

For later use, we also define $r^{N+j} = r^j$ and $\ell^{N+j} = \ell^j$; that is, the superscripts numbering the eigenvectors are in fact viewed in $\mathbb{Z}/N\mathbb{Z}$. Furthermore, we assume that there is a $k \in \{1,\dots,N\}$ such that the kth characteristic field is *genuinely nonlinear*. We thus choose r^k so that

$$da^k(u)\cdot r^k(u) \;>\; 0 \quad \forall u \in \mathcal{U}. \tag{2.9}$$

The exact system of conservation laws (2.4) is known to admit shock wave solutions of the form $u(x,t) = U(x-\sigma t)$ with U being a step function connecting a "left state" u_- to a "right state" u_+, and the speed σ being related to u_- and u_+ by the Rankine–Hugoniot condition

$$f(u_+) - f(u_-) \;=\; \sigma\,(u_+ - u_-). \tag{2.10}$$

Shock wave solutions are also submitted to an admissibility criterion. A convenient criterion is the one due to Lax [11]. It is relevant at least for shocks of small strength, that is, when the endstates are close enough.

DEFINITION 2.1 (Lax [11]). *A triple (u_-,u_+,σ) is called a k-shock if and only if the Rankine–Hugoniot condition* (2.10) *holds, together with the following inequalities:*

$$\begin{aligned} a^k(u_+) < \sigma < a^k(u_-), \\ a^{k-1}(u_-) < \sigma < a^{k+1}(u_+). \end{aligned} \tag{2.11}$$

The existence of k-shocks (of small strength) is implied by the genuine nonlinearity of the kth characteristic field (see [11] or any textbook on hyperbolic conservation laws).

2.2. Assumptions on the scheme. We proved in [3] the existence of semidiscrete shocks of small strength under two basic assumptions on the scheme, namely dissipativeness and nonresonance, plus a "technical" one. More precisely, we denote for $l \in \{-p+1,\dots,q\}$

$$C^l(u) := \partial_l g(u,\dots,u). \tag{2.12}$$

Note that C^l is matrix valued (in $\mathbb{R}^{N\times N}$) and because of (2.5) we have

$$\sum_{l=-p+1}^{q} C^l(u) \;=\; A(u).$$

We also introduce the so-called viscosity matrix

$$Q(u) := \sum_{l=-p+1}^{q} (1-2l)\,C^l(u). \tag{2.13}$$

(H1) *Dissipativeness.* There exists $\mu > 0$ so that

$$Q(u) \geq \mu I \qquad \forall u \in \mathcal{U}. \tag{2.14}$$

(H2) *Nonresonance.* For all $u \in \mathcal{U}$, $\sigma \neq 0$, and z in $i\mathbb{R} \setminus \{0\}$,

$$\Delta(\sigma; z, u) := \sigma z I + (\mathrm{e}^{-z} - 1) \sum_{l=-p+1}^{q} \mathrm{e}^{zl} C^l(u) \quad \in \mathrm{GL}_N(\mathbb{C}). \tag{2.15}$$

(H3) *Joint reduction.* For all $l \in \{-p+1, \dots, q\}$ and all $u \in \mathcal{U}$,

$$C^l(u) A(u) = A(u) C^l(u). \tag{2.16}$$

These assumptions are discussed in [3]. For the existence of semidiscrete shocks, dissipativeness is required only in the kth characteristic direction. But we need more for stability. This is why we have strengthened the assumption (H1) here. However, the examples given in [3], the generalized Lax–Friedrichs schemes (or Rusanov scheme) and the Godunov scheme away from sonic points, are still valid. More generally, it is to be noted that for any 3-point scheme the assumptions (H1) and (H3) together imply (H2). As a matter of fact, in that case we can adopt the simpler (and more standard) notations

$$D = C^0 \qquad \text{and} \qquad C := -C^1$$

in such a way that

$$Q = C + D \qquad \text{and} \qquad A = D - C.$$

Then the "characteristic matrix" defined in (2.15) just reads

$$\Delta(\sigma; z, u) = \sigma z I + \mathrm{e}^{-z} D(u) + \mathrm{e}^{z} C(u) - Q(u).$$

After simultaneous diagonalization, the coefficients of this matrix are, for $z = i\xi$ with $\xi \in \mathbb{R}$, of the form

$$i(\sigma\xi - a \sin\xi) + q(\cos\xi - 1).$$

This is clearly nonzero for $\sigma \neq 0$ and $q \neq 0$, unless ξ equals zero.

2.3. Existence of semidiscrete shocks. Semidiscrete shocks are by definition traveling wave solutions of (2.1) connecting an endstate u_- at $-\infty$ to another endstate u_+ at $+\infty$. Of course, stationary waves are also solutions of the fully discrete scheme, and they enter the rational framework of Majda and Ralston [15]. Here we consider only nonstationary waves.

DEFINITION 2.2. *A semidiscrete profile associated with a k-shock (u_-, u_+, σ) with $\sigma \neq 0$ is a solution of* (2.1) *of the form*

$$v_j(t) = V(j - st), \qquad s := \sigma/\Delta x,$$

such that

$$v_j(t) \xrightarrow[j \to \pm\infty]{} u_\pm. \tag{2.17}$$

Equivalently, V must solve the "boundary value problem"

$$\sigma V'(x) = g(V(x-p+1),\dots,V(x+q)) - g(V(x-p),\dots,V(x+q-1)), \tag{2.18}$$

$$V(\pm\infty) = u_\pm. \tag{2.19}$$

We recall from [3] the following.

THEOREM 2.3. *Assuming that the flux f is strictly hyperbolic, that the kth characteristic field is genuinely nonlinear, and that the scheme satisfies the requirements in* (H1)–(H3), *there is a $\varepsilon > 0$ such that all k-shocks (u_-, u_+, σ) of strength smaller than ε admit a (one-parameter family of) semidiscrete profiles.*

2.4. Properties of semidiscrete shocks. In addition to the pure existence result in Theorem 2.3, we can show the following, which will be extensively used in the stability analysis.

PROPOSITION 2.4. *In the framework of Theorem* 2.3, *there exists $C > 0$ such that the profiles associated with shocks lying in a small enough neighborhood, say, of size ε, of a nonsonic point satisfy*

$$|V(x) - u_\pm| \le C\varepsilon, \quad |V'(x)| \le C\varepsilon, \quad |V''(x)| \le C\varepsilon^2 \tag{2.20}$$

for all $x \in \mathbb{R}$ and

$$\begin{aligned} &|V'(x+l)| \le C\,|V'(x)| \quad \forall x \in \mathbb{R} \text{ and } l \in \{-p,\dots,q\},\\ &|V'(x)| \quad\ \le C\,\big(a^k(V(x-1)) - a^k(V(x))\big) \quad \forall x \in \mathbb{R}. \end{aligned} \tag{2.21}$$

Furthermore, those profiles are exponentially decaying. There exist $\theta_\pm$ so that $\mp\theta_\pm > 0$ and

$$|V(x) - u_\pm| \le C\,\mathrm{e}^{\theta_\pm x} \tag{2.22}$$

for all $x \in \mathbb{R}$.

The estimates in (2.20) are the same as for viscous profiles. The estimates in (2.21) are designed to deal with the mixing of continuous and discrete derivatives.

Proof. We recall that semidiscrete profiles were found in [3] on a $(N+1)$-dimensional center manifold, containing the N-dimensional vector space of constant solutions of (2.18). It is not difficult to see that the dynamics on that manifold is given by a system of ODEs of the form

$$\begin{cases} y_m' = \ell^m F(y), & m \ne k,\ m \ne N+1,\\ y_k' = y_{N+1} + c\,\ell^k F(y),\\ y_{N+1}' = \ell^k F(y), \end{cases} \tag{2.23}$$

where the ℓ^m are suitably normalized left eigenvectors of $A(u_i)$ and u_i denotes the (nonsonic) bifurcation point, c is a scalar, and the nonlinear term $F(y)$ is such that $F(y_1,\dots,y_N,0) = 0$ (the N-dimensional vector space of fixed points being precisely $\{y_{N+1} = 0\}$). The first estimate in (2.20) comes from the fact that connecting orbits stay globally in the neighborhood of u_i, and the second one follows in a standard way from the profile equation (2.18) and the mean value theorem. The third one is, as for viscous profiles, a little bit more subtle. We first note that, in a neighborhood of 0 of size ε,

$$F(y) = \mathcal{O}(\varepsilon^2).$$

This is due only to the fact that, by definition of F, $F(0) = 0$ and $dF(0) = 0$. Hence, for any (global) solution of (2.23) lying in that neighborhood, we have

$$y'_m = \mathcal{O}(\varepsilon^2) \quad \text{for} \quad m \neq k \qquad \text{and} \quad y'_k = \mathcal{O}(\varepsilon)$$

(uniformly on $\mathbb{R}$). Now, differentiating (2.23), we infer that

$$y''_m = \mathcal{O}(\varepsilon^2) \quad \forall m.$$

Recalling that $V(x)$ can be recovered from its coordinates $y(x)$ through a formula

$$V(x) = \sum_{m=1}^{N} y_m(x)\, r_m(u_i) + h(y(x)),$$

where the function h is such that $h(0) = 0$, $dh(0) = 0$ (by construction of the center manifold), and differentiating twice, we see that $V'' = \mathcal{O}(\varepsilon^2)$ uniformly on $\mathbb{R}$.

The estimates in (2.21) require a finer knowledge of the behavior of profiles at infinity. If u_0 is close to, but different from, u_i, and the corresponding constant solution of (2.18), $V \equiv u_0$, has coordinates $(y_1^0, \dots, y_N^0, 0)$, we have

$$\ell^k \frac{\partial F}{\partial y_{N+1}}\,(y_1^0, \dots, y_N^0, 0) = \vartheta(u_0),$$

where $\vartheta(u_0)$ is the root of $\det\Delta(\sigma, \vartheta, u_0)$ bifurcating from 0. And, as already shown in [3], we have $\vartheta(u_-) > 0$ and $\vartheta(u_+) < 0$ if $u_\pm$ are the endstates of a k-shock close to u_i. Additionally, the corresponding connecting orbits lie on an invariant curve, which crosses the space $\{y_{N+1} = 0\}$ transversally at points $(y_1^\pm, \dots, y_N^\pm, 0)$ (see [3, Lemma 6.2]). This is due to the fact that solutions of (2.18)–(2.19) must solve the integrated equation

$$\sigma\, V(x) - \int_{-1}^{0} g(V(x+\theta-p+1), \dots, V(x+\theta+q))\, d\theta = \sigma\, u_\pm - f(u_\pm). \tag{2.24}$$

Observe that this integration procedure restores hyperbolicity of the endpoints. As a matter of fact, linearizing (2.24) about $V \equiv u_\pm$, we get

$$\sigma\, V(x) - \int_{-1}^{0} \sum_{l=-p+1}^{q} C^l(u_\pm)\, V(x+\theta+l)\, d\theta = 0. \tag{2.25}$$

We obtain the "characteristic matrix" of this equation by looking for exponential solutions, in the same way as the matrix Δ is obtained from the linearized version of the original equation (2.18). Unsurprisingly, we find the matrix $\Delta(\sigma; z, u_\pm)/z$, which is nonsingular for all $z \in i\mathbb{R}$. And in the neighborhood of $z = 0$, there is only one point, $z = \vartheta(u_\pm)$, for which it is singular, its kernel being spanned by $r^k(u_\pm)$. In other words, $x \mapsto r^k(u_\pm)\, \mathrm{e}^{\vartheta(u_\pm)x}$ is a solution of (2.25). Returning to the center manifold, the invariant curve containing connecting orbits can be parametrized by y_{N+1} in the neighborhood of $(y_1^\pm, \dots, y_N^\pm, 0)$ Therefore, those orbits are *scalar-like*, and we are just led to study the behavior of solutions tending to 0 as $x \to \pm\infty$ of scalar ODEs of the form $y'_{N+1} = f^\pm(y_{N+1})$, with $f^\pm(0) = 0$ and $(f^\pm)'(0) = \vartheta(u_\pm)$. An elementary lemma then shows the existence of nonzero constants $c_\pm$ so that

$$y_{N+1}(x) \sim c_\pm\, \mathrm{e}^{\vartheta(u_\pm)x} \quad \text{as} \quad x \to \pm\infty.$$

This implies that $V(x) - u_\pm$ behaves accordingly. Therefore, the observation made above on solutions of (2.25) implies that

$$V(x) - u_\pm \sim c\, r^k(u_\pm)\, \mathrm{e}^{\vartheta(u_\pm)x}, \quad x \to \pm\infty,$$

for some nonzero constant c. In particular, this implies (2.22). Besides this asymptotic behavior, we also know that V' does not vanish, since solutions of (2.23) do not reach any fixed point in finite time. Combining these two facts, we get the first estimate in (2.21). Next, we observe that, for ε small enough, y_k' vanishes only at fixed points (characterized by $y_{N+1} = 0$). Furthermore, (2.23) implies that

$$da^k(V) \cdot V' = y_{N+1}\,(\, da^k(u_i) \cdot r^k(u_i) + \mathcal{O}(\varepsilon)\,)$$

does not vanish either, for ε small enough, because of (2.9). By (2.11), this means that $a^k \circ V$ is monotonically decaying. Using the mean value theorem and the previous observations, we get the second estimate in (2.21). ∎

3. Spectral stability of semidiscrete shocks. There are basically two approaches to the spectral stability of a semidiscrete traveling wave. One way was proposed by Chow, Mallet-Paret, and Shen [6]. They exhibited the linear operator, which we denote by $\mathcal{R}$ below, that encodes the spectral stability of rather general semidiscrete traveling waves. Some features of their approach are presented in section 3.1. However, their abstract theorems are intended to be applied to equations of reaction-diffusion type. We show in section 3.5 below that their spectral condition fails for conservative LDSs. The other possible approach seems more classical. It consists of forcing the wave to be stationary through a change of frame, as usual in traveling wave analysis. This approach is at first glance questionable since the lattice is not preserved by the change of frame. But it will appear to be complementary to the first one and especially useful to gain insight.

3.1. The LDS point of view. From now on, we fix a semidiscrete shock, denoted by $u_j(t) = U(j - st)$, in order to avoid confusion with general solutions of (2.1). Unless otherwise specified, it is not necessarily of small amplitude.

Without loss of generality, we assume that s, or equivalently σ, is positive. We introduce

$$T := 1/s = \Delta x/\sigma,$$

which is the time taken for the shock to travel across one mesh.[2] This is characteristic of all waves traveling with speed σ, as it can be formalized using the *shift* operator. If $\mathcal{T}$ is defined by

$$(\mathcal{T} \cdot v)_j := v_{j-1}$$

for any sequence $v = (v_j)_{j\in\mathbb{Z}}$, then traveling waves of speed σ are characterized by

$$v(t+T) = \mathcal{T} \cdot v(t). \tag{3.1}$$

As a matter of fact, (3.1) obviously holds if $v_j(t) = V(j - st)$ and, conversely, if (3.1) holds, then, defining $V(x) := v_0(-Tx)$, we have by induction

$$v_j(t) = V(j - st)$$

for all $j \in \mathbb{Z}$ and all t. This simple remark is nonetheless crucial in what follows.

[2]From place to place, we shall do a rescaling in order to have $\Delta x = 1$ and thus simplify the writing. But it is worth keeping in mind the dimensionality of the various quantities: T (time), s (frequency), Δx (distance), σ (speed).

Other important observations are that $\mathcal{T}$ is an isometry in all the spaces $\mathcal{L}^\alpha$ and that $\mathcal{T}$ commutes with the nonlinear mapping $\mathcal{G}$, that is,

$$\mathcal{T}\,\mathcal{G} \;=\; \mathcal{G}\,\mathcal{T}.$$

Following along the lines of [6], a first step in the stability analysis of the special solution u is to linearize (2.2) about u. This yields the *nonautonomous* linear LDS

$$\frac{dv}{dt} \;=\; D\mathcal{G}(u)\cdot v. \tag{3.2}$$

We denote by $\mathcal{S}$ the solution operator of (3.2). By definition, for all solutions v of (3.2) and all times $t \geq t_0$ we have

$$v(t) \;=\; \mathcal{S}(t,t_0)\cdot v(t_0).$$

Differentiating once the relation $du/dt \;=\; \mathcal{G}(u)$ and applying (3.1) to $v = u'$ we easily see that

$$\mathcal{T}\cdot u'(0) \;=\; \mathcal{S}(T,0)\cdot u'(0).$$

This means that $u'(0)$ is an eigenvector of the operator

$$\mathcal{R} \;:=\; \mathcal{T}^{-1}\,\mathcal{S}(T,0) \tag{3.3}$$

for the eigenvalue 1. We point out that since U is exponentially decaying (see section 2.4), this holds true in any space $\mathcal{L}^\alpha$, $1 \leq \alpha \leq \infty$.

Another simple remark makes use of the following properties.

PROPOSITION 3.1 (Chow, Mallet-Paret, and Shen [6]). *The shift operator $\mathcal{T}$ and the solution operator $\mathcal{S}$ of* (3.2) *satisfy the identities*

$$\mathcal{S}(t+T,T) \;=\; \mathcal{T}\,\mathcal{S}(t,0)\,\mathcal{T}^{-1} \qquad \forall t \geq 0, \tag{3.4}$$

$$(\,\mathcal{T}^{-1}\,\mathcal{S}(T,0)\,)^n \;=\; \mathcal{T}^{-n}\,\mathcal{S}(nT,0). \tag{3.5}$$

Proof. Regarding (3.4), it is sufficient to show that both sides applied to any sequence satisfy the linear LDS

$$\frac{dv}{dt} \;=\; D\mathcal{G}(\mathcal{T}u)\cdot v$$

and coincide at $t = 0$. And (3.5) is proved by induction, using (3.4). ∎

Equation (3.5) also reads

$$\mathcal{R}^n \;=\; \mathcal{T}^{-n}\,\mathcal{S}(nT,0).$$

Now assume that the operator $\mathcal{R}$ has an eigenvalue ζ of modulus greater than 1. There is a sequence w so that $\mathcal{R}w \;=\; \zeta\,w$. Then we have by (3.5)

$$\mathcal{S}(nT,0)\cdot w \;=\; \zeta^n\,\mathcal{T}^n\cdot w,$$

which shows that the solution $v(t) \;= \mathcal{S}(t,0)\cdot w$ of (3.2) blows up in all spaces $\mathcal{L}^\alpha$ as t tends to infinity.

Chow, Mallet-Paret, and Shen have proved in [6] the following nonlinear result, which shows that the operator $\mathcal{R}$ fully encodes the stability of the semidiscrete traveling wave u.

THEOREM 3.2 (Chow, Mallet-Paret, and Shen [6]). *If a traveling wave solution u of* (2.2) *satisfies the following spectral stability conditions, in terms of the $\mathcal{R}$ defined in* (3.3) *acting on $\mathcal{L}^\infty$,*

(i) *the spectrum of* $\mathcal{R}$ *is*

$$\Sigma(\mathcal{R}) \subset \{\zeta;\ |\zeta| < 1\} \cup \{1\} \tag{3.6}$$

and

(ii) *the eigenvalue* 1 *is simple,*

then u *is orbitally stable in* $\mathcal{L}^\infty$.

Chow, Mallet-Paret, and Shen have applied their theorem to a semidiscrete (scalar) Nagumo equation. In what follows we investigate (i) and (ii) in the context of conservative systems.

We shall show that (i) holds at least in two cases, namely, for scalar shocks and for shocks of small enough amplitude. The proof will proceed in several steps. First of all, we show in section 3.4 that the location of the essential spectrum of $\mathcal{R}$ can be determined in a rather standard way by Fourier analysis. Using an observation of Chow, Mallet-Paret, and Shen [6], the essential spectrum of $\mathcal{R}$ is seen to coincide with the essential spectrum of a modified operator $\mathcal{R}_0$ associated with a "fake profile." Together with some Fourier analysis of the autonomous systems about the endstates, this enables us to prove that (i) holds for the essential spectrum of $\mathcal{R}$. With regard to the point spectrum of $\mathcal{R}$, we observe that it is tightly related to the point spectrum of a differential operator with delay and advance. This operator L appears naturally when we look at (2.1) as a delayed and advanced PDE and perform a change of frame that makes the wave stationary; see section 3.2. We show that any nonzero eigenvalue of $\mathcal{R}$ reads $\exp(\Lambda/\sigma)$ with λ an eigenvalue of L and that their multiplicities coincide. Constructing an Evans function for L [2] then yields a necessary condition for the point spectrum of $\mathcal{R}$ to lie in the unit disk. In other respects, for shocks of small strength, we can perform some energy estimates on the LDS (2.1), in the spirit of the work of Goodman on viscous shocks [9]. This enables us to show that the point spectrum of $\mathcal{R}$ does lie in the unit disk.

But we shall also show that (ii) must fail. In fact, this is reminiscent of the viscous shock wave analysis. Computing the local behavior of the Evans function about 0 yields a sufficient condition for (ii) to hold true in $\mathcal{L}^2$. This condition is the same as for viscous shocks and is satisfied for shocks of small strength. But we show that, like for viscous shocks, there are too many bounded solutions of the equation $L \cdot Y = 0$. Hence condition (ii) does not hold in $\mathcal{L}^\infty$, contrary to what is required by Chow, Mallet-Paret, and Shen.

3.2. The change of frame point of view. The other point of view alluded to above was already used in [2]. It starts from the following (formal) observation. Solutions of the LDS (2.1) may be regarded as restrictions to $\mathbb{Z} \times \mathbb{R}^+$ of solutions of a "finite difference-PDE." Indeed, rewriting $v_j(t) = V(j\Delta x, t)$ as a solution of (2.1), we have

$$\begin{aligned} \frac{\partial V}{\partial t} + \frac{1}{\Delta x}\ & (g(V(x-(p-1)\Delta x,t),\dots,V(x+q\Delta x,t)) \\ & - g(V(x-p\Delta x,t),\dots,V(x+(q-1)\Delta x,t))) = 0. \end{aligned} \tag{3.7}$$

Introducing the change of variables $(\widetilde{x} := (x - \sigma t)/\Delta x, \widetilde{t} := t/\Delta x)$, $V(x,t) = \widetilde{V}(\widetilde{x},\widetilde{t})$ is a solution of (3.7) if and only if $\widetilde{V}$ solves

$$\frac{\partial \widetilde{V}}{\partial \widetilde{t}} - \sigma \frac{\partial \widetilde{V}}{\partial \widetilde{x}} + g(\widetilde{V}(\widetilde{x}-p+1,\widetilde{t}),\dots,\widetilde{V}(\widetilde{x}+q,\widetilde{t})) - g(\widetilde{V}(\widetilde{x}-p,\widetilde{t}),\dots,\widetilde{V}(\widetilde{x}+q-1,\widetilde{t})) = 0.$$

Dropping the tildes, this equation reads

$$\partial_t V - \sigma\,\partial_x V + g(V(x-p+1,t),\dots,V(x+q,t)) - g(V(x-p,t),\dots,V(x+q-1,t)) = 0, \tag{3.8}$$

of which semidiscrete profiles, solving (2.18), are obviously stationary solutions. Linearizing (3.8) about a stationary solution U yields the linear finite-difference PDE

$$\partial_t V - L\cdot V = 0, \tag{3.9}$$

where L is the spatial operator defined by

$$(L\cdot V)(x) = \sigma\,V'(x) - \sum_{l=-p}^{q} (\partial_l g(\,U(x-p+1),\dots,U(x+q)\,) - \partial_{l+1} g(\,U(x-p),\dots,U(x+q-1)\,)\cdot V(x+l). \tag{3.10}$$

We have adopted here the convention that

$$\partial_{-p} g \equiv 0 \qquad \text{and} \qquad \partial_{q+1} g \equiv 0.$$

To facilitate the reading, we shall use more compact notations, just writing

$$(L\cdot V)(x) = \sigma\,V'(x) - \sum_{l=-p}^{q} (\,C^l(x) - C^{l+1}(x-1)\,)\cdot V(x+l)$$

with

$$C^l(x) := \partial_l g(\,U(x-p+1),\dots,U(x+q)\,).$$

This is an abuse of notation since the matrices C^l were originally defined as functions of a (single) state $u\in\mathbb{R}^N$. Consistently, $C^l(\pm\infty)$ will stand for $C^l(u_\pm)$ and will merely be denoted by $C^l_\pm$.

We see on its definition (3.10) that L is a closed unbounded operator on $L^2(\mathbb{R})$, with dense domain H^1 in L^2. An easy preliminary result is the following.

PROPOSITION 3.3. *The operator L defined by* (3.10) *is the infinitesimal generator of a strongly continuous semigroup on $L^2(\mathbb{R})$.*

Proof. This follows from a consequence of the Hille–Yosida theorem. Since L is closed with dense domain, it is sufficient to prove that the resolvent set of L contains a ray $\{\lambda\in(M,+\infty)\}$ and that the resolvent enjoys the estimate

$$\|\,(\lambda - L\,)^{-1}\| \le \frac{1}{\lambda - M}$$

for $\lambda > M$ (see [19, p. 12]).[3] So assume that $(\lambda - L)\cdot V = W$. Taking the inner product with V gives the equality

$$\lambda\,\|V\|^2_{L^2(\mathbb{R})} + \int_{\mathbb{R}} V(x)\sum_{l=-p}^{q}(C^l(x) - C^{l+1}(x-1))\,dx = \langle\,V\,,\,W\,\rangle_{L^2(\mathbb{R})}.$$

[3] In fact, this result follows from our more precise Lemma 4.1 on the Green's function of $\lambda - L$; this is why we just show the estimate of the resolvent here.

Then by Cauchy–Schwarz inequality and the invariance of the L^2 norm under translation we get

$$(\,\lambda\,-\,2(p+q+1)\,\gamma\,)\,\|V\|_{L^2(\mathbb{R})}\,\le\,\|W\|_{L^2(\mathbb{R})},$$

where

$$\gamma\,:=\,\max\{\,\|\,C^l(x)\,\|_{\mathbb{R}^{N\times N}}\,;\;l\in\{-p+1,\dots,q\}\,,\;x\in\mathbb{R}\,\}.\qquad\square$$

In fact, a slightly more general energy estimate than in the above proof shows that any complex number λ satisfying $\mathrm{Re}\lambda > 2(p+q+1)\,\gamma$ belongs to the resolvent set of L. Of course, a natural spectral stability condition for the stationary solution U is that the whole open right half-plane lies in the resolvent set of L. This actually holds true simultaneously with the spectral condition (i) of Theorem 3.2 on the operator $\mathcal{R}$, due to the spectral mapping results that we show in the next subsections.

3.3. Link between the two points of view. It is clear that solutions V of (3.9) are related to solutions of (3.2) through the equality

$$v_j(t)\,=\,V(j\,-\,\sigma t/\Delta x\,,\,t/\Delta x\,).$$

In particular, at $t = T = \Delta x/\sigma$ we find that

$$v_j(T)\,=\,V(j-1,1/\sigma).$$

We can interpret this relation in terms of operators. Denoting by e^{tL} the semigroup associated with L and by P the mapping

$$\begin{array}{rccc} P\,: & \mathcal{C}(\mathbb{R}) & \to & \mathcal{L}^\infty \\ & V & \mapsto & (V(j))_{j\in\mathbb{Z}}, \end{array}$$

the previous equality means that

$$\mathcal{R}P\,=\,P\,\mathrm{e}^{\sigma^{-1}L}$$

on the domain ($H^1\subset\mathcal{C}(\mathbb{R})$) of L. This suggests that the spectrum of $\mathcal{R}$ is related to that of L. Even though P is obviously not invertible, we can prove some spectral mapping results. We study separately the essential spectrum and the point spectrum (i.e., the set of eigenvalues).

For clarity, we recall that the essential spectrum of an operator L can be defined by

$$\Sigma_{\mathrm{ess}}(L)\,=\,\{\,\lambda\in\mathbb{C}\,;\;(L\,-\,\lambda)\;\text{is }\textit{not}\text{ Fredholm index }0\,\}. \tag{3.11}$$

An alternative definition is that λ belongs to $\Sigma_{\mathrm{ess}}(L)$ if and only if λ belongs to the spectrum of $L+K$ for any compact operator K. Both definitions are equivalent for closed operators (see [22, p. 15]). With the first definition, it is clear that the whole spectrum $\Sigma(L)$ consists of the union of $\Sigma_{\mathrm{ess}}(L)$ and the set of eigenvalues.

3.4. Essential spectrum.

LEMMA 3.4. *The essential spectrum of the operator $\mathcal{R}$ defined in* (3.3) *and the essential spectrum of the operator L defined in* (3.10) *are such that*

$$\begin{aligned} &\Lambda(u_\pm)\,\subset\,\Sigma_{ess}(L)\,\subset\,\bigcup_{u\in\mathcal{U}}\Lambda(u)\,=:\,\Lambda,\\ &\Lambda(u)\,:=\,\{\,\lambda\in\mathbb{C}\,;\;\exists\xi\in\mathbb{R},\;\det\,(\,\Delta(\sigma;i\xi,u)\,-\,\lambda\,)\,=\,0\,\}, \end{aligned} \tag{3.12}$$

where the matrix Δ is defined as in (2.15)*, and*

$$\mathrm{e}^{\sigma^{-1}\Lambda(u_\pm)} \subset \Sigma_{ess}(\mathcal{R})\backslash\{0\} \subset \mathrm{e}^{\sigma^{-1}\Lambda}. \tag{3.13}$$

Proof. We begin with (3.12). The basic step consists of showing that $\Lambda(u_\pm)$ is exactly the spectrum of the asymptotic operator, $L_\pm$, defined by

$$(L_\pm \cdot V)(x) = \sigma V'(x) + \sum_{l=-p}^{q} (C_\pm^{l+1} - C_\pm^l) \cdot V(x+l). \tag{3.14}$$

Using Fourier transform, we readily see that λ belongs to $\Sigma(L_\pm)$ if and only if there exists $\xi \in \mathbb{R}$ so that the characteristic equation

$$\det(\Delta(\sigma; i\xi, u_\pm) - \lambda) = 0 \tag{3.15}$$

holds. This is precisely the definition of $\Lambda(u_\pm)$. Hence $\Sigma(L_\pm) = \Lambda(u_\pm)$. In fact, this is pure essential spectrum, as the following argument shows when applied to $L_\pm$ instead of L. For λ belongs to $\Sigma_{\mathrm{ess}}(L)$ as soon as there exists a sequence $(V^n)_{n\in\mathbb{N}}$, with $V^n \in H^1$ (the domain of L), which is orthonormal in L^2 and such that $(L-\lambda)V^n$ tends to 0 as n goes to ∞ (this easily implies that λ belongs to the spectrum of $L+K$ for any compact operator K). Now, (3.15) implies the existence of a vector r such that the function $V : x \mapsto \mathrm{e}^{i\xi x} r$ is an approximate eigenfunction of $L_\pm$; i.e., $(L_\pm - \lambda)V = 0$, but V does not belong to the domain of $L_\pm$. To construct a sequence V^n that fulfills the above requirements, it suffices to choose suitably supported $\mathcal{C}^\infty$ functions χ^n and take $V^n = \chi^n V/\|\chi^n V\|$. This is rather classical. The reader may check that a possible choice is

$$\begin{cases} \chi^n(x) = 1, & |x \mp 2n^2| \le n, \\ \chi^n(x) = \chi^1(x), & n \le |x \mp 2n^2| \le n+1, \\ \chi^n(x) = 0, & |x \mp 2n^2| \ge n+1. \end{cases}$$

This works because the supports of the χ^n do not intersect and the L^2 norms of $\chi^n V$ are $\mathcal{O}(n^{1/2})$, while the L^2 norms of $(L-\lambda)(\chi^n V)$ remain bounded, since $(L_\pm - \lambda)V = 0$ and the coefficients of L are converging rapidly enough (namely, exponentially fast) to the coefficients of $L_\pm$ at $\pm\infty$.

Hence all points of $\Sigma(L_\pm) = \Lambda(u_\pm)$ belong to $\Sigma_{\mathrm{ess}}(L)$. The other inclusion in (3.12) requires a tougher piece of analysis, for which we refer to a work by Mallet-Paret [16]. For if λ does not belong to Λ, in particular it is not in $\Sigma(L_\pm)$, and then the variational operator $(L-\lambda)$ is asymptotically hyperbolic according to the definition of Mallet-Paret. Therefore, by Theorems A and B in [16], $(L-\lambda)$ is Fredholm and its index depends only on $L_\pm$. The computation of the index is based on the spectrum flow formula (Theorem C in [16]), considering a homotopy connecting L_- to L_+. The simplest one is given by L_θ, the constant coefficient operator about $u(\theta)$, where $\theta \mapsto u(\theta)$ is any smooth curve between u_- and u_+. If $\lambda \notin \Lambda$, then, for all θ, $(L_\theta - \lambda)$ is hyperbolic in the sense of Mallet-Paret. Consequently, we find that the index of $(L-\lambda)$ equals 0, which precisely means that λ is not in the essential spectrum of L. This completes the proof of (3.12).

Regarding $\mathcal{R}$, a similar procedure works. It is to some extent easier due to the boundedness of $\mathcal{R}$. We first look at the spectrum of $\mathcal{R}_\pm = \mathcal{T}^{-1}\mathcal{S}_\pm(T)$, with $\mathcal{S}_\pm(t)$ the solution operator of the autonomous LDS

$$\frac{dv}{dt} = D\mathcal{G}(u_\pm)\cdot v. \tag{3.16}$$

By a slight abuse of notation, $u_\pm$ here stands for the constant sequence $(u_\pm)$. Pointwisely, (3.16) reads

$$\frac{dv_j}{dt} + \frac{1}{\Delta x}\sum_{l=-p}^{q} (C_\pm^l - C_\pm^{l+1})\cdot v_{j+l} = 0. \tag{3.17}$$

Equation (3.17) can be solved through discrete Fourier transform. To be precise, we fix some notations. Any slowly growing sequence w (for instance, bounded) can be associated with a tempered distribution of period 2π

$$\mathcal{F}w := \sum_{j\in\mathbb{Z}} w_j\, e^{-ij\xi}.$$

And, conversely, a tempered distribution g of period 2π is associated with the slowly growing sequence $\overline{\mathcal{F}}g$ (of its Fourier coefficients) defined by

$$\big(\overline{\mathcal{F}}g\big)_j := \frac{1}{2\pi}\langle\, g\,,\, e^{ij\xi}\,\rangle_{\mathbb{R}/2\pi\mathbb{Z}}.$$

Of course, we have $\overline{\mathcal{F}}\mathcal{F} = I$ and $\mathcal{F}\overline{\mathcal{F}} = I$ on the corresponding spaces.

Using that $\mathcal{F}\mathcal{T}^{-l} = \mathrm{e}^{il\xi}\mathcal{F}$ for all l, we find that the solution operator of (3.16) is given by

$$\mathcal{S}_\pm(t) = \overline{\mathcal{F}}\exp\left(\frac{t}{\Delta x}\sum_{l=-p}^{q}\mathrm{e}^{il\xi}(C_\pm^{l+1} - C_\pm^l)\right)\mathcal{F}.$$

Therefore, for all v and w in $\mathcal{L}^\alpha(\mathbb{Z})$, the relation

$$(\mathcal{R}_\pm - \zeta)\cdot w = v \tag{3.18}$$

is by definition of $\mathcal{R}_\pm$ equivalent to

$$\exp\left(i\xi + \frac{1}{\sigma}\sum_{l=-p}^{q}\mathrm{e}^{il\xi}(C_\pm^{l+1} - C_\pm^l)\right)\cdot\mathcal{F}w - \zeta\,\mathcal{F}w = \mathcal{F}v.$$

Rewriting this with $\zeta = e^{\lambda/\sigma}$ and using the matrix Δ introduced in (2.15), we obtain that (3.18) is equivalent to

$$\zeta\, B_\pm(i\xi,\lambda)\,\mathcal{F}w = \mathcal{F}v, \qquad B_\pm(i\xi,\lambda) := \mathrm{e}^{(\,\Delta(\sigma;i\xi,u_\pm) - \lambda)/\sigma} - I.$$

The special form of Δ and Lyapunov's theorem show that the matrix $B_\pm(i\xi,\lambda)$ is singular for $\xi\in\mathbb{R}$ if and only if $(\,\Delta(\sigma;i\widetilde{\xi},u_\pm) - \lambda\,)$ is singular for some $\widetilde{\xi}\equiv\xi[2\pi]$. Equivalently, $(\,\Delta(\sigma;i\xi,u_\pm) - \widetilde{\lambda}\,)$ is singular for some $\widetilde{\lambda}\equiv\lambda[2i\pi\sigma]$; that is, ζ belongs to $\exp(\Lambda(u_\pm)/\sigma)$.

This implies that

$$\Sigma(\mathcal{R}_\pm)\backslash\{0\} = \mathrm{e}^{\sigma^{-1}\Lambda(u_\pm)}.$$

As a matter of fact, elements of $\exp(\Lambda(u_\pm)/\sigma)$ are obviously eigenvalues of $\mathcal{R}_\pm$ (here the eigenvectors are genuine ones in $\mathcal{L}^\infty$). Conversely, if ζ is not in $\exp(\Lambda(u_\pm)/\sigma)$,

according to an observation already made by Rustichini [21], there exists $\gamma > 0$ such that $(\,\Delta(\sigma; z, u_\pm) - \lambda\,)$ is not singular in the strip $\{\,|\mathrm{Re} z| \le \gamma\,\}$. Indeed, the roots of the holomorphic function

$$z \mapsto \det(\,\Delta(\sigma; z, u_\pm) - \lambda\,)$$

have their imaginary parts bounded in terms of their real parts. Hence, $B_\pm(z, \lambda)$ is nonsingular for $|\mathrm{Re} z| \le \gamma$. As a consequence (see, for instance, Lemma 9 in [2]), $\varphi^\pm := \overline{\mathcal{F}}(\,B_\pm(i\xi, \lambda)\,)$ is exponentially decaying at infinity (with rate $\beta < \gamma$). Therefore, $(3.18)_\pm$ is equivalent to

$$w = \varphi^\pm * v, \quad \text{i.e.,} \quad w_j = \sum_{k\in\mathbb{Z}} \varphi^\pm_{j-k}\, v_k \quad \forall j \in \mathbb{Z},$$

and we have a constant C so that

$$\|w\|_\infty \le C\, \|v\|_\infty.$$

This means that ζ belongs to the resolvent set of $\mathcal{R}_\pm$.

Now we make an observation that is useful to get information on the essential spectrum of $\mathcal{R}$. Considering the simpler operator $\mathcal{R}_0 = \mathcal{T}^{-1}\,\mathcal{S}_0(T)$, where $\mathcal{S}_0$ is the solution operator of the autonomous equation

$$\frac{dv}{dt} = D\mathcal{G}(u^0)\cdot v \tag{3.19}$$

associated with the (stationary) fake profile defined by $u^0_j := u_-$, $j < 0$, $u^0_j := u_+$, $j \ge 0$, it is known from [6, Lemma 7.2] that $\mathcal{R} - \mathcal{R}_0$ is compact. Therefore, $\Sigma_{\mathrm{ess}}(\mathcal{R}) = \Sigma_{\mathrm{ess}}(\mathcal{R}_0)$.

So (3.13) is equivalent to

$$\mathrm{e}^{\sigma^{-1}\,\Lambda(u_\pm)} \subset \Sigma_{\mathrm{ess}}(\mathcal{R}_0)\backslash\{0\} \subset \mathrm{e}^{\sigma^{-1}\,\Lambda}.$$

To show that $\zeta \in \mathrm{e}^{\sigma^{-1}\Lambda(u_\pm)}$ belongs to $\Sigma_{\mathrm{ess}}(\mathcal{R}_0)$, it is sufficient to find a sequence (w^n) with no limit point and $\|w^n\|_{\mathcal{L}^\infty} = 1$ such that $(\mathcal{R}_0 - \zeta)\, w^n$ tends to 0. (This is the same argument as the one used for L, except that we cannot use orthonormal sequences here.) On the contrary, if $\zeta \notin \mathrm{e}^{\sigma^{-1}\Lambda}$, we can show that $(\mathcal{R}_0 - \zeta)$ is Fredhlom index 0. The proof relies on pointwise estimates for the resolvent of both $\mathcal{R}_0$ and its *antiadjoint* $\mathcal{R}_0^*$. By this we mean, doing a slight abuse of notation, that $\mathcal{R}_0$ is the adjoint of $\mathcal{R}_0^*$. The latter is in fact related to the adjoint LDS

$$\frac{dz}{dt} = -\,(\,D\mathcal{G}(u^0)\,)^*\cdot z \tag{3.20}$$

through the formula $\mathcal{R}_0^* := \mathcal{S}_0^*(-T)\,\mathcal{T}$, where $\mathcal{S}_0^*$ is the solution operator on $\mathcal{L}^1(\mathbb{Z})$ of (3.20). For completeness, we recall those pointwise estimates in Proposition 3.5. That they imply $\mathcal{R}_0$ is Fredholm is a standard exercise. The calculation of the index uses again a homotopy argument. For more details, the reader may refer to [2, Section 4]. □

PROPOSITION 3.5. *For $\zeta \notin \mathrm{e}^{\sigma^{-1}\Lambda(u_-)} \bigcup \mathrm{e}^{\sigma^{-1}\Lambda(u_+)}$, there exists $C_0 > 0$ and $\beta > 0$ such that for $w \in \mathcal{L}^\infty$,*

$$|\,w_j\,| \le C_0\ (\,\|(\mathcal{R}_0 - \zeta)\cdot w\|_\infty + \ e^{-\,\beta\,|j|}\,\|w\|_\infty\,) \quad \forall j \in \mathbb{Z}, \tag{3.21}$$

and for $z \in \mathcal{L}^1$,

$$|\, z_j \,| \,\le\, C_0 \;(\, \|(\mathcal{R}_0^* - \bar{\zeta}) \cdot z\|_1 \,+\, e^{-\,\beta\,|j|}\, \|z\|_1 \,) \quad \forall j \in \mathbb{Z}. \tag{3.22}$$

Proof. The proof relies on Duhamel's formula

$$\mathcal{S}_0(t) \cdot w \,=\, \mathcal{S}_+(t) \cdot w \,+\, \int_0^t \mathcal{S}_+(t-\tau) \cdot (\,(D\mathcal{G}(u^0) \,-\, D\mathcal{G}(u_+))\mathcal{S}_0(\tau) \cdot w)\, d\tau \tag{3.23}$$

and convolution estimates. The rate β is imposed by the width of the vertical strip in which $(\,\Delta(\sigma; z, u_\pm) \,-\, \lambda\,)$ is nonsingular. See Propositions 3.4.1 and 3.4.2 in [2]. ∎

The proof of Lemma 3.4 actually shows that (3.12) holds with a restricted definition of Λ, $\cup\Lambda(u)$ for u describing a curve connecting u_- to u_+ in $\mathcal{U}$. We have left the whole set $\mathcal{U}$ for simplicity.

We complete this subsection by showing that our assumptions on the scheme eliminate the possibility of having unstable essential spectrum. By unstable we mean spectrum of positive real part regarding L and spectrum of modulus greater than 1 regarding $\mathcal{R}$. In view of Lemma 3.4, this amounts to showing that the set Λ lies in the right half-plane. In fact, we have a slightly more precise result.

THEOREM 3.6. *For* 3*-point schemes (i.e.,* $p = q = 1$*) satisfying* (H1) *and* (H3), *the operators* $\mathcal{R}$*, defined in* (3.3)*, and* L*, defined in* (3.10)*, are such that*

$$\Sigma_{ess}(\mathcal{R}) \,\subset\, \{\,\zeta \in \mathbb{C};\; |\zeta| < 1\} \cup \{1\},$$

$$\Sigma_{ess}(L) \,\subset\, \{\,\lambda \in \mathbb{C};\; \mathrm{Re}\lambda < 0\} \cup 2\,i\,\pi\,\sigma\,\mathbb{Z}.$$

Proof. In view of (3.12)–(3.13), this amounts to showing that for any $u \in \mathcal{U}$, for $\mathrm{Re}\lambda \ge 0$, and for $\lambda \notin 2\,i\,\pi\,\sigma\,\mathbb{Z}$,

$$\det(\,\Delta(\sigma; i\xi, u) \,-\, \lambda\,) \,\neq\, 0 \quad \forall\, \xi \in \mathbb{R}.$$

For 3-point schemes, this is easy. As a matter of fact, the coefficients of the diagonalized $\Delta(\sigma; i\xi, u)$ are of nonpositive real part, as already noticed in section 2.2. Hence $\Delta(\sigma; i\xi, u)$ does not have any eigenvalue of positive real part. Furthermore, $\Delta(\sigma; i\xi, u)$ has a purely imaginary eigenvalue only if ξ belongs to $2\pi\mathbb{Z}$, and then that eigenvalue is $\lambda = \sigma\xi$. For schemes with larger stencil, (H1) and (H3) imply that the real parts of the coefficients of the diagonalized $\Delta(\sigma; i\xi, u)$ have a local maximum at $\xi = 0$. With the additional assumption that those coefficients achieve their global maximum on $[-\pi, \pi]$ at $\xi = 0$, then we would get the same conclusion as for 3-point schemes. ∎

3.5. Point spectrum. With regard to the point spectrum of $\mathcal{R}$, an interesting preliminary result is the following.

PROPOSITION 3.7. *If* $\zeta \notin \exp(\Lambda/\sigma)$*, where* Λ *is defined as in* (3.12)*, and* ζ *is an eigenvalue of* $\mathcal{R}$ *in* $\mathcal{L}^\infty$*, then the corresponding eigenvector* w *is actually exponentially decaying and thus belongs to all the spaces* $\mathcal{L}^\alpha(\mathbb{Z})$ *for* $\alpha \in \mathbb{N}^*$.

Proof. The proof relies on a pointwise estimate similar to (3.21) but for $\mathcal{R}$ instead of $\mathcal{R}_0$. As a matter of fact, we can show the existence of a constant C and $\alpha > 0$ such that for all $w \in \mathcal{L}^\infty$

$$|\, w_j \,| \,\le\, C \;(\, \|(\mathcal{R} - \zeta) \cdot w\|_\infty \,+\, e^{-\,\alpha\,|j|}\, \|w\|_\infty \,) \quad \forall j \in \mathbb{Z}. \tag{3.24}$$

The proof of (3.24) is very similar to that of Proposition 3.5. Rewrite (3.2) as

$$\frac{dv}{dt} \,=\, D\mathcal{G}(u_\pm)\cdot v \,+\, \mathcal{E}^\pm v, \quad \mathcal{E}^\pm \,=\, D\mathcal{G}(u) \,-\, D\mathcal{G}(u_\pm),$$

and deduce that $v \,=\, (\mathcal{R} \,-\zeta)\cdot w$ is equivalent to

$$w \,=\, \varphi^\pm \,*\, \left(v \,-\, \mathcal{T}^{-1} \int_0^T \psi^\pm(T-t) * (\mathcal{E}^\pm \mathcal{S}(t)\cdot w)\, dt \right),$$

where the kernels $\varphi^\pm$ and $\psi^\pm$ decay exponentially fast at $\pm\infty$. More precisely, $\varphi^\pm$ decay at most like $e^{-\beta|j|}$, where $\beta > 0$ is determined by a spectral gap argument, and $\psi^\pm$ is an $\mathcal{O}(e^{-\delta|j|})$ for δ arbitrarily large. Using a uniform bound for $\mathcal{S}(t)$ for $t \in [0,T]$ and the exponential decay (2.22) of the profile, we obtain by a convolution estimate that

$$\int_0^T \psi^\pm(T-t) * (\mathcal{E}^\pm \mathcal{S}(t)\cdot w)\, dt \,\le\, C\, \mathrm{e}^{\theta_\pm\, j}\, \|w\|_\infty, \quad \pm j > 0.$$

Convolution by $\varphi^\pm$ eventually yields (3.24) with $\alpha = \min(\beta, \theta^-, -\theta^+)$. ∎

We can now prove a kind of spectral mapping theorem regarding the point spectrum of the operators $\mathcal{R}$ and L. It is in fact partly contained in [6]. The result about multiplicities seems to be new though.

THEOREM 3.8. *We consider the point spectrum, $\Sigma_p^\alpha(L)$, of the operator L defined by* (3.10) *on $W^{1,\alpha}$, $1 \le \alpha \le \infty$, and the point spectrum of the operator $\mathcal{R}$ defined in* (3.3) *and acting on $\mathcal{L}^\alpha$, that is,*

$$\Sigma_p^\alpha(\mathcal{R}) \,:=\, \{\zeta \in \mathbb{C}\,;\; \exists w \neq 0 \in \mathcal{L}^\alpha\,,\; \mathcal{R}\cdot w \,=\, \zeta\, w\,\}.$$

Then

$$\Sigma_p^\alpha(\mathcal{R})\backslash\{0\} \,=\, \mathrm{e}^{\sigma^{-1}\,\Sigma_p^\alpha(L)}. \tag{3.25}$$

Furthermore, if ζ is a nonzero eigenvalue of $\mathcal{R}$ and $\zeta = \mathrm{e}^{\lambda/\sigma}$, then the multiplicity of λ as an eigenvalue of L equals the multiplicity of ζ as an eigenvalue of $\mathcal{R}$.

Proof. Assume that $\mathcal{R}\cdot w \,=\, \zeta\, w$ and consider any λ such that $\zeta = \mathrm{e}^{\lambda/\sigma}$. We may preferably write

$$\mathrm{e}^{\lambda/\sigma} \,=\, \mathrm{e}^{\underline{\lambda}\, T}, \quad \text{with} \quad \underline{\lambda} := \frac{\lambda}{\Delta x}.$$

We associate with λ and w the evolving sequence

$$y(t) \,:=\, \mathrm{e}^{-\underline{\lambda} t}\, \mathcal{S}(t,0)\cdot w.$$

It obviously solves the LDS

$$\frac{dy}{dt} \,=\, (\, D\mathcal{G}(u) \,-\, \underline{\lambda}\,)\cdot y. \tag{3.26}$$

Furthermore, y is a traveling wave of speed s. As a matter of fact, it is easy to check that (3.1) holds for $v = y$. We have

$$y(t+T) \,=\, \mathrm{e}^{-\underline{\lambda}(t+T)}\, \mathcal{S}(t+T,0)\cdot w \,=\, \frac{\mathrm{e}^{-\underline{\lambda} t}}{\zeta}\, \mathcal{S}(t+T,T)\mathcal{S}(T,0)\cdot w.$$

By assumption on ζ and w and because of (3.4) we have

$$\mathcal{S}(t+T,T)\mathcal{S}(T,0)\cdot w \;=\; \zeta\,\mathcal{S}(t+T,T)\mathcal{T}\cdot w \;=\; \zeta\,\mathcal{T}\mathcal{S}(t,0)\cdot w,$$

and thus we conclude that $y(t+T) \;=\; \mathcal{T}\cdot y(t)$. Therefore, we have $y_j(t) \;=\; Y(j-st)$. Substituting in (3.26) and multiplying by Δx, we get

$$-\sigma\,Y'(x) \;=\; \sum_{l=-p}^{q} \partial_l G(\,U(x-p),\dots,U(x+q)\,)\cdot Y(x+l) \;-\; \lambda Y(x)$$

for $x = j - st$, that is, $(L-\lambda)\cdot Y \;=\; 0$. (Recall that $s\Delta x = \sigma$.)

Conversely, if $(L-\lambda)\cdot Y \;=\; 0$, then the sequence w defined by $w_j := Y(j)$ is an eigenvector of $\mathcal{R}$ for the eigenvalue $\mathrm{e}^{\lambda/\sigma}$. This follows from the fact that $y_j(t) := Y(j-st)$ solves (3.26); hence

$$\mathrm{e}^{\lambda t/\Delta x}\,y(t) \;=\; \mathcal{S}(t,0)\,y(0).$$

At $t = T$ this precisely means, since y is by construction a traveling wave, that

$$\mathrm{e}^{\lambda/\sigma}y(0) \;=\; \mathcal{R}\cdot y(0).$$

Observe that if $\mathrm{e}^{\widetilde{\lambda}/\sigma} = \mathrm{e}^{\lambda/\sigma}$, that is, if $\widetilde{\lambda} = \lambda + 2i\pi m\sigma$ with m an integer, then $(L-\widetilde{\lambda})$ and $(L-\lambda)$ are conjugated under the multiplication by $\mathrm{e}^{2i\pi mx}$. Therefore, the multiplicity of λ as an eigenvalue of L does not depend on the chosen representation of ζ. Furthermore, if Y and $\widetilde{Y}$ are, respectively, associated with λ and $\widetilde{\lambda}$, they yield the same eigenvector $w_j \;=\; Y(j) \;=\; \widetilde{Y}(j)$ of $\mathcal{R}$ since $\mathrm{e}^{2i\pi mj} \;=\; 1$. This argument shows that the geometric multiplicities of ζ and λ coincide.

With regard to the algebraic multiplicities, a similar but more technical argument works to show that they do coincide. We leave the general, cumbersome proof to the reader. For clarity, we present the proof in the case of double nonsemisimple eigenvalues. Assume that λ is a double, nonsemisimple eigenvalue of L. Then there exist Y and Z such that

$$L\cdot Y \;=\; \lambda Y, \qquad L\cdot Z \;=\; \lambda Z \;+\; Y.$$

Let y and z be defined by

$$y_j(t) \;:=\; Y(j-st), \qquad z_j(t) \;:=\; Z(j-st).$$

Then y solves (3.26), while z solves the LDS

$$\frac{dz}{dt} \;=\; (\,D\mathcal{G}(u) \;-\; \underline{\lambda}\,)\cdot z \;-\; \underline{y}, \tag{3.27}$$

where $\underline{\lambda} \;=\; \lambda/\Delta x$, as denoted above, and similarly $\underline{y} \;=\; y/\Delta x$. Hence by Duhamel's formula we get

$$\mathrm{e}^{\underline{\lambda}t}\,z(t) \;=\; \mathcal{S}(t,0)\cdot z(0) + \int_0^t \mathcal{S}(t,\tau)\cdot(-\mathrm{e}^{\underline{\lambda}\tau}\,\underline{y}(\tau))\,d\tau \;=\; \mathcal{S}(t,0)\cdot z(0) - \frac{t}{\Delta x}\,\mathcal{S}(t,0)\cdot y(0).$$

At $t = T$ this gives

$$\mathrm{e}^{\lambda T}z(0) \;=\; \mathcal{R}\cdot z(0) \;-\; \frac{1}{\sigma}\,\mathrm{e}^{\lambda/\sigma}\,y(0).$$

Therefore, $z(0)$ belongs to $\mathrm{Ker}(\mathcal{R}-\zeta)^2\backslash\mathrm{Ker}(\mathcal{R}-\zeta)$, which means that the algebraic multiplicity of ζ is at least 2. Conversely, if $\zeta \neq 0$ is a double, nonsemisimple eigenvalue of $\mathcal{R}$, there exist w and v such that

$$\mathcal{R}\cdot w \,=\, \zeta w, \qquad \mathcal{R}\cdot v \,=\, \zeta(\, v \,+\, w/\sigma\,).$$

Then both

$$y(t) \,:=\, \mathrm{e}^{-\frac{\lambda}{\Delta}t}\,\mathcal{S}(t,0)\cdot w$$

and

$$z(t) \,:=\, \mathrm{e}^{-\frac{\lambda}{\Delta}t}\left(\mathcal{S}(t,0)\cdot v \,-\, \frac{t}{\Delta x}\,\mathcal{S}(t,0)\cdot w\right)$$

are traveling waves of speed σ—this is proved using (3.4) again. Introducing Y and Z so that

$$y_j(t) \,=\, Y(j-st), \qquad z_j(t) \,=\, Z(j-st)$$

and reversing the computation done before, we see that

$$L\cdot Y \,=\, \lambda Y, \qquad L\cdot Z \,=\, \lambda Z \,+\, Y.$$

Thus the algebraic multiplicity of λ is at least 2. ∎

In view of Theorems 3.6 and 3.8, checking the stability condition in (3.6) amounts to showing that

$$\Sigma_{\mathrm{p}}(L) \,\subset\, \{\,\lambda\in\mathbb{C}\,;\ \mathrm{Re}\lambda<0\}\cup 2\,i\,\pi\,\sigma\,\mathbb{Z}.$$

In some special cases (scalar conservation laws or shocks of small amplitude), (3.6) can be checked through energy estimates. This is performed in the next section. For arbitrary shocks, $\Sigma_{\mathrm{p}}(L)$ is encoded in an Evans function, the construction of which was derived in [2] for the upwind scheme. There is some indication in very recent results due independently to Mallet-Paret and Verduyn Lunel [17] and Härterich, Sandstede, and Scheel [10] that an Evans function might also be constructed for general schemes. This is discussed in section 3.7. With regard to condition (ii) in Theorem 3.2, we also show in section 3.7 that it must fail. We propose a weaker condition, in terms of the Evans function, which ensures stability though.

3.6. Energy estimates. The aim of this section is to deduce from energy estimates the spectral condition (i) of Theorem 3.2 for the point spectrum of $\mathcal{R}$ in the case of small shocks. Our method is very much inspired from that of Goodman [9] (also see [13]). In particular, it relies on a special diagonalization process and on the trick that consists of "integrating" the variational system (2.2) *before* deriving estimates. However, some additional difficulties are due to the mixing of true derivatives with discrete derivatives. This is why we need rather fine properties of profiles, as they are stated in Proposition 2.4.

To simplify the writing, we perform here a rescaling and set $\Delta x = 1$. Also, for the sake of clarity, we present the energy estimates procedure in the case of the *upwind scheme*, for which $p = 1$, $q = 0$. But the very same method works for the more general scheme (2.1).

In the case $p = 1$, $q = 0$, the semidiscrete system (2.1), or (2.2), then reduces to

$$\frac{dv_j}{dt} + f(v_j) - f(v_{j-1}) = 0$$

and the linearized system (3.2) reduces to

$$\frac{dv_j}{dt} + A_j\, v_j - A_{j-1}\, v_{j-1} = 0, \tag{3.28}$$

where $A_j := A(u_j)$. The viscosity matrix here reduces to $Q = A$, and the dissipativeness requirement (H1) is that A has positive eigenvalues (which is, by the way, the appropriate condition for the linear stability of constant states). As for viscous shocks, a direct estimate on (3.28) is useless. To be convincing, let us multiply (3.28) by v_j and sum on j. We get

$$\frac{d}{dt}\left(\sum_{j=-\infty}^{+\infty} |v_j|^2\right) + 2\sum_{j=-\infty}^{+\infty} v_j\,(A_j\, v_j - A_{j-1}\, v_{j-1}) = 0.$$

Assuming for awhile that the matrices A_j are symmetric (for instance diagonal), we have the identity

$$2\langle v, Av - \mathcal{T}(Av)\rangle = \langle v - \mathcal{T}v, (\mathcal{T}A)(v - \mathcal{T}v)\rangle + \langle v, (A - \mathcal{T}A)v\rangle. \tag{3.29}$$

Here and below in this section, the brackets $\langle\ ,\ \rangle$ stand for the inner product in $\mathcal{L}^2(\mathbb{Z})$, and the associated norm will be denoted by $\|\ \|$. Substituting (3.29) in the previous equality, we obtain

$$\frac{d}{dt}\|v\|^2 + \langle v - \mathcal{T}v, (\mathcal{T}A)(v - \mathcal{T}v)\rangle + \langle v, (A - \mathcal{T}A)v\rangle = 0.$$

By (H1) and thus $A = Q > 0$, the term $\langle v - \mathcal{T}v, (\mathcal{T}A)(v - \mathcal{T}v)\rangle$ is nonnegative. But the term $\langle v, (A - \mathcal{T}A)v\rangle$ is typically not. For instance, in the scalar case, the characteristic speed a is decreasing along the profile and thus $(A - \mathcal{T}A)$ is nonpositive.

Similarly as for viscous shocks [9], this problem can be fixed by considering an "integrated" version of (3.28). Assume that $|\zeta| \geq 1$ and $\zeta \neq 1$ is an eigenvalue of $\mathcal{R}$, and w is a corresponding eigenvector. By Theorem 3.6, $\zeta \notin \exp(\Lambda/\sigma)$. Then Proposition 3.7 shows that w is summable and thus $v(t) := \mathcal{S}(t, 0)\cdot w$ is summable for all t. Furthermore, we see by summing (3.28) that $\sum_j v_j(t) = \sum_j w_j$. This is due to the conservativity of the scheme. In particular, this shows that $\sum \mathcal{R}\cdot w = \sum w$. But we also have by assumption on w that $\sum \mathcal{R}\cdot w = \zeta \sum w$. Therefore, since $\zeta \neq 1$, we must have

$$\sum_{j=-\infty}^{+\infty} w_j = 0.$$

This allows us to define

$$W_j := \sum_{i=-\infty}^{j} w_i = -\sum_{i=j+1}^{+\infty} w_i.$$

Furthermore, this new sequence W is exponentially decaying and thus belongs to all the spaces $\mathcal{L}^\alpha(\mathbb{Z})$ by Proposition 3.7. Now it is easy to see that $V_j(t) := \sum_{i=-\infty}^{j} v_i(t)$ satisfies the "integrated" LDS

$$\frac{dV_j}{dt} + A_j\,(V_j - V_{j-1}) = 0. \tag{3.30}$$

Performing a similar computation as on the original LDS and using the identity

$$2\,\langle v\,,\,A(v-\mathcal{T}v)\rangle = \langle v-\mathcal{T}v\,,\,A(v-\mathcal{T}v)\rangle + \langle v\,,\,(A-\mathcal{T}^{-1}A)\,v\rangle, \tag{3.31}$$

which is valid for symmetric matrices A_j, we obtain

$$\frac{d}{dt}\|V\|^2 + \langle V-\mathcal{T}V\,,\,A(V-\mathcal{T}V)\rangle + \langle V\,,\,(A-\mathcal{T}^{-1}A)\,V\rangle = 0.$$

We observe that the sign of $\langle V\,,\,(A-\mathcal{T}^{-1}A)\,V\rangle$ is now the good one, at least in the scalar case. For systems we must cope with the fact that, in general, the matrices A_j are not symmetric and that only $\langle V\,,\,(a^k-\mathcal{T}^{-1}a^k)\,V\rangle$ is nonnegative. This is done below by using two essential tools originally due to Goodman [9], a special diagonalization process and weights on the outgoing waves. We first state the theorem.

THEOREM 3.9. *Assuming* (H1), (H2), *and* (H3), *for sufficiently small shock profiles, the* $\mathcal{L}^1$ *solutions of* (3.2) *with zero mass are such that* $V_j := \sum_{i=-\infty}^{j} v_j$ *satisfy an estimate*

$$\tilde{\|}\,\mathcal{T}^{-1}V(T)\,\tilde{\|}^2_{\mathcal{L}^2} + \omega\int_0^T \|\,V(t) - \mathcal{T}V(t)\,\|^2_{\mathcal{L}^2}\,dt \le \tilde{\|}\,V(0)\,\tilde{\|}^2_{\mathcal{L}^2} \tag{3.32}$$

for some positive ω, *where* $\tilde{\|}\cdot\tilde{\|}$ *is an equivalent norm on* $\mathcal{L}^2$. *As a consequence,* $\mathcal{R}$ *does not have any eigenvalue* $\zeta \neq 1$ *with* $|\zeta| \ge 1$.

Proof. The last statement follows directly from (3.32) and the above argument. As a matter of fact, if $\mathcal{R}\cdot w = \zeta w$ with $|\zeta| \ge 1$ and $\zeta \neq 1$, then $v(t) = \mathcal{S}(t,0)\cdot w$ is such that $v(T) = \zeta\mathcal{T}\,v(0)$. Thus, taking the sum, $V(T) = \zeta\mathcal{T}\,V(0)$, and substituting in (3.32) we get a contradiction unless $|\zeta| = 1$ and $V \equiv \mathcal{T}V$. The latter equality implies by definition of V that $v = 0$ and thus also $w = 0$. So the whole proof is devoted to the energy estimate (3.32), which will appear to be a direct consequence of Lemma 3.10. More precisely, with the notations introduced below, (3.32) follows from (3.35) with

$$\tilde{\|}V\tilde{\|}^2_{\mathcal{L}^2} := \sum_{m\neq k} \|\phi^m(0)\,\ell^m(u(0))\,V\|^2 + \|\ell^k(u(0))\,V\|^2$$

and using the characterization of traveling waves in (3.1) for $\phi^m\,\ell^m$ and ℓ^k. ∎

Let us first describe the diagonalization process. To simplify the notations we write

$$\ell^m_j = \ell^m(u_j) \quad\text{and}\quad r^m_j = r^m(u_j).$$

The important property to have in mind is that the ℓ^m and r^m defined this way are traveling with speed s. Since the profile is exponentially decaying, it is now possible to renormalize ℓ^m and r^m in such a way that

$$\left(\frac{d}{dt}\ell^m_j\right)\cdot r^m_j \equiv \ell^m_j\cdot\frac{d}{dt}r^m_j \equiv 0. \tag{3.33}$$

This is achieved by considering

$$\widetilde{\ell}_j^m \,=\, \frac{1}{\nu_j^m}\,\ell_j^m \quad \text{and} \quad \widetilde{r}_j^m \,=\, \nu_j^m\, r_j^m,$$

with

$$\nu_j^m(t) \,:=\, \nu^m(j-st) \,=\, \exp\,\left(-\int_0^{j-st} \ell^m(U(x))\cdot\frac{d}{dx}\,r^m(U(x))\,dx\right).$$

It is important to note that the weight ν^m is traveling with the speed s, and thus also the new vectors $\widetilde{\ell}^m$ and $\widetilde{r}^m$. Furthermore, we easily check that $\widetilde{\ell}^m$ and $\widetilde{r}^m$ satisfy (3.33).

From now on, we drop the tildes and assume that (3.33) holds for all $m \in \{1,\dots,N\}$.

Further notations. We denote

$$L_j := \begin{pmatrix} -\ell_j^1- \\ \vdots \\ -\ell_j^N- \end{pmatrix} \quad \text{and} \quad R_j := \begin{pmatrix} | & & | \\ r_j^1 & \cdots & r_j^N \\ | & & | \end{pmatrix},$$

the square matrices made of eigenvectors satisfying (3.33), and

$$a_j \,=\, L_j\, A_j\, R_j \,=\, \text{diag}\,(a_j^1,\dots,a_j^N).$$

This is consistent with the notations introduced in section 2.1. And from now on in this section,

$$v_j \,:=\, L_j\cdot V_j,$$

which has nothing to do with the $v = V - \mathcal{T}V$—satisfying (3.28)—originally considered. We hope that this choice of notation will not be confusing to the reader. Observe that, conversely,

$$V_j \,=\, R_j\cdot v_j.$$

Multiplying (3.30) on the left by L_j, we get the LDS satisfied by v,

$$\frac{dv_j}{dt} \,+\, a_j\,(\,v_j \,-\, v_{j-1}\,) \,=\, \frac{dL_j}{dt}\,R_j\,v_j \,+\, a_j\,L_j\,(\,R_{j-1} \,-\, R_j\,)\,v_{j-1}. \tag{3.34}$$

Lemma 3.10. *Assuming that* (H1)–(H3) *hold,*[4] *there exist* $\varepsilon_0 > 0$, *weights* $\phi^m(t)$, *uniformly bounded from below in* $\mathcal{L}^\infty$ *and traveling with speed* s, *and* $\omega > 0$ *so that if the* k*-shock has an amplitude* $\varepsilon = |u_+ - u_-|$ *less than* ε_0 *and if* v *solves* (3.34), *then*

$$\begin{aligned}\sum_{m\neq k} \|\phi^m(T)v^m(T)\|^2 \,+\, \|v^k(T)\|^2 \,+\, \omega\int_0^T \|\,v \,-\, \mathcal{T}v\,\|^2 \\ \leq \sum_{m\neq k} \|\phi^m(0)v^m(0)\|^2 \,+\, \|v^k(0)\|^2.\end{aligned} \tag{3.35}$$

[4]We recall that for the upwind scheme, (H2) is implied by (H1), and (H3) is trivial.

Proof. The proof proceeds in two steps. We begin by showing that

(3.36)
$$\frac{d}{dt}\|v\|^2 + \frac{\mu}{2}\|v - \mathcal{T}v\|^2 + \frac{1}{2}\langle v^k, (a^k - \mathcal{T}^{-1}a^k)v^k\rangle \leq cst\,\langle \check{v}, (a^k - \mathcal{T}^{-1}a^k)\check{v}\rangle,$$

where $\check{v}$ is defined by $\check{v}^m = v^m$, $m \neq k$, and $\check{v}^k = 0$. Afterwards, we shall construct weights ϕ^m yielding estimates of the "outgoing waves" that can absorb the right-hand side of (3.36), thus showing that

$$\sum_{m\neq k} \|\phi^m(t)v^m(t)\|^2 + \|v(t)\|^2$$

is strictly decreasing with t.

Proof of (3.36). Taking the inner product of (3.34) with v we get

$$\frac{d}{dt}\|v\|^2 + 2\langle v, a(v - \mathcal{T}v)\rangle = 2\left\langle v, \frac{dL}{dt}Rv\right\rangle + 2\langle v, aL(\mathcal{T}R - R)\mathcal{T}v\rangle.$$

Applying the identity (3.31) to the diagonal matrix a, this also reads

$$\begin{aligned}\frac{d}{dt}\|v\|^2 + \langle v - \mathcal{T}v, a(v - \mathcal{T}v)\rangle + \langle v, (a - \mathcal{T}^{-1}a)v\rangle \\ = 2\left\langle v, \frac{dL}{dt}Rv\right\rangle + 2\langle v, aL(\mathcal{T}R - R)\mathcal{T}v\rangle.\end{aligned} \tag{3.37}$$

By (2.21) there is a constant such that

$$|a - \mathcal{T}^{-1}a| \leq cst\,(a^k - \mathcal{T}^{-1}a^k) = cst\,|a^k - \mathcal{T}^{-1}a^k|.$$

Hence we see that, if the right-hand side in (3.37) equals 0 (which occurs if the system (3.30) is already diagonal), (3.37) trivially implies (3.36). In order to deal with the general right-hand side, we use in a crucial way the special normalization of the eigenvectors (3.33). As a matter of fact, (3.33) implies that the diagonal coefficients of the matrices $\frac{dL_j}{dt}R_j$ are zero. (In fact, we need only that the coefficient $\frac{d\ell_j^k}{dt}r_j^k$ be zero.) And by (2.21) all the other coefficients are bounded by a constant times $(a_j^k - a_{j+1}^k)$. Therefore, for all $\gamma > 0$ there exists $c_\gamma > 0$ such that

$$2\left\langle v, \frac{dL}{dt}Rv\right\rangle \leq \gamma\langle v^k, (a^k - \mathcal{T}^{-1}a^k)v^k\rangle + c_\gamma\langle \check{v}, (a^k - \mathcal{T}^{-1}a^k)\check{v}\rangle.$$

The last term in (3.37) is the most complicated to deal with. We first split it as follows:

$$\langle v, aL(\mathcal{T}R - R)\mathcal{T}v\rangle = \langle v, aL(\mathcal{T}R - R)v\rangle + \langle v, aL(\mathcal{T}R - R)(\mathcal{T}v - v)\rangle.$$

The diagonal of the matrices $a_j L_j (R_{j-1} - R_j)$ occurring here is not necessarily zero. However, the diagonal coefficients are of order 2 in terms of the shock strength ε, and the other coefficients are of order 1, as we show just after. Consequently, for all $\gamma > 0$ there is a constant still denoted c_γ (up to augmenting the previous one) so that, on the one hand,

$$2\langle v, aL(\mathcal{T}R - R)v\rangle \leq (\varepsilon + \gamma)\langle v^k, (a^k - \mathcal{T}^{-1}a^k)v^k\rangle + c_\gamma\langle \check{v}, (a^k - \mathcal{T}^{-1}a^k)\check{v}\rangle,$$

and, on the other hand,

$$2\,\langle\, v\,,\, a\,L\,(\,\mathcal{T}R\,-\,R\,)\,(\,\mathcal{T}v\,-\,v\,)\,\rangle\,\le\,\gamma\,\|\,\mathcal{T}v\,-\,v\,\|^2+\,c_\gamma\,\varepsilon\,\langle\, v\,,\,(\,a^k\,-\,\mathcal{T}^{-1}a^k\,)\,v\,\rangle.$$

Collecting these estimates altogether and substituting the resulting inequality in (3.37), we obtain (3.36) by choosing $\gamma \le \mu/2$ (the lower bound of a which is positive by (H1)) and $\varepsilon \le \varepsilon_0 = (1-2\gamma)/(2(1+c_\gamma))$. □

Estimates on the matrices $L_j\,(\,R_{j-1}\,-\,R_j\,)$. We first note that

$$r^m_{j-1}(t)\,-\,r^m_j(t)\,=\,r^m_j(t+T)\,-\,r^m_j(t)\,=\,\int_t^{t+T}\,\frac{d}{d\tau}r^m(U(j-s\tau))\,d\tau$$

is bounded by a constant times $(\,a^k_j\,-\,a^k_{j+1}\,)$, in view of (2.21). Then, leading further the procedure, we see that

$$\ell^m_j(t)\,(\,r^m_{j-1}(t)\,-\,r^m_j(t)\,)\,=\,\int_t^{t+T}\,(\,\ell^m_j(t)\,-\,\ell^m_j(\tau)\,)\,\frac{d}{d\tau}r^m(U(j-s\tau))\,d\tau$$

because of (3.33), and thus

$$\ell^m_j(t)\,(\,r^m_{j-1}(t)\,-\,r^m_j(t)\,)\,=\,\int_t^{t+T}\,\int_\tau^t\,\frac{d}{d\theta}\ell^m(U(j-s\theta))\,d\theta\,\frac{d}{d\tau}r^m(U(j-s\tau))\,d\tau$$

is bounded by a constant times $(\,a^k_j\,-\,a^k_{j+1}\,)^2$, using twice (2.21). □

Proof of the outgoing waves estimates. The mth component of (3.34) reads

$$\frac{dv^m_j}{dt}\,+\,a^m_j\,(\,v^m_j\,-\,v^m_{j-1}\,)\,=\,\frac{d\ell^m_j}{dt}\,R_j\,v_j\,+\,a^m_j\,\ell^m_j\,(\,R_{j-1}\,-\,R_j\,)\,v_{j-1}.$$

We consider a weight of the form

$$\phi^m_j(t)\,:=\,\Phi^m(j-st)\,=\,\exp\,\left(\,\frac{M}{\underline{a}-s}\,\int_0^{j-st}\,\frac{d}{dx}\,a^k(U(x))\,dx\,\right),$$

where the constant M will be chosen large enough later on, and $\underline{a}$ is chosen in a way that either

$$a^m \le \underline{a} < s \quad \text{if} \quad m \le k-1$$

or

$$a^m \ge \underline{a} > s \quad \text{if} \quad m \ge k+1.$$

Observe that in all cases

$$\frac{a^m-s}{\underline{a}-s}\,\ge\,1.$$

Also observe that $\phi^m(t)$ is uniformly bounded from below and above in $\mathcal{L}^\infty$, *uniformly* in M, ε when $M\varepsilon < 1$. Then multiplying the previous equality by $\phi^m_j\,v^m_j$ and summing, we get

$$\frac{d}{dt}\langle\,v^m\,,\,\phi^m\,v^m\,\rangle\,+\,2\,\langle\,v^m\,,\,\phi^m\,a^m\,(\,v^m\,-\,\mathcal{T}v^m\,)\,\rangle$$

$$=\,2\,\left\langle\,v^m\,,\,\frac{d\phi^m}{dt}\,v^m\,\right\rangle\,+\,2\,\left\langle\,v\,,\,\phi^m\,\frac{d\ell^m}{dt}\,R\,v\,\right\rangle\,+\,2\,\langle\,v^m\,,\,\phi^m a^m\,\ell^m\,(\,\mathcal{T}R\,-\,R\,)\,\mathcal{T}v\,\rangle.$$

Applying the identity (3.31) to the scalar valued multiplier $\phi^m a^m$, this also reads

$$\begin{aligned}\frac{d}{dt}\langle\, v^m\,,\,\phi^m\, v^m\,\rangle\ &+\langle\, v^m\,-\mathcal{T}v^m\,,\,\phi^m a^m\,(\,v^m\,-\,\mathcal{T}v^m)\,\rangle\\ &+\left\langle\, v^m\,,\left(\,\phi^m a^m\,-\,\mathcal{T}^{-1}(\phi^m a^m)\,-\,\frac{d\phi^m}{dt}\,\right)v^m\,\right\rangle\\ &=2\left\langle\, v^m\,,\,\phi^m\frac{d\ell^m}{dt}\,R\,v\,\right\rangle+\,2\,\langle\, v^m\,,\,\phi^m a^m\,\ell^m\,(\,\mathcal{T}R\,-\,R\,)\,\mathcal{T}v\,\rangle.\end{aligned}\tag{3.38}$$

Due to the boundedness of ϕ^m, both terms in the right-hand side of (3.38) can be bounded exactly in the same way as their counterparts in (3.37). We are going to see that choosing M large enough and ε small enough such that $M\varepsilon<1$ makes the last term of the left-hand side larger than and thus absorbing the right-hand side. As a matter of fact, we have

$$\phi_j^m a_j^m\,-\,\phi_{j+1}^m a_{j+1}^m\,-\,\frac{d\phi_j^m}{dt}\,=\,(\,\phi_j^m\,-\,\phi_{j+1}^m\,)\,a_j^m\,-\,\frac{d\phi_j^m}{dt}\,+\,\phi_{j+1}^m\,(\,a_j^m\,-\,a_{j+1}^m\,).$$

The last term is bounded by $c\,(\,a_j^k-a_{j+1}^k\,)$, c a positive constant (independent of M). The remaining terms can be rearranged as

$$(\,\phi_j^m\,-\,\phi_{j+1}^m\,)\,a_j^m\,-\,\frac{d\phi_j^m}{dt}\,=\,\int_{t-T}^{t}\left((a_j^m(t)-s)\,\frac{d\phi_j^m}{dt}(\tau)\,+\,s\int_t^\tau\,\frac{d^2\phi_j^m}{dt^2}(\tau')\right)d\tau'\,d\tau.$$

By definition of ϕ^m and (2.21) we thus see that there are constants $\alpha>0$ and $c'>0$ so that for $M\varepsilon<1$

$$(\,\phi_j^m\,-\,\phi_{j+1}^m\,)\,a_j^m\,-\,\frac{d\phi_j^m}{dt}\,\geq\,\alpha\,M\,(1-c'\,\varepsilon)\,(\,a_j^k\,-\,a_{j+1}^k\,).$$

Choosing, for instance, $M\geq 4c/\alpha$ and $\varepsilon\,<\,1/4c'$, we get

$$\phi_j^m a_j^m\,-\,\phi_{j+1}^m a_{j+1}^m\,-\,\frac{d\phi_j^m}{dt}\,\geq\,\alpha\,\frac{M}{2}\,(\,a_j^k\,-\,a_{j+1}^k\,).$$

Substituting this in (3.38) and using the usual bound for the right-hand side, we have for all $\gamma>0$ a constant $c_\gamma>0$ so that

$$\begin{aligned}&\frac{d}{dt}\|\phi^m v^m\|^2\,+\,\alpha\,\frac{M}{2}\,\langle\, v^m\,,\,(\,a^k\,-\,\mathcal{T}^{-1}a^k\,)\,v^m\,\rangle\\ &\leq\,\gamma\,\|\,v\,-\mathcal{T}v\|^2\,+\,(\,\varepsilon(1+c_\gamma)\,+\,\gamma\,)\langle\, v^k\,,\,(\,a^k\,-\,\mathcal{T}^{-1}a^k\,)\,v^k\,\rangle\\ &+c_\gamma\,\langle\,\check{v}\,,\,(\,a^k\,-\,\mathcal{T}^{-1}a^k)\,\check{v}\,\rangle\end{aligned}\tag{3.39}$$

for $m\neq k$. □

Proof of (3.35). Summing (3.39) on m and adding the result to (3.36), we obtain the inequality

$$\sum_{m\neq k} \frac{d}{dt}\|\phi^m v^m\|^2 + \frac{d}{dt}\|v^k\|^2 + \frac{\mu}{2}\|v - \mathcal{T}v\|^2$$

$$+\frac{1}{2}\langle v^k, (a^k - \mathcal{T}^{-1}a^k)v^k,\rangle + \alpha\frac{M}{2}\sum_{m\neq k}\langle v^m, (a^k - \mathcal{T}^{-1}a^k)v^m\rangle$$

$$\leq \gamma\|v - \mathcal{T}v\|^2 + N(\varepsilon(1+c_\gamma)+\gamma)\langle v^k, (a^k - \mathcal{T}^{-1}a^k)v^k\rangle$$

$$+c'_\gamma\langle \check{v}, (a^k - \mathcal{T}^{-1}a^k)\check{v}\rangle,$$

where we have changed ϕ^m into $1+\phi^m$, which is harmless. Then, taking $\gamma \leq \min(\mu/4, 1/8N)$, $M\alpha \geq \max(4c'_\gamma, 1)$, and $\varepsilon < \min(1/(8N(1+c_\gamma)), 1/M)$, we find that

$$\frac{d}{dt}\left(\sum_{m\neq k}\|\phi^m v^m\|^2 + \|v^k\|^2\right) + \frac{\mu}{4}\|v - \mathcal{T}v\|^2$$

$$+\frac{1}{4}\langle v, (a^k - \mathcal{T}^{-1}a^k)v\rangle \leq 0.$$

The last term of the left-hand side being nonnegative, this proves (3.35) by integration in time. ∎

3.7. The Evans function. The notion of Evans function extends that of characteristic polynomial to infinite dimensional operators L. It must be analytic in $\mathrm{Re}\lambda > 0$ and vanish only at eigenvalues of L. Its construction is based on a shooting method, using solutions of the eigenvalue equations $(L-\lambda)\cdot W = 0$ that tend to 0 either at $-\infty$ or $+\infty$. In general, there are infinite dimensional subspaces of such solutions (called the unstable/stable manifolds), which seems helpless. However, in some cases, at least one of these subspaces is finite dimensional. This is the case for the operator L associated with the upwind scheme (i.e., for $p=1$, $q=0$), of which the unstable manifold is always of finite dimension. This enabled us to construct an Evans function in a previous work [2].

We shall recall hereafter some material from [2] that will be useful in section 5. Before that, we address the question for more general schemes. The first, convenient step consists of reformulating the eigenvalue equations $(L-\lambda)\cdot W = 0$ as a linear system of the form

$$\frac{d\mathbb{W}}{dx} = \mathbb{A}(x;\lambda)\,\mathbb{W}, \tag{3.40}$$

where $\mathbb{A}(x;\lambda)$ is an unbounded operator on the Hilbert space

$$H_0 := L^2((-p,q);\mathbb{C}^N)\times\mathbb{C}^N,$$

with dense domain

$$H_1 := \left\{(\phi,c)\in H_0\,;\ \phi\in H^1((-p,q);\mathbb{C}^N) \text{ and } \phi(0)=c\right\}.$$

Denoting by $\dot\phi$ the derivative of $\phi \in H^1(-p,q)$, $\mathbb{A}(x;\lambda)$ is defined by

$$\begin{array}{rccc} (3.41)\ \mathbb{A}(x;\lambda): & H_1 & \to & H_0 \\ & \begin{pmatrix} \phi \\ c \end{pmatrix} & \mapsto & \begin{pmatrix} \dot\phi \\ \sigma^{-1}\,\lambda\,c + \sigma^{-1}\,\displaystyle\sum_{l=-p}^{q} (C^l(x) - C^{l+1}(x-1))\cdot\phi(l) \end{pmatrix}, \end{array}$$

which makes sense because of the embedding of H^1 into $\mathcal{C}^0$. Accordingly, we define the asymptotic operators

$$\begin{array}{rccc} (3.42)\,\mathbb{A}_\pm(\lambda): & H_1 & \to & H_0 \\ & \begin{pmatrix} \phi \\ c \end{pmatrix} & \mapsto & \begin{pmatrix} \dot\phi \\ \sigma^{-1}\,\lambda\,c \,+\, \sigma^{-1}\,\displaystyle\sum_{l=-p}^{q} (\,C_\pm^l \,-\, C_\pm^{l+1}\,)\cdot\phi(l) \end{pmatrix}. \end{array}$$

It is not difficult to show (see Lemma 4.3 in [3]) that $\mathbb{A}_\pm(\lambda)$ has only isolated eigenvalues of finite multiplicities and that they are precisely the roots μ of the characteristic equation

$$\det\,(\,\Delta(\sigma;\mu,u_\pm) \,-\, \lambda\,) \,=\, 0. \tag{3.43}$$

In general, there is an infinity of such roots on either side of any vertical line. This is an argument against the construction of an Evans function. Another, though related, argument is that both the forward and the backward Cauchy problems for (3.40) are ill-posed. However, Evans functions basically rely on exponential dichotomies. Recent works about exponential dichotomies for mixed-type functional differential operators [10, 17] indicate that it might be possible to overcome the difficulty.

As far as we are concerned, we concentrate on the case with delay only. With $p=1$, $q=0$, the operator $\mathbb{A}$ reduces to

$$\begin{array}{rccc} (3.44)\ \mathbb{A}(x;\lambda): & H_1 & \to & H_0 \\ & \begin{pmatrix} \phi \\ c \end{pmatrix} & \mapsto & \begin{pmatrix} \dot\phi \\ \sigma^{-1}\,(\,\lambda\,c \,+\, A(x)\,\phi(0) \,-\, A(x-1)\,\phi(-1)\,) \end{pmatrix}. \end{array}$$

Then the system (3.40) is associated with a solution operator $(T_\lambda(x,y))_{x\ge y}$, depending analytically on λ; see Lemma 4 in [2]. The link with the delay differential equation $(\,L \,-\, \lambda)\,W \,=\, 0$, which reads

$$\sigma\,W'(x) \,-\, (\,A(x) \,+\, \lambda\,)\,W(x) \,+\, A(x-1)\,W(x-1) \,=\, 0,$$

is merely the following. We denote by Γ the projection

$$\begin{array}{rccc} \Gamma: & H_0 & \to & \mathbb{C}^N \\ & \begin{pmatrix} \phi \\ c \end{pmatrix} & \mapsto & c. \end{array}$$

Then for all $\phi \in H^1$, the problem

$$(\,L \,-\, \lambda)\,W \,=\, 0, \quad W(y+\theta) \,=\, \phi(\theta) \quad \forall\,\theta \in [-1,0],$$

admits a single solution, defined for $x \geq y$ by

$$W(x) \,=\, \Gamma\, \mathbb{W}(x) \quad \text{with} \quad \mathbb{W}(x) \,=\, T_\lambda(x,y) \begin{pmatrix} \phi \\ \phi(0) \end{pmatrix}.$$

We shall also need to consider the adjoint system

$$\frac{d\mathbb{Y}}{dx} \,=\, -\,\mathbb{A}(x;\lambda)^*\,\mathbb{Y}. \tag{3.45}$$

Although the operator $\mathbb{A}(x;\lambda)^*$ is awkward (its domain depends on x), the system (3.45) is associated with a backward solution operator $(T^*_\lambda(x,y))_{x\leq y}$, with the standard relationship

$$T^*_\lambda(x,y) \,=\, T_\lambda(y,x)^*.$$

For convenience, we identify the adjoint space H_0^* with $L^2(]0,1[;(\mathbb{C}^N)^*) \times (\mathbb{C}^N)^*$ through the inner product

$$\left\langle \begin{pmatrix} \psi & b \end{pmatrix}, \begin{pmatrix} \phi \\ c \end{pmatrix} \right\rangle \,:=\, \overline{b}\cdot c \,+\, \int_{-1}^{0} \overline{\psi(\theta+1)}\,\phi(\theta)\,d\theta.$$

Note that if $\mathbb{Y}(x,\lambda)$ is a solution of (3.45) and $\mathbb{W}(x,\lambda)$ is a solution of (3.40), then

$$\langle \mathbb{Y}(x,\lambda), \mathbb{W}(x,\lambda)\rangle \text{ is independent of } x. \tag{3.46}$$

Then we can see that $T^*_\lambda(x,y)$ is linked to solutions of the advanced differential equation $(L-\lambda)^* Y = 0$ as follows. By a slight abuse of notation, Γ also stands for

$$\begin{array}{rccc} \Gamma: & H_0^* & \to & \mathbb{C}^N \\ & \begin{pmatrix} \psi & , & b \end{pmatrix} & \mapsto & b. \end{array}$$

Then the problem

$$(L-\lambda)^* Y \,=\, 0, \quad Y(y+\theta) \,=\, -\,A(y+\theta-1)^{-1}\,\psi(\theta) \quad \forall\,\theta\in[0,1],$$

admits a single solution, defined for $x \leq y$ by

$$Y(x) \,=\, \Gamma\,\mathbb{Y}(x) \quad \text{with} \quad \mathbb{Y}(x) \,=\, T^*_\lambda(x,y)\,\big(\psi, -\,A(y-1)^{-1}\,\psi(0)\big).$$

See [2, pp. 636–639] for more details.

In the case where we are looking at $(p=1,\, q=0)$, (3.43) is equivalent to

$$\prod_{m=1}^{N} (\,\sigma\mu \,-\, a^m_\pm \,+\, a^m_\pm e^{-\mu} \,-\, \lambda\,) \,=\, 0,$$

with $a^m_\pm := a^k(u_\pm)$. We give below some preliminary results about this characteristic equation. Some of them are not required for constructing the Evans function but will be useful later.

LEMMA 3.11. *For $1 \leq m \leq N$, we consider the equation*

$$\sigma\mu - a^m_\pm + a^m_\pm e^{-\mu} - \lambda = 0. \tag{3.47}$$

For $\mathrm{Re}\lambda > 0$, (3.47) *admits a unique solution* μ *of positive real part, which depends analytically on* λ. *We denote it by* $\mu_\pm^m(\lambda)$. *It can be continued to the strip*

$$\Omega := \{\lambda,\, \mathrm{Re}\lambda \geq -\eta, |\mathrm{Im}\lambda| \leq \pi\sigma\}$$

for $\eta > 0$ *small enough.*

The root $\mu_\pm^m(\lambda)$ *remains bounded away from the imaginary axis when* λ *goes to* 0 *if and only if* $a_\pm^m > \sigma$.

On the contrary, if $a_\pm^m < \sigma$, $\mu_\pm^m(\lambda)$ *goes to* 0 *when* λ *goes to* 0 *and we have the expansion*

$$\mu_\pm^m(\lambda) = -\frac{\lambda}{a_\pm^m - \sigma} + \frac{a_\pm^m\,\lambda^2}{2(a_\pm^m - \sigma)^3} + \mathcal{O}(\lambda^3). \tag{3.48}$$

Returning to the case $a_\pm^m > \sigma$, *we also find a solution of* (3.47) *that behaves as in* (3.48) *when* λ *goes to* 0. *We denote this solution by* $\mu_\pm^{N+m}$, *which is of course of negative real part when* $\mathrm{Re}\lambda > 0$. *With the convention that* $a^{N+m} = a^m$, *these new roots satisfy the expansion in* (3.48). *In view of* (2.11), *there are* $N-k+1$ *such roots,* $\mu_-^{N+k}, \dots, \mu_-^{2N}$, *for the* $\pm = -$ *sign, and* $N-k$ *such roots,* $\mu_+^{N+k+1}, \dots, \mu_+^{2N}$, *for the* $\pm = +$ *sign.*

The root $\mu_\pm^m(\lambda)$ *is simple, provided that* $\lambda \neq \sigma\,(\log(a_\pm^m/\sigma) - a_\pm^m/\sigma + 1)$.

If $a_\pm^m > \sigma$, *then two of the above roots collide at* $\lambda = \sigma\,(\log(a_\pm^m/\sigma) - a_\pm^m/\sigma + 1)$, *namely,* $\mu_\pm^m = \mu_\pm^{N+m} = \log(a_\pm^m/\sigma)$ *(the indices* m *lying in* $\{k, \dots, N\}$ *for the* $\pm = -$ *sign and in* $\{k+1, \dots, N\}$ *for the* $\pm = +$ *sign). So, up to diminishing* η, *all roots* $\mu_\pm^m(\lambda)$ *are simple and analytic in* $\lambda \in \Omega$.

Finally, up to diminishing again η, *we can find* $r > \eta$ *so that*

$$(\Omega \backslash C_r) \cap \Sigma_{ess}(L) = \emptyset, \tag{3.49}$$

with

$$C_r := \{\lambda,\, \sup(|\mathrm{Re}\lambda|, |\mathrm{Im}\lambda|) \leq r\} \tag{3.50}$$

Proof. The first part is contained in [2, Lemma 12]. The next part follows from bifurcation analysis of (3.47) at $(\lambda, \mu) = (0, 0)$ and elementary calculations. The last assertion is a consequence of Lemma 3.4. The reader may find it helpful to refer to Figures 3.1 and 3.2. ∎

Thanks to these properties of the eigenvalues of $\mathbb{A}_\pm(\lambda)$, we can obtain decompositions of H_0 similar to those used in [2] for constructing our Evans function D. In fact, the construction of D made in [2], valid in $\{\mathrm{Re}\lambda > 0\} \cup \{0\}$, did not require eigenvectors associated with the $\mu_\pm^{N+m}$. But we shall need the modes $\mu_\pm^{N+m}$ to extend the Green's function to a full neighborhood of 0 (namely, C_r). This is why we state a result that is slightly different from Lemma 3 in [2].

LEMMA 3.12. *In the framework of Lemma* 3.11, *we introduce for* $\lambda \in \Omega$,

$$\Phi_\pm^m(\lambda) := \begin{pmatrix} \theta \mapsto e^{\mu_\pm^m(\lambda)\theta}\, r_\pm^m \\ r_\pm^m \end{pmatrix},$$

$$\Psi_\pm^m(\lambda) := \begin{pmatrix} \theta \mapsto -\,e^{-\overline{\mu_\pm^m(\lambda)}\theta}\, a_\pm^m\, \ell_\pm^m \quad , \quad \sigma\, \ell_\pm^m \end{pmatrix} / \sigma,$$

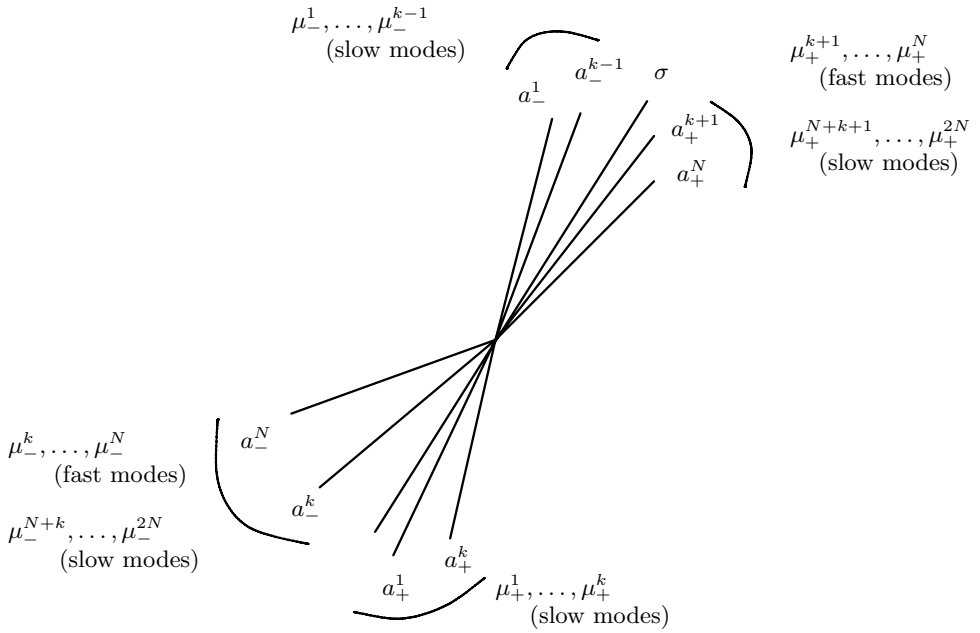


FIG. 3.1. *Characteristic speeds (in the (x,t)-plane) and the corresponding modes.*

which are eigenvectors of $\mathbb{A}_\pm(\lambda)$ and $\mathbb{A}_\pm(\lambda)^*$, respectively, associated with the eigenvalues $\mu_\pm^m(\lambda)$ and $\overline{\mu_\pm^m(\lambda)}$, and such that

$$\langle\, \Psi_\pm^n(\lambda)\,,\, \Phi_\pm^m(\lambda)\,\rangle \,=\, \delta_m^n.$$

(Here and below δ_m^n stands for the usual Kronecker symbol.) We have the decompositions

$$H_0 \,=\, \mathit{Span}\{\Phi_-^1(\lambda), \dots, \Phi_-^N(\lambda), \Phi_-^{N+k}(\lambda), \dots, \Phi_-^{2N}(\lambda)\}$$
$$\oplus\, \mathit{Span}\{\Psi_-^1(\lambda), \dots, \Psi_-^N(\lambda), \Psi_-^{N+k}(\lambda), \dots, \Psi_-^{2N}(\lambda)\}^\perp,$$

$$H_0 \,=\, \mathit{Span}\{\Phi_+^1(\lambda), \dots, \Phi_+^N(\lambda), \Phi_+^{N+k+1}(\lambda), \dots, \Phi_+^{2N}(\lambda)\}$$
$$\oplus\, \mathit{Span}\{\Psi_+^1(\lambda), \dots, \Psi_+^N(\lambda), \Psi_+^{N+k+1}(\lambda), \dots, \Psi_+^{2N}(\lambda)\}^\perp,$$

and the corresponding projections onto $\mathit{Span}\{\Phi_\pm^m(\lambda)\}$, $\mathit{Span}\{\Psi_\pm^m(\lambda)\}^\perp$ are analytic in λ and $\overline{\lambda}$, respectively, being defined by

$$Q_\pm(\lambda)\cdot\Phi \,=\, \sum_m \langle\, \Psi_\pm^m(\lambda)\,,\, \Phi\,\rangle\, \Phi_\pm^m(\lambda),$$

$$P_\pm(\lambda) \,=\, I_{H_0} \,-\, Q_\pm(\lambda).$$

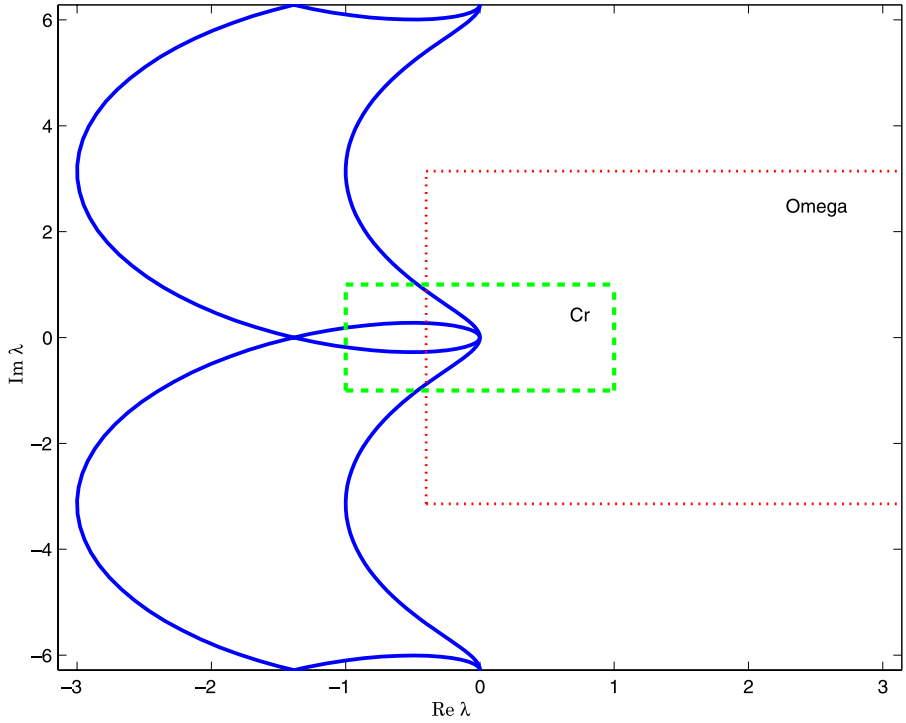


FIG. 3.2. *The sets Ω and C_r in the λ-plane. Thick curves represent typical plottings of $\Lambda(u)$ (with $\sigma = 1$), having equations of the form $\lambda = i\sigma\xi - a + ae^{-i\xi}$ ($\xi \in \mathbb{R}$) with either $a < \sigma$ (case without double points) or $a > \sigma$. For $u = u_\pm$, they signal that a root μ of (3.47) crosses the imaginary axis. Directions of crossings at $\lambda = 0$ are explicitly given in Lemma 3.11. From left to right across such a curve, there is a root μ of $\sigma\mu - a + ae^{-\mu} = \lambda$ crossing from left to right if $a < \sigma$ and from right to left if $a > \sigma$.*

Additionally, since all the eigenvalues of $\mathbb{A}_\pm(\lambda)$ but possibly the $\mu_\pm^m(\lambda)$ (the indices m lying in $\{1,\dots,N\} \cup \{N+k,\dots,2N\}$ for the $\pm = -$ sign and in $\{1,\dots,N\} \cup \{N+k+1,\dots,2N\}$ for the $\pm = +$ sign) lie in some fixed half-plane $\{\mathrm{Re}\mu \le \gamma_\pm < 0\}$ for all $\lambda \in C_r$, the semigroup associated with $\mathbb{A}_\pm(\lambda)$, $T_\lambda^\pm$, is exponentially decaying when restricted to the range of $P_\pm(\lambda)$. More precisely, for all $\alpha_\pm < \gamma_\pm$ there exists $C > 0$ so that

$$\left\| T_\lambda^\pm(x)\, P_\pm(\lambda) \right\| \le Ce^{-\alpha_\pm x} \quad \forall x \ge 0, \quad \forall \lambda \in C_r. \tag{3.51}$$

Proof. The proof is identical to that of Lemma 3 in [2]. ∎

Now, using the gap lemma construction, we can deduce decompositions of the space H_0 adapted to $\mathbb{A}(x,\lambda)$ on both half-lines $\mathbb{R}^\pm$.

For convenience, we shall use the following shortcuts:

$$M_- := \{1,\dots,N\} \cup \{N+k,\dots,2N\},$$

$$M_+ := \{1,\dots,N\} \cup \{N+k+1,\dots,2N\}.$$

LEMMA 3.13. *In the framework of Lemma 3.12, for all $m \in M_\pm$ there exist unique $\mathbb{W}_\pm^m$ and $\mathbb{Y}_\pm^m$ that are solutions on $\mathbb{R}^\pm$ of (3.40) and (3.45), respectively, having the*

following asymptotics:

$$\mathbb{W}^m_\pm(x,\lambda) = e^{\mu^m_\pm(\lambda)x}\Big(\Phi^m_\pm(\lambda) + \mathcal{O}(e^{\omega_\pm x})\Big) \quad \forall x \in \mathbb{R}^\pm, \tag{3.52}$$

$$\mathbb{Y}^m_\pm(x,\lambda) = e^{-\overline{\mu^m_\pm(\lambda)}x}\Big(\Psi^m_\pm(\lambda) + \mathcal{O}(e^{\omega_\pm x})\Big) \quad \forall x \in \mathbb{R}^\pm, \tag{3.53}$$

for some uniform $\omega_+ < 0 < \omega_-$ on compact subsets of Ω, are analytic in λ and satisfy the orthogonality relations

$$\langle\, \mathbb{Y}^n_\pm(x,\lambda)\,,\, \mathbb{W}^m_\pm(x,\lambda)\,\rangle \,=\, \delta^n_m. \tag{3.54}$$

Additionally, the projection onto Span$\{\mathbb{Y}^m_\pm(x,\lambda)\}^\perp$ defined by

$$P_\pm(x,\lambda)\cdot\Phi \,=\, \Phi \,-\, \sum_{m\in M_\pm} \langle\, \mathbb{Y}^m_\pm(x,\lambda)\,,\, \Phi\,\rangle\, \mathbb{W}^m_\pm(x,\lambda)$$

is such that the evolution operator $T_\lambda(x,y)$ associated with (3.40) *enjoys the following*

(i) *commutation property, $T_\lambda(x,y)P_\pm(y,\lambda) = P_\pm(x,\lambda)T_\lambda(x,y)$ for all $x \geq y \in \mathbb{R}_\pm$,*

(ii) *and decay estimate*

$$\|T_\lambda(x,y)P_\pm(y,\lambda)\| \leq Ce^{-\alpha_\pm(x-y)} \quad \forall x \geq y \in \mathbb{R}^\pm \tag{3.55}$$

for some $C \geq 0$ and $\alpha_\pm > 0$ independent of $\lambda \in C_r$.

Proof. The construction of the $\mathbb{W}^m_-$ (respectively, the $\mathbb{Y}^m_+$) was made in Proposition 2 (respectively, Proposition 3) and Lemma 7 in [2]. The construction of $\mathbb{W}^m_+$ for $x \geq x_0$ and of $\mathbb{Y}^m_-$ for $x \leq -x_0$ works the same way for x_0 large enough, through (revisited) Duhamel's principle and a fixed point argument. The main difficulty lies in extending $\mathbb{W}^m_+$ to the whole half-line $x \geq 0$ and $\mathbb{Y}^m_-$ to $x \leq 0$, which amounts to solving the backward Cauchy problem for (3.40) with data $\mathbb{W}^m_+(x_0)$ and the (forward) Cauchy problem for (3.45) with data $\mathbb{Y}^m_-(x_0)$. This can be overcome by using Lin's method [12, Theorem 3.3], provided that his basic assumptions hold, namely,

$$T_\lambda(x,y)\cdot\Phi \,\neq\, 0 \quad \forall x \geq y \in \mathbb{R}^-, \quad \forall \Phi \in H_0\backslash\{0\},$$

$$T^*_\lambda(x,y)\cdot\Psi \,\neq\, 0 \quad \forall x \leq y \in \mathbb{R}^+, \quad \forall \Psi \in H^*_0\backslash\{0\}.$$

In terms of the operator L, this requires that no (nontrivial) eigenfunction of L or L^* vanishes on an interval of length 1. Because of the invertibility of $A(x)$, an elementary calculation shows that this is obviously true.[5] ∎

Away from the essential spectrum, and in particular in $\Omega\backslash C_r$, we recover usual exponential dichotomies on $\mathbb{R}^\pm$ through the following alternate version of Lemma 3.13 (already pointed out in [2]).

LEMMA 3.14. *In the framework of Lemma* 3.13*, we define*

$$\Pi_\pm(x,\lambda) = I - \sum_{1\leq m\leq N} \mathbb{W}^m_\pm(x,\lambda)\mathbb{Y}^m_\pm(y,\lambda)$$

[5] For the more general, mixed-type operator, the same is true, provided that the extreme matrices, $C^q(x)$ and $C^{-p+1}(x)$, are invertible for all $x \in \mathbb{R}$. This is a crucial hypothesis, called *atomicity* in [17], to get exponential dichotomies; also see Hypothesis 1 in [10].

and

$$\Xi_\pm(x,\lambda) = I - \Pi_\pm(x,\lambda) = \sum_{1\le m\le N} \mathbb{W}^m_\pm(x,\lambda)\mathbb{Y}^m_\pm(x,\lambda), \tag{3.56}$$

where by abuse of notation all terms of the kind $\mathbb{W}\,\mathbb{Y}$ *stand for the operator*

$$\Phi \;\mapsto\; \langle\, \mathbb{Y}\,,\, \Phi\,\rangle\, \mathbb{W}.$$

Then we have

(i) $T_\lambda(x,y)\,\Pi_\pm(y,\lambda) = \Pi_\pm(x,\lambda)\,T_\lambda(x,y)$ *for all* $x \ge y \in \mathbb{R}^\pm$,

(ii) *and for every* $0 < r < R$, *there exist* $\alpha(r,R) > 0$ *and* $C(r,R) > 0$ *such that for all* $\lambda \in \Omega\backslash C_r$, $\operatorname{Re}\lambda \le R$,

$$\begin{aligned} &||T_\lambda(x,y)\,\Pi_\pm(y,\lambda)|| \le C(r,R)e^{-\alpha(r,R)(x-y)} \quad \forall x \ge y,\ x,y \in \mathbb{R}^\pm,\\ &||T_\lambda(y,x)\,\Xi_\pm(x,\lambda)|| \le C(r,R)e^{-\alpha(r,R)(x-y)} \quad \forall x \ge y,\ x,y \in \mathbb{R}^\pm. \end{aligned} \tag{3.57}$$

In the above inequality, $T_\lambda(y,x)\,\Xi_\pm(x,\lambda)$ *is to be understood as*

$$\big(T_\lambda(x,y)|_{\mathcal{R}\Xi_\pm(y,\lambda)}\big)^{-1}\,\Xi_\pm(x,\lambda).$$

The construction of the Evans function relies on this lemma. We recall it for completeness. By Lemma 3.4, Proposition 3.7, and Theorem 3.8, $\lambda \in \{\mathrm{Re}\lambda \ge 0\}\backslash\Sigma_{ess}(L)$ is an eigenvalue of L if and only if there exists a nonzero solution $\mathbb{W}$ of (3.40) vanishing at $\pm\infty$. Using the exponential dichotomies given by (3.57), we must have

$$\mathbb{W}(x) = T_\lambda(x,0)\cdot\Phi \quad \forall x \in \mathbb{R}$$

and $\Phi \in \mathcal{R}\,\Pi_+(0,\lambda) \,\cap\, \mathcal{R}\Xi_-(0,\lambda)$. Consequently, we have a nonzero solution if and only if there is a nonzero linear combination $\Phi = \sum_{m=1}^N \varphi^m \mathbb{W}^m_-(0,\lambda)$ belonging to $\mathcal{R}\,\Pi_+(0,\lambda) = \mathrm{Span}\{\mathbb{Y}^m_+(x,\lambda);\ m = 1,\dots,N\}^\perp$. This equivalently means that the linear system

$$B(\lambda)\,\varphi = 0,$$

where $\varphi = (\varphi^1,\dots,\varphi^N)$ and $B(\lambda)$ is the square matrix of coefficients

$$B(\lambda)_{n,m} = \langle\, \mathbb{Y}^n_+(0,\lambda)\,,\, \mathbb{W}^m_-(0,\lambda)\,\rangle, \quad 1 \le n\,,\, m \le N,$$

has a nontrivial solution. This yields as in [1] (also see [4] for a discussion of mixed-type Evans functions in finite dimensions) the following definition.

Definition 3.15. *With the notations introduced in Lemma* 3.13*, we define an Evans function associated with* L *by*

$$D(\lambda) := \det(B(\lambda)) = \det(\langle\, \mathbb{Y}^n_+(0,\lambda)\,,\, \mathbb{W}^m_-(0,\lambda)\,\rangle)_{1\le m,n\le N}. \tag{3.58}$$

A useful remark in practical computations is the following. By construction of $\mathbb{W}^m_-$, a solution of (3.40), and of $\mathbb{Y}^n_+$, a solution of the adjoint system (3.45), we see that

$$B(\lambda)_{n,m} = \langle\, \mathbb{Y}^n_+(x,\lambda)\,,\, \mathbb{W}^m_-(x,\lambda)\rangle$$

for all x, where both $\mathbb{W}^m_-$ and $\mathbb{Y}^n_+$ are well defined (which does occur for some of the indices, as we shall recall below). Not only D vanishes at the eigenvalues of L, but

we have the following result relating the multiplicity of an eigenvalue to its order as a root of D.

PROPOSITION 3.16. *An eigenvalue of L of positive real part in L^2 has algebraic multiplicity at most equal to the order of vanishing of the Evans function D, defined in* (3.58)*, at that point.*

Proof. The proof is inspired by that of Gardner and Jones [7] in the more classical, finite dimensional case.

For clarity, we first show that *geometric* multiplicity is at most the order of vanishing. Assume that a given eigenvalue λ_0 is of geometric multiplicity K. Then there exist independent $W^1, \dots, W^K$ and independent $Y^1, \dots, Y^K$ such that

$$(L - \lambda_0) \cdot W^m = 0, \quad (L - \lambda_0)^* \cdot Y^m = 0$$

for all $m \le K$. Using the standard notation $W_x(\theta) = W(x + \theta)$ and introducing the mappings

$$\begin{array}{rccl} \mathcal{I}: & H^1(\mathbb{R}) & \to & \mathcal{C}(\mathbb{R}; H_1) \\ & W & \mapsto & \mathbb{W} : x \mapsto \begin{pmatrix} W_x \\ W(x) \end{pmatrix}, \end{array}$$

$$\begin{array}{rccl} \mathcal{J}: & H^1(\mathbb{R}) & \to & \mathcal{C}(\mathbb{R}; H^*_{1,\cdot}) \\ & Y & \mapsto & \mathbb{Y} : x \mapsto \begin{pmatrix} -Y_x A_{x-1} & , & \sigma Y(x) \end{pmatrix}, \end{array}$$

we have that $\mathcal{I}W^n$ (respectively, $\mathcal{J}Y^n$) are global solutions of (3.40) (respectively, (3.45)) (see [2] for more details), which vanish at both $+\infty$ and $-\infty$. This means we can modify the $\mathbb{Y}^m$ and $\mathbb{W}^m$ without changing the Evans function in such a way that

$$\mathbb{Y}^m(\pm\infty, \lambda_0) = 0, \quad \mathbb{W}^m(\pm\infty, \lambda_0) = 0 \quad \forall m \le K.$$

This implies that the matrix

$$(\langle\, \mathbb{Y}^n_\pm(x, \lambda)\,,\, \mathbb{W}^m_\pm(x, \lambda)\,\rangle)_{1 \le m,n \le N}$$

has at least K null rows (and K null columns) at $\lambda = \lambda_0$. Differentiating, at most $m-1$ times, its determinant row by row, there always remains a null row. This shows that

$$\frac{d^l D}{d\lambda^l}(\lambda_0) = 0 \quad \forall l \le K - 1.$$

To show that *algebraic* multiplicity is at most the order of vanishing, we should use Jordan chains for L and L^*. This is a little cumbersome in general. For simplicity, we present the proof in the case when the algebraic multiplicity of λ_0 equals 2 (and the geometric multiplicity equals 1; otherwise the previous case applies). Then there exist independent V and W, and independent Y and Z, such that

$$(L - \lambda_0) \cdot W = 0, \quad (L - \lambda_0)^* \cdot Y = 0,$$

$$(L - \lambda_0) \cdot V = W, \quad (L - \lambda_0)^* \cdot Z = Y.$$

Similarly as before, we replace one of the $\mathbb{W}^m$ and one of the $\mathbb{Y}^m$, for instance with $m = 1$, by $\mathbb{W} = \mathcal{I}W$ and $\mathbb{Y} = \mathcal{J}Y$, respectively. We easily see that

$$\mathcal{I}V - \frac{\partial \mathbb{W}^1}{\partial \lambda}(\lambda_0)$$

is a solution of (3.40), which vanishes at $-\infty$. Therefore, there exist coefficients $\varphi^1, \dots, \varphi^N$ so that

$$\mathbb{V} - \frac{\partial \mathbb{W}^1}{\partial \lambda}(\lambda_0) = \sum_{m=1}^{N} \varphi^m \, \mathbb{W}^m(\lambda_0)$$

with $\mathbb{V} := \mathcal{I}V$. So, replacing $\mathbb{W}^1$ by

$$\mathbb{W}^1 + \sum_{m=1}^{N} \varphi^m (\lambda - \lambda_0) \mathbb{W}^m$$

does not alter D and makes

$$\mathbb{V} = \frac{\partial \mathbb{W}^1}{\partial \lambda}(\lambda_0).$$

Exactly the same argument shows that we can assume without loss of generality that[6]

$$\mathbb{Z} := \mathcal{J}Z = \frac{\partial \mathbb{Y}^1}{\partial \lambda}(\lambda_0).$$

The interest is again that they vanish at both $+\infty$ and $-\infty$. This way, we find that $D'(\lambda_0)$ is proportional to

$$\frac{d}{d\lambda} \langle \mathbb{Y}^1, \mathbb{W}^1 \rangle (\lambda_0) = \langle \mathbb{Z}, \mathbb{W} \rangle + \langle \mathbb{Y}, \mathbb{V} \rangle = 0.$$

Hence $D'(\lambda_0) = 0$.

This proves that multiplicity of λ_0 is at most its order as a root of D. Due to the infinite dimensions, it is not clear whether there is equality though. As a matter of fact, assuming, for instance, that multiplicity equals 1 and order equals 2, we should get a contradiction by constructing independent V and W vanishing at $\pm\infty$ such that

$$(L - \lambda_0) \cdot W = 0, \quad (L - \lambda_0) \cdot V = W.$$

But we have been able only to construct V such that

$$\langle \mathbb{Y}^m, \mathbb{V} \rangle (+\infty) = 0 \quad \forall m \le N,$$

which is from far being sufficient for $\mathbb{V}$ to vanish at $+\infty$. (This would be sufficient if $\mathbb{V}$ were a solution of the homogeneous system (3.40).) □

In fact, the very same proof works at $\lambda = 0$, which is of course a zero of D because of the translation invariance (this means that we may choose one of the $\mathbb{W}^m$ to be $\mathcal{I}U'$, the derivative of the profile, since $L \cdot U' = 0$ and $U'(-\infty) = 0$). Thus, if $D'(0)$ is nonzero, 0 is a simple eigenvalue of L in L^2. Under the additional, generic assumption

(H4) $\mathrm{Span}(\ell_+^1, \dots, \ell_+^k, \ell_-^k, \dots, \ell_-^N) = (\mathbb{R}^N)^*$,

it was shown in [2] (also see [4]) how to compute $D'(0)$. We found that

$$D'(0) = h^k \cdot (u_+ - u_-) \det(\langle \mathbb{Y}_+^n(0,\lambda), \mathbb{W}_-^m(0,\lambda) \rangle)_{k+1 \le n, m \le N}$$

[6]The awkward notation $\mathbb{Z}$ here has nothing to do with the set of integers.

up a to a nonzero factor, where h^k belongs to $\mathrm{Span}(\ell_+^1,\dots,\ell_+^k)\cap\mathrm{Span}(\ell_-^k,\dots,\ell_-^N)$. The reduced determinant

$$\det(\langle\, \mathbb{Y}_+^n(0,\lambda)\,,\, \mathbb{W}_-^m(0,\lambda)\,\rangle)_{k+1\le n\,,\,m\le N}$$

can be viewed as a transversality coefficient. For its vanishing means the existence of a square integrable eigenfunction of L independent of $\mathbb{U}'=\mathbb{W}_k$ (recall that $\mathbb{Y}_+^n$ for $n\ge k+1$ and W_-^m for $m\ge k$ are exponentially decaying at $\pm\infty$). Of course, in the case $k=N$, that reduced determinant is irrelevant. In general, the other factor in $D'(0)$, $h^k\cdot(u_+-u_-)$, is related to Majda's one-dimensional stability condition

$$\det(\, r_-^1,\dots,r_-^{k-1},u_+-u_-,r_+^{k+1},\dots,r_+^N\,)\ \neq\ 0 \tag{3.59}$$

for the exact shock. Indeed, elementary linear algebra shows that

$$\begin{aligned}(h^k)^\perp\ &=\ \mathrm{Span}(\ell_+^1,\dots,\ell_+^k)^\perp\ \oplus\ \mathrm{Span}(\ell_-^k,\dots,\ell_-^N)^\perp\\ &=\ \mathrm{Span}(r_+^{k+1},\dots,r_+^N)\oplus\mathrm{Span}(r_-^1,\dots,r_-^{k-1}).\end{aligned}$$

Hence $h^k\cdot(u_+-u_-)\neq 0$ if and only if Majda's condition (3.59) holds true. This is in particular always the case for weak shocks.

To summarize, we have the following.

PROPOSITION 3.17. *Under the assumption* (H4), *the Evans function D defined in* (3.58) *admits a simple zero at $\lambda=0$ if and only if* 0 *is a simple eigenvalue of L in L^2* and *the underlying shock is Majda stable; that is,* (3.59) *is satisfied. For (extreme) N-shocks, $D'(0)\neq 0$ is merely equivalent to Majda's stability condition, which equivalently reads*

$$\ell_-^N\cdot(u_+-u_-)\ \neq\ 0.$$

Theorem 3.2 *does not apply.* Anyway, $D'(0)\neq 0$ does not imply that 0 is a simple eigenvalue of L in L^∞. As we show hereafter, it cannot be that 0 is simple in L^∞, which in view of Theorem 3.8 implies that the requirement (ii) in Theorem 3.2 fails, unless $N=1$. The argument is similar as for viscous shocks [24, 20] and holds for general schemes, just assuming that 0 is a simple eigenvalue of L in L^2.

It is based on the fact that the eigenvalue equation $L\cdot Y=0$ for $Y\in W^{1,\infty}$ is equivalent to $M\cdot Y=\mathrm{const}$, where M is the integral operator defined by

$$(M\cdot Y)(x)\ :=\ \sigma\,Y(x)\ -\ \int_{x-1}^{x}\ \sum_{l=-p}^{q}\ C^l(s)\ \cdot V(s+l)\,ds.$$

In fact, this operator already arose in the proof of Proposition 2.4 (see (2.25)). Contrary to L, the operator M is asymptotically hyperbolic, in the sense that the limit operators

$$(M_\pm\cdot Y)(x)\ :=\ \sigma\,Y(x)\ -\ \int_{-1}^{0}\ \sum_{l=-p+1}^{q}\ C_\pm^l\ \cdot V(x+\theta+l)\,d\theta$$

have no purely imaginary spectrum. As a matter of fact, by Fourier transform (in $\mathcal{S}'$), we easily see that the spectrum of $M_\pm$ is made of the roots μ of $\det(\Delta(\sigma;\mu,u_\pm)/\mu)$. As already observed, there are no such μ on the imaginary axis. By exactly the same method as Mallet-Paret in [16], one can show that M is Fredholm and its index

depends only on $M_\pm$. Now, as in the proof of Lemma 3.4 concerning L, we can compute the index of M by a homotopy argument. This works, provided that there exists a smooth curve $\theta \in [-1,1]$ connecting u_- at $\theta = -1$ to u_+ at $\theta = 1$, along which a^k is monotone (decreasing). Note that this assumption is generically harmless. For instance, that curve can simply be taken as the profile itself in the framework of Proposition 2.4. The constant coefficient operator about $u(\theta)$, M_θ, is hyperbolic for all θ but one, the unique θ such that $a^k(u(\theta)) = \sigma$, as computations in the proof of Theorem 3.6 show. This means there is exactly one eigenvalue μ of M_θ crossing the imaginary axis when θ is varied from -1 to 1. Therefore, M is Fredholm index 1.

Because of the translation invariance, the derivative of the profile, U', belongs to the kernel of M. Because of the asymptotic hyperbolicity of M, we expect that its kernel in L^∞ coincides with its kernel in L^2. Assuming that 0 is a simple eigenvalue of L in L^2, we get that the kernel of M in L^∞ is exactly of dimension 1. By the Fredholm property, this implies that M is onto. More precisely, there is a codimension 1 subspace of L^∞, say H, such that $M_{|H}$ is an isomorphism.

In particular, for any constant $c \in \mathbb{C}^N$, there exists a unique $Y \in H$ such that $M \cdot Y = c$, and thus, differentiating once, $L \cdot Y = 0$. This means that the L^∞ kernel of L is at least N-dimensional.

4. Pointwise Green's function bounds for the linear LDS.

4.1. Derivation of the Green's function of the linear LDS. In the previous section, we have shown that the spectral assumption (i) of Theorem 3.2 is true for weak shocks and also that $D'(0) \neq 0$ at least for extreme shocks. We claim that $D'(0) \neq 0$ is an appropriate, weaker assumption in place of (i) to get nonlinear orbital stability in conservative semidiscrete schemes, provided that an Evans function is available. To support this, we show how to adapt to our setting the method of Zumbrun and Howard [24], relying on pointwise Green's functions estimates.

The first difficulty lies in the fact that the linearized system (3.2) is *not* autonomous, which prevents us from directly applying Laplace transform. This brings back into favor the change of frame point of view. Indeed, the evolution system (3.9) is autonomous. But we need a more precise link, in terms of Green's functions, between the two approaches. This part works for any scheme.

We call Green's function of the linear LDS (3.2) the solution $v_j(t) = G_j^{j_0}(t,t_0)$ of

$$\left\{ \begin{array}{l} \dfrac{dv}{dt} = D\mathcal{G}(u) \cdot v, \\ \\ v_j(t_0) = \delta_j^{j_0}. \end{array} \right. \tag{4.1}$$

This definition is natural. The general solution of the nonhomogeneous equation

$$\frac{dv}{dt} = D\mathcal{G}(u) \cdot v + F(t)$$

is indeed given by the convolution relation

$$v_j(t) = \sum_{j_0} G_j^{j_0}(t,0) v_{j_0}(0) + \int_0^t \sum_{j_0} G_j^{j_0}(t,t_0) F_{j_0}(t_0)\, dt_0.$$

Consequently, the behavior of the Green's function determines the behavior of general solutions of the LDS.

By definition, we have $G_j^{j_0}(t,t_0) = V(j-st,t/\Delta x)$, where V solves (3.9) and satisfies $V(x,t_0/\Delta x) = 1$ if $x = j_0 - st_0$ and $V(x,t_0/\Delta x) = 0$ if $|x - j_0 - st_0| \geq 1$. To simplify the writing, we perform a rescaling in space and time so that $\Delta x = 1$ and $\sigma = s$.

Taking the Laplace transform in t of (3.9), we get

$$(\lambda - L)\,\mathcal{L}[V(\cdot + t_0)](\lambda) = V(t_0).$$

(As a usual abuse of notation, $V(t)$ stands for the function $x \mapsto V(x,t)$.) We know from Proposition 3.3 that $(\lambda - L)$ is invertible for $\mathrm{Re}\lambda > M$ large enough. Then using the Laplace inversion formula, we get for $\gamma > M$

$$V(t) = \frac{1}{2\,i\,\pi}\int_{\gamma-i\infty}^{\gamma+i\infty} \mathrm{e}^{\lambda(t-t_0)}\,(\lambda - L)^{-1}\,V(t_0)\,d\lambda, \quad t > t_0. \tag{4.2}$$

(Since L is the infinitesimal generator of a strongly continuous semigroup, by Proposition 3.3, this formula holds true for all $t > t_0$, provided that $V(t_0)$ belongs to the domain of L, that is, H^1.) At this stage, we may try to use directly (4.2) to get estimates of $V(x,t)$ and then deduce estimates of $G_j^{j_0}(t,t_0) = V(j-st,t)$. But, as shown in Theorem 3.6, the essential spectrum of L intersects the imaginary axis infinitely many times on $2i\pi\sigma\mathbb{Z}$. This way, it seems difficult to use the method of [24], which relies on a very careful continuation of $(\lambda - L)^{-1}V(t_0)$ into the essential spectrum. This is why we use only (4.2) as an intermediate step for obtaining a representation formula of $G_j^{j_0}(t,t_0)$.

We begin with a formal derivation based on the obvious and nevertheless crucial conjugation formula

$$(\lambda + 2i\pi\sigma m - L)^{-1} = E^m\,(\lambda - L)^{-1}\,E^{-m},$$

where E^m denotes the multiplication operator by the function $x \mapsto \mathrm{e}^{2i\pi m\,x}$. We deduce from (4.2) that

$$V(x,t) = \frac{1}{2i\pi}\sum_{m\in\mathbb{Z}}\int_{\gamma-i\pi\sigma}^{\gamma+i\pi\sigma} \mathrm{e}^{(\lambda+2i\pi\sigma m)(t-t_0)}\mathrm{e}^{2i\pi mx}\left\{(\lambda - L)^{-1}(E^{-m}V(t_0))\right\}(x)\,d\lambda.$$

And, in particular,

$$V(j-st,t) = \frac{1}{2i\pi}\sum_{m\in\mathbb{Z}}\int_{\gamma-i\pi\sigma}^{\gamma+i\pi\sigma} \mathrm{e}^{\lambda(t-t_0)}\,\left\{(\lambda - L)^{-1}\,(\mathrm{e}^{-2i\pi smt_0}E^{-m}V(t_0))\right\}(j-st)\,d\lambda.$$

Now if we (formally) permute the sum and the integral, we get by linearity

$$V(j-st,t) = \frac{1}{2i\pi}\int_{\gamma-i\pi\sigma}^{\gamma+i\pi\sigma} \mathrm{e}^{\lambda(t-t_0)}\left\{(\lambda - L)^{-1}\left(\sum_{m\in\mathbb{Z}}\mathrm{e}^{-2i\pi smt_0}E^{-m}V(t_0)\right)\right\}(j-st)\,d\lambda.$$

Finally, recalling the definition of the Dirac comb and the alternative formula

$$\sum_{m\in\mathbb{Z}}\mathrm{e}^{-2i\pi smt_0}\,E^{-m} = \sum_{j\in\mathbb{Z}}\delta_{j-st_0},$$

we find that

$$\sum_{m\in\mathbb{Z}}\mathrm{e}^{-2i\pi smt_0}\,E^{-m}\,V(t_0) = \delta_{j_0-st_0}$$

in the sense of distributions. Consequently, the expected formula for the solution of (4.1) is

$$G_j^{j_0}(t,t_0) \,=\, \frac{1}{2\,i\,\pi\,\sigma}\int_{\gamma-i\pi\sigma}^{\gamma+i\pi\sigma} \mathrm{e}^{\lambda(t-t_0)}\,G_\lambda(j-st,j_0-st_0)\,d\lambda, \tag{4.3}$$

where G_λ denotes the Green function of $(\lambda - L)$, defined by

$$(\lambda - L)\,G_\lambda(\cdot,y) \,=\, \delta_y. \tag{4.4}$$

A useful remark is that the conjugation property of L translates to G_λ into the equality

$$G_{\lambda+2i\pi\sigma m}(x,y) \,=\, \mathrm{e}^{2i\pi m(x-y)}\,G_\lambda(x,y). \tag{4.5}$$

Also observe that the formula in (4.3) requires only integration on compact sets in λ. Moreover, if G_λ is holomorphic for $\mathrm{Re}\lambda > M$, then, because of (4.5) and Cauchy's theorem, the integral in (4.3) is independent of $\gamma > M$.

To justify the previous computation, we need estimates on G_λ. Revisiting Proposition 5.2 in [16], we have the following.

LEMMA 4.1. *We assume that* (H1)–(H3) *hold and consider the operator L defined in* (3.10). *Then for* $\mathrm{Re}\lambda > M$ *large enough, the Green's function G_λ defined by* (4.4) *enjoys a uniform estimate*

$$|G_\lambda(x,y)| \,\le\, K\,, \qquad (x,y)\in\mathbb{R}^2, \tag{4.6}$$

for some positive constant K. Furthermore, for all $\alpha > 0$ there exists M_α such that for $\mathrm{Re}\lambda > M_\alpha$

$$|G_\lambda(x,y)| \,\le\, K\,\mathrm{e}^{-\alpha|x-y|}\,, \qquad (x,y)\in\mathbb{R}^2. \tag{4.7}$$

Proof. The proof is based on finding G_λ through a fixed point argument. Recalling that $L \,=\, \sigma d/dx \,+\, B$ with

$$(B\cdot V)(x) \,=\, -\sum_{l=-p}^{q} (\,C^l(x) \,-\, C^{l+1}(x-1)\,)\cdot V(x+l)$$

and introducing the Green's function G_λ^0 of $(\lambda - \sigma d/dx)$, we can look for the solution of (4.4) as a solution of the implicit equation

$$G_\lambda(x,y) \,=\, G_\lambda^0(x,y) \,+\, \int_{\mathbb{R}} G_\lambda(x,z)\,(BG_\lambda^0)(z,y)\,dz. \tag{4.8}$$

Of course, the advantage of using G_λ^0 is that it is explicitly known, and in particular it meets the estimate

$$|G_\lambda^0(x,y)| \,\le\, \frac{1}{\sigma}\,\mathrm{e}^{-\mathrm{Re}\lambda|x-y|/\sigma}.$$

Now a rough estimate shows that for all $G\in L^\infty$

$$\left|\int_{\mathbb{R}} G(x,z)\,(BG_\lambda^0)(z,y)\,dz\right| \,\le\, c\,\|G\|_{L^\infty}\sum_{l=-p}^{q}\int_{\mathbb{R}}\mathrm{e}^{-\mathrm{Re}\lambda|z+l-y|/\sigma}\,dz \,\le\, \frac{c'}{|\mathrm{Re}\lambda|}\,\|G\|_{L^\infty}.$$

Therefore, the mapping $G \mapsto \int_{\mathbb{R}} G(\cdot,z)\,(BG^0_\lambda)(z,y)\,dz$ is a contraction in L^∞ for $\mathrm{Re}\lambda > c'$ large enough so that (4.8) has a unique bounded solution, which must be G_λ. Hence G_λ is bounded, as claimed in (4.6). The proof of (4.7) is identical, applying a fixed point argument in the Banach space

$$\mathcal{E}^\alpha := \{\, G\,;\, \sup_{x,y} |G(x,y)|\, \mathrm{e}^{\alpha|x-y|} < \infty \,\}.$$

We just evaluate

$$\left| \int_{\mathbb{R}} G(x,z)\,(BG^0_\lambda)(z,y)\,dz \right| \le c\,\|G\|_{\mathcal{E}^\alpha} \sum_{l=-p}^{q} \int_{\mathbb{R}} \mathrm{e}^{-\alpha|x-z|}\,\mathrm{e}^{-\mathrm{Re}\lambda|z+l-y|/\sigma}\,dz$$

$$\le c'\,\mathrm{e}^{-\alpha|x-y|}\,\|G\|_{\mathcal{E}^\alpha} \int_{\mathbb{R}} \mathrm{e}^{-(\mathrm{Re}\lambda-\alpha\sigma)|z-y|/\sigma}\,dz \le \frac{c''}{|\mathrm{Re}\lambda-\alpha\sigma|}\,\mathrm{e}^{-\alpha|x-y|}\,\|G\|_{\mathcal{E}^\alpha}.$$

Hence the mapping $G \mapsto \int_{\mathbb{R}} G(\cdot,z)\,(BG^0_\lambda)(z,\cdot)\,dz$ is a contraction in $\mathcal{E}^\alpha$ for $\mathrm{Re}\lambda > \alpha\sigma + c''$. ∎

THEOREM 4.2. *Assuming* (H1)–(H3), *for all* $t > t_0$ *the Green's function* $G^{j_0}_j(t,t_0)$ *of* (3.2) *is given by* (4.3) *in terms of the Green's function* G_λ *of* $(\lambda - L)$, *where* L *is the operator defined in* (3.10).

Proof. It will be useful to reformulate the definition (4.4) of G_λ as

$$\int_{\mathbb{R}} \sigma\,\zeta'(x)\,G_\lambda(x,y) + \lambda\,\zeta(x)\,G_\lambda(x,y) \tag{4.9}$$

$$+\zeta(x) \sum_{l=-p}^{q} (\,C^l(x) - C^{l+1}(x-1)\,)\cdot G_\lambda(x+l,y)\,dx = \zeta(y)$$

for all $\zeta \in \mathcal{C}^\infty_c(\mathbb{R})$.

By Cauchy–Lipschitz theorem (recall that (3.2) is merely a linear ODE in the Banach space $\mathcal{L}^\alpha$), we know that (4.1) admits a unique solution. We want to show that $v_j(t) := G^{j_0}_j(t,t_0)$ defined by (4.3) coincides with that solution on $\{\, t \ge t_0 \,\}$. Because of (4.6), we see by making γ tend to $+\infty$ in (4.3) that $v(t) = 0$ for $t < t_0$. Therefore, it is sufficient to prove that

$$\int_{-\infty}^{\infty} \sum_{j\in\mathbb{Z}} -\frac{dz_j}{dt}\,v_j + z_j \sum_{l=-p}^{q} (\,C^l(j-st) - C^{l+1}(j-1-st)\,)\cdot v_{j+l}\,dt = z_{j_0}(t_0) \tag{4.10}$$

for all $z \in \mathcal{C}^\infty_c(\mathbb{R}\,;\,\mathcal{L}^2)$. And this is a matter of simple calculus. We have

$$\int_{-\infty}^{\infty} \sum_{j\in\mathbb{Z}} -\frac{dz_j}{dt}\,v_j\,dt + z_j \sum_{l=-p}^{q} (\,C^l(j-st) - C^{l+1}(j-1-st)\,)\cdot v_{j+l}\,dt$$

$$= \frac{1}{2\,i\,\pi} \int_{\gamma-i\pi\sigma}^{\gamma+i\pi\sigma} d\lambda \sum_{j\in\mathbb{Z}} \int_{-\infty}^{\infty} dt\,\mathrm{e}^{\lambda(t-t_0)} \left\{ -\frac{dz_j}{dt}\,G_\lambda(j-st, j_0-st_0) \right.$$

$$\left. + z_j \sum_{l=-p}^{q} \big(\,C^l(j-st) - C^{l+1}(j-1-st)\,\big)\cdot G_\lambda(j+l-st, j_0-st_0) \right\}.$$

The permutation of summations is allowed by the estimate in (4.6), the fact that z is compactly supported, and Lebesgue's theorem. Now we see by change of variable that

$$\int_{t=-\infty}^{t=\infty} \mathrm{e}^{\lambda(t-t_0)} \left\{ -\frac{dz_j}{dt}\, G_\lambda(j-st, j_0-st_0) \right.$$

$$\left. +\, z_j \sum_{l=-p}^{q} \big(C^l(j-st) - C^{l+1}(j-1-st) \big) \cdot G_\lambda(j+l-st, j_0-st_0) \right\} dt$$

$$= \frac{1}{\sigma} \int_{x=-\infty}^{x=\infty} \sigma\, \frac{d\zeta_j}{dx}\, G_\lambda(x, j_0-st_0) + \lambda\, \zeta_j(x,\lambda)\, G_\lambda(x, j_0-st_0)$$

$$+ \zeta_j(x,\lambda) \sum_{l=-p}^{q} \big(C^l(x) - C^{l+1}(x-1) \big) \cdot G_\lambda(x+l, j_0-st_0)\, dx,$$

with

$$\zeta_j(x,\lambda) := \mathrm{e}^{\lambda(\,(j-x)/s - t_0\,)}\, z_j((j-x)/s).$$

Since z is compactly supported, (4.9) holds in particular for $\zeta_j(\cdot,\lambda)$. Hence we have

$$\int_{t=-\infty}^{t=\infty} \mathrm{e}^{\lambda(t-t_0)} \left\{ -\frac{dz_j}{dt}\, G_\lambda(j-st, j_0-st_0) \right.$$

$$\left. +\, z_j \sum_{l=-p}^{q} \big(C^l(j-st) - C^{l+1}(j-1-st) \big) \cdot G_\lambda(j+l-st, j_0-st_0) \right\} dt$$

$$= \frac{1}{\sigma}\, \zeta_j(j_0-st_0, \lambda).$$

To get (4.10) it remains to show that

$$\frac{1}{2\,i\,\pi\,\sigma} \int_{\gamma-i\pi\sigma}^{\gamma+i\pi\sigma} \sum_{j\in\mathbb{Z}} \zeta_j(j_0-st_0,\lambda)\, d\lambda = z_{j_0}(t_0).$$

But by definition of ζ_j we have

$$\int_{\gamma-i\pi\sigma}^{\gamma+i\pi\sigma} \sum_{j\in\mathbb{Z}} \zeta_j(j_0-st_0,\lambda)\, d\lambda = \int_{\gamma-i\pi\sigma}^{\gamma+i\pi\sigma} \sum_{j\in\mathbb{Z}} \mathrm{e}^{\lambda(j-j_0)/s}\, z_j((j-j_0)/s+t_0)\, d\lambda.$$

Noting that

$$\frac{1}{2\,i\,\pi\,\sigma} \int_{\gamma-i\pi\sigma}^{\gamma+i\pi\sigma} \mathrm{e}^{\lambda\, n/\sigma}\, d\lambda = \delta_n^0$$

we get the result by permuting the integral with the sum. ∎

4.2. Study of the Green's function G_λ for the upwind scheme. Similarly as in [24], our method consists of deriving estimates on the evolutionary Green's function $G_j^{j_0}(t,t_0)$ from precise bounds on the Green's function $G_\lambda(x,y)$, thanks to the formula (4.3).

From now on, we focus on the upwind scheme and make use of the Evans function defined in (3.58). We assume linear stability, and more precisely

(H5) $D(\lambda) \neq 0$ for all $\lambda \notin 2\,i\,\pi\,\sigma\,\mathbb{Z}$, $\mathrm{Re}\lambda \geq 0$,
(H6) $D'(0) \neq 0$.

Indeed, the same argument as in [24, Corollary 8.2] shows that both assumptions (H5) and (H6) are necessary for the linear stability of the semidiscrete shock.

Remark 4.3. Note that, thanks to (H5)–(H6), we can actually assume that

$$D(\lambda) \neq 0 \quad \forall \lambda \in \Omega \backslash C_r$$

and that 0 is the only root of the Evans function in C_r by choosing r and η smaller, thanks to the analyticity of the Evans function in Ω.

We already obtained in Lemma 4.1 some uniform estimates on $G_\lambda(x,y)$ for large $\mathrm{Re}\lambda$. The next step consists of constructing and decomposing G_λ into controllable terms for bounded $\mathrm{Re}\lambda$. This amounts to constructing and decomposing the Green's function $\mathbb{G}_\lambda(x,y)$ of the operator

$$\mathbb{L}_\lambda := \sigma \left(-\frac{d}{dx} + \mathbb{A}(x,\lambda) \right).$$

As a matter of fact, we have

$$(\lambda - L) \cdot W = f \quad \Longleftrightarrow \quad \mathbb{L}_\lambda \cdot \mathbb{W} = F,$$

where

$$\mathbb{W} = \mathcal{I} W, \quad \text{i.e.,} \quad \mathbb{W}(x) = \begin{pmatrix} W_x \\ W(x) \end{pmatrix} \quad \text{and} \quad F = \begin{pmatrix} 0 \\ f \end{pmatrix}.$$

Therefore, denoting as before

$$\begin{array}{rccc} \Gamma : & H_0 & \to & \mathbb{C}^N \\ & \begin{pmatrix} \phi \\ c \end{pmatrix} & \mapsto & c \end{array}$$

and its right inverse

$$\begin{array}{rccc} \Upsilon : & \mathbb{C}^N & \to & H_0 \\ & c & \mapsto & \begin{pmatrix} 0 \\ c \end{pmatrix}, \end{array}$$

we have

$$G_\lambda(x,y) = \Gamma\, \mathbb{G}_\lambda(x,y)\, \Upsilon. \tag{4.11}$$

Our approach is close to the original one of Zumbrun and Howard [24], Zumbrun [23], and Mascia and Zumbrun [18] and may be translated to other frameworks. However,

special care is required because of the infinite dimensions. In this context, the accurate definition of $\mathbb{G}_\lambda(x,y)$ is that for any $\mathbb{F} \in L^2(\mathbb{R}, H_0)$,

$$\mathbb{W}(x) = \int_{\mathbb{R}} \mathbb{G}_\lambda(x,y)\, \mathbb{F}(y)\, dy \tag{4.12}$$

is the unique *mild solution* (see, for instance, [19]) in $L^2(\mathbb{R}, H_0)$ of the nonhomogeneous problem

$$\sigma \Big(-\frac{d}{dx} + \mathbb{A}(x,\lambda) \Big) \mathbb{W}(x) \,=\, \mathbb{F}(x). \tag{4.13}$$

To get an expression of $\mathbb{G}_\lambda$ in $\Omega \backslash C_r$, we shall make a crucial use of the fact that $\mathbb{L}_\lambda$ has exponential dichotomies on $\mathbb{R}^+$ and $\mathbb{R}^-$ (see (3.57)).

PROPOSITION 4.4. *We define the projections* $\Pi_\pm$ *and* $\Xi_\pm$ *as in* (3.56). *Under the assumption* (H5), *we have*

$$\mathcal{R}\Pi_+(0,\lambda) \oplus \mathcal{R}\Xi_-(0,\lambda) \,=\, H_0 \quad \forall \lambda \in \Omega \backslash C_r. \tag{4.14}$$

Therefore, there exist $M_1(\lambda) : H_0 \to \mathcal{R}\Pi_+(0,\lambda)$ *and* $M_2(\lambda)$*:* $H_0 \to \mathcal{R}\Xi_-(0,\lambda)$ *such that*

$$M_1(\lambda) \cdot \Phi - M_2(\lambda) \cdot \Phi = \Phi \qquad \forall \Phi \in H_0, \qquad \forall \lambda \in \Omega \backslash C_r, \tag{4.15}$$

and the Green's function $\mathbb{G}_\lambda$ *for* $\lambda \in \Omega \backslash C_r$ *is analytic in* λ *and can be decomposed as follows.*

- *If* $y \geq 0$,

$$\begin{aligned}
\mathbb{G}_\lambda(x,y) \;&=\; T_\lambda(x,y)\, \Pi_+(y,\lambda) \,+\, T_\lambda(x,0)\, M_1(\lambda)\, T_\lambda(0,y)\, \Xi_+(y,\lambda), && x \,>\, y,\\
&=\; -T_\lambda(x,y)\, \Xi_+(y,\lambda) \,+\, T_\lambda(x,0)\, M_1(\lambda) T_\lambda(0,y)\, \Xi_+(y,\lambda), && 0 \leq x \,<\, y,\\
&=\; T_\lambda(x,0)\, M_2(\lambda)\, T_\lambda(0,y) \Xi_+(y,\lambda), && x \leq 0,
\end{aligned}$$

- *and if* $y \leq 0$,

$$\begin{aligned}
\mathbb{G}_\lambda(x,y) \;&=\; -T_\lambda(x,y)\, \Xi_-(y,\lambda) \,+ T_\lambda(x,0)\, M_2(\lambda)\, T_\lambda(0,y)\, \Pi_-(y,\lambda), && x \,<\, y,\\
&=\; T_\lambda(x,y)\, \Pi_-(y,\lambda) \,+\, T_\lambda(x,0)\, M_2(\lambda) T_\lambda(0,y)\, \Pi_-(y,\lambda), && y \,<\, x \leq 0,\\
&=\; T_\lambda(x,0)\, M_1(\lambda)\, T_\lambda(0,y)\, \Pi_-(y,\lambda), && x \,\geq\, 0.
\end{aligned}$$

Proof. Concerning M_1 and M_2 it is sufficient to determine $M_2\, \Xi_+$ and then define

$$M_1 \,=\, \Pi_+ \,+\, (\, I \,+\, M_2\,)\, \Xi_+, \quad M_2 \,=\, M_2\, \Xi_+.$$

But $M_2(\lambda)\, \Xi_+(0,\lambda)$ is fully determined by $M_2(\lambda)\, \mathbb{W}^m_+(0,\lambda)$, $m \in \{1, \ldots, N\}$, which must be looked for in

$$\mathcal{R}\Xi_-(0,\lambda) \,=\, \mathrm{Span}\{\mathbb{W}^1_-(0,\lambda), \ldots, \mathbb{W}^N_-(0,\lambda)\}.$$

Introducing coefficients $\beta_{n,m}(\lambda)$ such that

$$M_2(\lambda)\, \mathbb{W}^m_+(0,\lambda) = \sum_{1 \leq n \leq N} \beta_{n,m}(\lambda)\, \mathbb{W}^n_-(0,\lambda)$$

and requiring that $(I + M_2(\lambda))\,\mathbb{W}_+^m(0,\lambda)$ belong to

$$\mathcal{R}\Pi_+(0,\lambda) = \mathrm{Span}\{\mathbb{Y}_+^1(0,\lambda),\dots,\mathbb{Y}_+^N(0,\lambda)\}^\perp,$$

we easily find that the matrix $\beta(\lambda) = (\beta_{n,m}(\lambda))_{n,m}$ must be equal to $-B(\lambda)^{-1}$, where $B(\lambda)$ is nothing but the matrix defining the Evans function in (3.58). We recall that, thanks to Remark 4.3, $B(\lambda)$ is really invertible in $\Omega\backslash C_r$. Hence, with the same abuse of notation as in (3.56), we obtain

$$M_2(\lambda)=M_1(\lambda)-I=\sum_{1\le n,m\le N}\beta_{n,m}(\lambda)\mathbb{W}_-^n(0,\lambda)\mathbb{Y}_+^m(0,\lambda),\quad \beta(\lambda):=-B(\lambda)^{-1}. \tag{4.16}$$

To find the Green's function $\mathbb{G}_\lambda$, the procedure is rather standard. We look for bounded solutions of the nonhomogeneous problem (4.13) through Duhamel-like formulas. Once it is under the form (4.12), it suffices to read the kernel $\mathbb{G}_\lambda$.

Because of (3.57),

$$\mathbb{W}_+(x,\lambda) := \int_0^x T_\lambda(x,y)\Pi_+(y,\lambda)\mathbb{F}(y)\,dy - \int_x^{+\infty} T_\lambda(x,y)\Xi_+(y,\lambda)\mathbb{F}(y)\,dy \quad \forall x\in\mathbb{R}^+$$

belongs to $L^2(\mathbb{R}^+,H_1)$, is bounded, and is a mild solution of (4.13). Consequently, if $\mathbb{W}$ is a bounded mild solution of (4.13), $\mathbb{W}-\mathbb{W}_+$ is a bounded mild solution of the homogeneous system (3.40). Using again the existence of the exponential dichotomy on $\mathbb{R}^+$ given by (3.57), we see that $(\mathbb{W}-\mathbb{W}_+)(0) =: \Phi_+ \in \mathcal{R}\,\Pi_+(0,\lambda)$; hence

$$\mathbb{W}(x,\lambda) = T_\lambda(x,0)\Phi_+ + \mathbb{W}_+(x,\lambda)$$

for $x\ge 0$. In other words,

$$\begin{aligned}\mathbb{W}(x,\lambda) = T_\lambda(x,0)\cdot\Phi_+ &+ \int_0^x T_\lambda(x,y)\,\Pi_+(y,\lambda)\,\mathbb{F}(y)\,dy\\ &- \int_x^{+\infty} T_\lambda(x,y)\,\Xi_+(y,\lambda)\,\mathbb{F}(y)\,dy \quad \forall x\in\mathbb{R}^+.\end{aligned} \tag{4.17}$$

Similarly, using the exponential dichotomy on $\mathbb{R}_-$, we find that every bounded solution of (4.13) on $\mathbb{R}^-$ is of the form

$$\begin{aligned}\mathbb{W}(x,\lambda) = T_\lambda(x,0)\Phi_- &+ \int_0^x T_\lambda(x,y)\,\Xi_-(y,\lambda)\,\mathbb{F}(y)\,dy\\ &+ \int_{-\infty}^x T_\lambda(x,y)\,\Pi_-(y,\lambda)\,\mathbb{F}(y)dy \quad \forall x\in\mathbb{R}^-,\end{aligned} \tag{4.18}$$

where $\Phi_-\in\mathcal{R}\Xi_-(0,\lambda)$.

For a bounded solution on $\mathbb{R}$, the two expressions (4.17) and (4.18) must agree at $x=0$. Hence we get

$$\Phi_+ - \Phi_- = \int_0^{+\infty} T_\lambda(0,y)\,\Xi_+(y,\lambda)\,\mathbb{F}(y)\,dy + \int_{-\infty}^0 T_\lambda(0,y)\,\Pi_-(y,\lambda)\,\mathbb{F}(y)\,dy.$$

By definition of $M_i(\lambda)$, this yields

$$\Phi_+ = M_1(\lambda)\left(\int_0^{+\infty} T_\lambda(0,y)\Xi_+(y,\lambda)\mathbb{F}(y)\,dy + \int_{-\infty}^0 T_\lambda(0,y)\Pi_-(y,\lambda)\mathbb{F}(y)\,dy\right)$$

and

$$\Phi_- = M_2(\lambda)\left(\int_0^{+\infty} T_\lambda(0,y)\Xi_+(y,\lambda)\mathbb{F}(y)\,dy + \int_{-\infty}^{0} T_\lambda(0,y)\Pi_-(y,\lambda)\mathbb{F}(y)\,dy\right).$$

Substituting these two expressions into (4.17) and (4.18) and interpreting the result in the various cases, we complete the proof of the decomposition of $\mathbb{G}_\lambda$. ∎

Remark 4.5. We have not used the well-posedness of the Cauchy problem for the operator $\mathbb{L}_\lambda$. The crucial point is the existence of exponential dichotomies on $\mathbb{R}^\pm$. So this result could extend to more general shemes satisfying (H1)–(H3). Indeed, for λ in $\Omega\backslash C_r$, $\mathbb{L}_\lambda$ is asymptotically hyperbolic, and the result of [10] ensures the existence of exponential dichotomies on both half-lines.

Estimating each term of $\mathbb{G}_\lambda$ separately, by means of (3.57), we easily show the following.

COROLLARY 4.6. *For all $R > 0$, there exists $\alpha > 0$ such that for every $\lambda \in \Omega\backslash C_r$, $\operatorname{Re}\lambda \le R$, we have the estimate*

$$|\mathbb{G}_\lambda(x,y)| \le C(r,R)e^{-\alpha|x-y|} \quad \forall x,y \in \mathbb{R}, \tag{4.19}$$

and thus, thanks to (4.11),

$$|G_\lambda(x,y)| \le C(r,R)e^{-\alpha|x-y|} \quad \forall x,y \in \mathbb{R}. \tag{4.20}$$

We have adopted here the simple notation $|\cdot|$ for both the norm in $\mathcal{L}(\mathbb{C}^N)$ and in $\mathcal{L}(H_0)$.

The next step is more delicate; it consists of showing that the Green's function $\mathbb{G}_\lambda$ has an analytic continuation to a whole vicinity of the origin, namely C_r. The precise description of this continuation is the crucial part of the analysis, since bounds on G_λ for small λ will determine the large time behavior of $G_j^{j_0}(t,t_0)$.

THEOREM 4.7. *Assuming* (H5)–(H6), $\mathbb{G}_\lambda$ *extends to C_r as follows. If $y \ge 0$,*

$$\begin{aligned}\mathbb{G}_\lambda(x,y) = {} & T_\lambda(x,y)P_+(y,\lambda) + \sum_{N+k+1\le l\le 2N} \mathbb{W}_+^l(x,\lambda)\mathbb{Y}_+^l(y,\lambda) \\ & + \sum_{1\le n,\,m\le N} \beta_{n,m}(\lambda)T_\lambda(x,0)P_+(0,\lambda)\mathbb{W}_-^n(0,\lambda)\mathbb{Y}_+^m(y,\lambda) \\ & + \sum_{1\le m\le N}\ \sum_{N+k+1\le l\le 2N} \gamma_{l,m}(\lambda)\mathbb{W}_+^l(x,\lambda)\mathbb{Y}_+^m(y,\lambda), \qquad x > y,\end{aligned} \tag{4.21}$$

$$\begin{aligned}\mathbb{G}_\lambda(x,y) = {} & -\sum_{1\le m\le N} \mathbb{W}_+^m(x,\lambda)\mathbb{Y}_+^m(y,\lambda) \\ & + \sum_{1\le n\,,\,m\le N} \beta_{n,m}(\lambda)T_\lambda(x,0)P_+(0,\lambda)\mathbb{W}_-^n(0,\lambda)\mathbb{Y}_+^m(y,\lambda) \\ & + \sum_{1\le m\le N}\ \sum_{N+k+1\le l\le 2N} \gamma_{l,m}(\lambda)\mathbb{W}_+^l(x,\lambda)\mathbb{Y}_+^m(y,\lambda), \qquad 0 \le x < y,\end{aligned} \tag{4.22}$$

$$\mathbb{G}_\lambda(x,y) = \sum_{1\le n\,,\,m\le N} \beta_{n,m}(\lambda)\mathbb{W}_-^n(x,\lambda)\mathbb{Y}_+^m(y,\lambda), \qquad x \le 0, \tag{4.23}$$

where the coefficients $\beta_{n,m}(\lambda)$ are defined as in (4.16) *by*

$$(\beta_{n,m}(\lambda)) = \beta(\lambda) = -B(\lambda)^{-1}, \tag{4.24}$$

$$B(\lambda)_{n,m} = \langle\, \mathbb{Y}^n_+(0,\lambda)\,,\, \mathbb{W}^m_-(0,\lambda)\,\rangle\,, \quad 1 \le n\,,\, m \le N,$$

and the coefficients $\gamma_{l,m}(\lambda)$ by

$$(\gamma_{l,m}(\lambda)) = \gamma(\lambda) = C(\lambda)\,\beta(\lambda), \tag{4.25}$$

$$C(\lambda)_{l,m} = \langle\, \mathbb{Y}^l_+(0,\lambda)\,,\, \mathbb{W}^m_-(0,\lambda)\,\rangle, \quad N+k+1 \le l \le 2N,\ 1 \le m \le N.$$

In C_r, the coefficients $\beta_{n,m}(\lambda)$ are holomorphic for $n \neq k$ and meromorphic with a single pole of order at most one at zero if $n = k$. The coefficients $\gamma_{l,m}(\lambda)$ are holomorphic.

For $y \le 0$ we have left similar expressions to the reader.

Proof. The expressions obtained in Proposition 4.4 can be extended to C_r by means of Lemma 3.13. Recalling the definition of P_+,

$$P_+(y,\lambda) = I - \sum_{m=1}^{N} \mathbb{W}^m_+(y,\lambda)\,\mathbb{Y}^m_+(y,\lambda) - \sum_{l=N+k+1}^{2N} \mathbb{W}^l_+(y,\lambda)\,\mathbb{Y}^l_+(y,\lambda), \tag{4.26}$$

and the definition of $\Pi_+(y,\lambda)$ (see (3.56)), we can write

$$\Pi_+(y,\lambda) = \sum_{N+k+1\le l\le 2N} \mathbb{W}^l_+(y,\lambda)\mathbb{Y}^l_+(y,\lambda) + P_+(y,\lambda). \tag{4.27}$$

Consequently, we easily get that

$$T_\lambda(x,y)\,\Pi_+(y,\lambda) = T_\lambda(x,y)\,P_+(y,\lambda) + \sum_{N+k+1\le l\le 2N} \mathbb{W}^l_+(x,\lambda)\mathbb{Y}^l_+(y,\lambda). \tag{4.28}$$

This is the first piece of $\mathbb{G}_\lambda$ for $x > y$ given in Proposition 4.4. We also need an expansion of $T_\lambda(x,0)\,M_1(\lambda)\,T_\lambda(0,y)\,\Xi_+(y,\lambda)$. By definition of $\Xi_+(y,\lambda)$ (see (3.56)), we have

$$T_\lambda(0,y)\,\Xi_+(y,\lambda) = \sum_{1\le m\le N} \mathbb{W}^m_+(0,\lambda)\mathbb{Y}^m_+(y,\lambda).$$

Hence, using (4.16), we find that

$$T_\lambda(x,0)\,M_1(\lambda)\,T_\lambda(0,y)\,\Xi_+(y,\lambda) = \sum_{1\le m\le N} \mathbb{W}^m_+(x,\lambda)\mathbb{Y}^m_+(y,\lambda) + T_\lambda(x,0)\,M_2(\lambda)\,T_\lambda(0,y)\,\Xi_+(y,\lambda), \tag{4.29}$$

$$T_\lambda(x,0)M_2(\lambda)T_\lambda(0,y)\,\Xi_+(y,\lambda) = \sum_{1\le n,m\le N} \beta_{n,m}(\lambda)\,T_\lambda(x,0)\mathbb{W}^n_-(0,\lambda)\mathbb{Y}^m_+(y,\lambda). \tag{4.30}$$

This formula is not sufficient to get estimates of $\mathbb{G}_\lambda$, because of the terms $T_\lambda(x,0)\mathbb{W}^n_-(0,\lambda) = \mathbb{W}^n_-(x,\lambda)$, for which Lemma 3.13 gives the behavior for $x \le 0$

only. We are going to rewrite (4.30) in a more suitable way. By definition of $P_+(0,\lambda)$ (see (4.26)), we have

$$\mathbb{W}^n_-(0,\lambda) = \sum_{l\in M_+} \langle\, \mathbb{Y}^l_+(0,\lambda)\,,\, \mathbb{W}^n_-(0,\lambda)\,\rangle\, \mathbb{W}^l_+(0,\lambda) \,+\, P_+(0,\lambda)\, \mathbb{W}^n_-(0,\lambda),$$

where $M_+ = \{1,\dots,N\} \cup \{N+k+1,\dots,2N\}$ as before. Therefore, the right-hand side in (4.30) equivalently reads

$$\sum_{1\le n\,,\, m\le N} \beta_{n,m}(\lambda)\, T_\lambda(x,0)\, \mathbb{W}^n_-(0,\lambda)\, \mathbb{Y}^m_+(y,\lambda)$$

$$= \sum_{1\le n\,,\, m\le N} \beta_{n,m}(\lambda)\, T_\lambda(x,0)\, P_+(0,\lambda)\, \mathbb{W}^n_-(0,\lambda)\, \mathbb{Y}^m_+(y,\lambda) \,+\, \Sigma_1 \,+\, \Sigma_2,$$

where

$$\Sigma_1 \;=\; \sum_{1\le n\,,\, m\,,\, l\le N} B_{l,n}(\lambda)\, \beta_{n,m}(\lambda)\, T_\lambda(x,0)\, \mathbb{W}^n_+(0,\lambda)\, \mathbb{Y}^m_+(y,\lambda)$$

$$= \,-\, \sum_{1\le m\le N} T_\lambda(x,0)\, \mathbb{W}^m_+(0,\lambda)\, \mathbb{Y}^m_+(y,\lambda)$$

because of (4.24), and

$$\Sigma_2 \;=\; \sum_{1\le n\,,\, m\le N} \sum_{l=N+k+1}^{2N} C_{l,n}(\lambda)\, \beta_{n,m}(\lambda)\, T_\lambda(x,0)\, \mathbb{W}^n_+(0,\lambda)\, \mathbb{Y}^m_+(y,\lambda)$$

$$= \sum_{1\le m\le N} \sum_{l=N+k+1}^{2N} \gamma_{l,m}(\lambda)\, T_\lambda(x,0)\, \mathbb{W}^l_+(0,\lambda)\, \mathbb{Y}^m_+(y,\lambda)$$

in view of the definition (4.25) of $\gamma(\lambda)$. This implies in particular that the first sum in the right-hand side of (4.29) cancels out. Eventually, we get

$$T_\lambda(x,0)\, M_1(\lambda)\, T_\lambda(0,y)\, \Xi_+(y,\lambda) \;=\; \sum_{1\le n,m\le N} \beta_{n,m} T_\lambda(x,0)\, P_+(0,\lambda)\, \mathbb{W}^n_-(0,\lambda)$$

$$+ \sum_{1\le m\le N} \sum_{l=N+k+1}^{2N} \gamma_{l,m}(\lambda) \mathbb{W}^l_+(x,\lambda)\, \mathbb{Y}^m_+(y,\lambda). \tag{4.31}$$

Collecting (4.28), (4.30), and (4.31), we get (4.21), (4.22), and (4.23) from the corresponding expressions in Proposition 4.4.

It remains to prove the statements about the analyticity of the coefficients. Regarding $\beta_{n,m}(\lambda)$, we have by (4.24) and Cramer's rule

$$\beta_{n,m}(\lambda) \;=\; -\,\frac{\det(B_1(\lambda),\dots,B_{n-1}(\lambda),e_m,B_{n+1}(\lambda),\dots,B_N(\lambda))}{\det B(\lambda)}, \tag{4.32}$$

where, for all $m\in\{1,\dots,N\}$, B_m denotes the mth *column* of B and e_m denotes the mth vector of the canonical basis in $\mathbb{C}^N$. As already noted above, we may assume

that one of the $\mathbb{W}^m$ coincides with $\mathcal{I}U' = \mathbb{U}'$ at $\lambda = 0$. This is a usual simplification in the study of the Evans function in the vicinity of $\lambda = 0$. For reasons that were explained in [2], we assume that the concerned index is k, i.e.,

$$\mathbb{W}^k_-(x,0) = \mathbb{U}'(x). \tag{4.33}$$

Consequently, $\mathbb{W}^k_-(x,0)$ is also decreasing when x tends to $+\infty$, and since $\mathbb{Y}^+_n(x,0)$ is bounded when x tends to $+\infty$ for $n \in M_+$, we deduce that

$$\langle\, \mathbb{Y}^n_+(x,0)\,,\, \mathbb{W}^k_-(x,0)\rangle \equiv 0 \quad \forall n \in M_+. \tag{4.34}$$

This means that $B_k(0) = 0$. Consequently, (4.32) and our assumption (H6) imply that $\beta_{n,m}$ is analytic in C_r, provided that $n \neq k$ and that β_{km} has a simple pole at $\lambda = 0$. Furthermore, the pole of β_{km} is compensated by the cancellation of $C_{lk}(\lambda) = \langle\, \mathbb{Y}^l_+(x,\lambda)\,,\, \mathbb{W}^k_-(x,\lambda)\,\rangle$ at $\lambda = 0$. Therefore, all terms in the sum defining $\gamma_{l,m}$ (see (4.25)) are analytic in C_r. ∎

At this stage, we did not really use the nature of the shock, that is, the dimension of Ker $P_\pm$. If we were to use the expansions of Theorem 4.7 to estimate the Green's function $G^{j_0}_j(t,t_0)$ of the evolution problem, we should get estimates similar to the ones obtained in [24] for viscous shocks, which have been pointed out in [23] as *not* being optimal for Lax shocks. This is why we now turn to refined expansions of the Green's function, inspired from [18] and [23].

4.3. Refined decomposition of the Green's function. In this section, we show the final refined decomposition of the Green's function. We define the "truncated" eigenvalues $\nu^m_\pm$ by

$$\nu^m_\pm(\lambda) = -\frac{\lambda}{\lambda^m_\pm} + \frac{a^m_\pm}{2(\lambda^m_\pm)^3}\lambda^2, \tag{4.35}$$

where $\lambda^m_\pm := a^m_\pm - \sigma$, in such a way that (3.48) merely reads

$$\mu^m_\pm(\lambda) = \nu^m_\pm(\lambda) + \mathcal{O}(\lambda^3). \tag{4.36}$$

THEOREM 4.8. *Assuming* (H4)–(H6), *we have for some constants independent of* λ, c^+_m, $\widetilde{\beta}^+_{n,m}$, $\widetilde{\gamma}^+_{l,m}$ *the following expansion of the Green's function* G_λ *in* C_r*:*

$$G_\lambda \,=\, E_\lambda \,+\, S_\lambda \,+\, R_\lambda,$$

where for $y > 0$,

$$E_\lambda(x,y) \,=\, \sum_{m=1}^{k} \frac{c^+_m}{\lambda} \mathrm{e}^{-\nu^m_+(\lambda)y} U'(x)\, \ell^m_+, \tag{4.37}$$

$$S_\lambda(x,y) = \sum_{l=N+k+1}^{2N} \mathrm{e}^{\nu^l_+(x-y)}\, r^l_+\, \ell^l_+ \,+\, + \sum_{m=1}^{k} \sum_{l=N+k+1}^{2N} \widetilde{\gamma}^+_{l,m}\, \mathrm{e}^{\nu^l_+ x}\, \mathrm{e}^{-\nu^m_+ y}\, r^l_+\, \ell^m_+ \tag{4.38}$$

if $x > y$,

$$S_\lambda(x,y) = -\sum_{m=1}^{k} \mathrm{e}^{\nu^m_+(x-y)}\, r^m_+\, \ell^m_+ \,+\, \sum_{m=1}^{k} \sum_{l=N+k+1}^{2N} \widetilde{\gamma}^+_{l,m}\, \mathrm{e}^{\nu^l_+ x}\, \mathrm{e}^{-\nu^m_+ y}\, r^l_+\, \ell^m_+ \tag{4.39}$$

if $0 < x < y$,

$$S_\lambda(x,y) = \sum_{n=1}^{k-1} \sum_{m=1}^{k} \widetilde{\beta}_{n,m}\, \mathrm{e}^{\nu^n_- x}\, \mathrm{e}^{-\nu^m_+ y}\, r^n_-\, \ell^m_+ \tag{4.40}$$

if $x < 0$,

and

$$R_\lambda = R_\lambda^E + R_\lambda^S,$$

where

$$\partial_y^\alpha R_\lambda^E = \mathcal{O}(\mathrm{e}^{-\omega|x-y|}) \tag{4.41}$$

$$+ \sum_{m=1}^{k} \mathrm{e}^{-\nu_+^m y}\mathcal{O}\Big(\lambda^{\alpha-1}(\mathrm{e}^{\mathcal{O}(\lambda^3)y}-1) + \lambda^{\alpha-1}(\mathrm{e}^{\mathcal{O}(\lambda^3)x}-1) + \mathcal{O}(\lambda^\alpha)\Big)\mathrm{e}^{-\omega|x|},$$

$$\partial_y^\alpha R_\lambda^S = \mathcal{O}(\mathrm{e}^{-\omega|x-y|}) \tag{4.42}$$

$$+ \sum_{N+k+1\le l\le 2N} \mathrm{e}^{\nu_+^l(x-y)}\mathcal{O}\Big(\lambda^\alpha \mathrm{e}^{-\omega x} + \lambda^\alpha(\mathrm{e}^{\mathcal{O}(\lambda^3)x}-1) + \lambda^\alpha(\mathrm{e}^{\mathcal{O}(\lambda^3)y}-1) + \lambda^{\alpha+1}\Big)$$
$$+ \sum_{m\le k, N+k+1\le l\le 2N} \mathcal{O}\mathrm{e}^{\nu_+^l x - \nu_+^m y}\Big(\lambda^\alpha(\mathrm{e}^{\mathcal{O}(\lambda^3)x}-1) + \lambda^\alpha(\mathrm{e}^{\mathcal{O}(\lambda^3)y}-1) + \lambda^\alpha \mathrm{e}^{-\omega|x|} + \lambda^{\alpha+1}\Big)$$
$$\text{if } x > y,$$

$$\partial_y^\alpha R_\lambda^S = \mathcal{O}(\mathrm{e}^{-\omega|x-y|}) \tag{4.43}$$

$$+ \sum_{m\le k} \mathrm{e}^{\nu_+^m(x-y)}\mathcal{O}\Big(\lambda^\alpha \mathrm{e}^{-\omega x} + \lambda^\alpha(\mathrm{e}^{\mathcal{O}(\lambda^3)x}-1) + \lambda^\alpha(\mathrm{e}^{\mathcal{O}(\lambda^3)y}-1) + \lambda^{\alpha+1}\Big)$$
$$+ \sum_{m\le k, N+k+1\le l\le 2N} \mathrm{e}^{\nu_+^l x - \nu_+^m y}\mathcal{O}\Big(\lambda^\alpha \mathrm{e}^{-\omega x} + \lambda^\alpha(\mathrm{e}^{\mathcal{O}(\lambda^3)x}-1) + \lambda^\alpha(\mathrm{e}^{\mathcal{O}(\lambda^3)y}-1) + \lambda^{\alpha+1}\Big)$$
$$\text{if } 0 < x < y,$$

$$\partial_y^\alpha R_\lambda^S \tag{4.44}$$

$$= \sum_{n\le k, m\le k} \mathrm{e}^{\nu_-^n x - \nu_+^m y}\mathcal{O}\Big(\lambda^\alpha \mathrm{e}^{\omega x} + \lambda^\alpha(\mathrm{e}^{\mathcal{O}(\lambda^3)x}-1) + \lambda^\alpha(\mathrm{e}^{-\mathcal{O}(\lambda^3)y}-1) + \lambda^\alpha \mathrm{e}^{-\omega|x|} + \lambda^{\alpha+1}\Big)$$
$$\text{if } x < 0.$$

All these estimates are uniform in λ, x, y. Symmetric expansions hold for $y \le 0$.

Similar expansions were found in [23, 18], thanks to a careful study of the slow modes of the adjoint system, that is to say, the $\mathbb{Y}_+^m$, $m \in \{1,\dots,k\} \cup \{N+k+1,\dots,2N\}$ (associated with the $\mu_+^m(\lambda)$ tending to zero when λ tends to zero), in order to get additional cancellations in the coefficients $\beta_{n,m}(\lambda)$, $\gamma_{l,m}(\lambda)$. These results were based on the properties of the standard (see [4]) Evans function for viscous and relaxation shock profiles, respectively. Here the proof will follow the same outlines, adapted to our mixed Evans function (see again [4] and [2]) in infinite dimensions. The generic assumption (H4) is needed to do precise computations at $\lambda = 0$, similarly as in [2, 4].

We make the choice (4.33). Furthermore, the constants are obviously in the kernel of the advanced differential operator

$$L^* : Y \mapsto L^* Y ; \; L^* Y(x) = -\sigma - \sigma Y'(x) + (Y(x+1) - Y(x)) A(x).$$

So, similarly as in [18, 2], we may choose the slow modes of the adjoint dynamical system in such a way that

$$\mathbb{Y}_+^n(x,0) = \mathcal{J} \ell_+^n = \left(-\ell_+^n A_{x-1}, \sigma \ell_+^n\right), \tag{4.45}$$

$$n \in \{1, \dots, k\} \cup \{N+k+1, \dots, 2N\}.$$

In particular, these slow modes $\mathbb{Y}_+^n(x,0)$ are well defined and bounded on the whole real line.

Thanks to (4.33) and (4.45), we can get more information on the coefficients $\beta_{n,m}(\lambda)$ and $\gamma_{l,m}(\lambda)$ at $\lambda = 0$.

LEMMA 4.9. *Assuming* (H4)–(H6), *we have the following additional properties:*

- *for all* $m \in \{k+1, \dots, N\}$, $\beta_{k,m}$ *is analytic in* C_r;
- *for all* $m \in \{k+1, \dots, N\}$ *and* $l \in \{N+K+1, \dots, 2N\}$, $\gamma_{l,m}(\lambda) = \mathcal{O}(\lambda)$ *in* C_r.

Proof of Lemma 4.9. We take benefit from (H4) in the same way as in [4, 2]. There exist independent row vectors $h^1, \dots, h^k$ such that

$$h^n \in \text{Span}\,\{\ell_+^1, \dots, \ell_+^k\}, \quad n \in \{1, \dots, k\}, \tag{4.46}$$

$$h^k \in \text{Span}\{\ell_-^k, \dots, \ell_-^N\}. \tag{4.47}$$

The isomorphism M, defined on $(\mathbb{R}^N)^*$ by

$$M \ell_+^n = h^n, \; n \le k, \quad M \ell_+^n = \ell_+^n, \quad n \ge k+1, \tag{4.48}$$

induces an isomorphism on H_0^*, defined by

$$\mathbb{M} : \Psi = (\psi, b) \mapsto (M\psi, Mb),$$

which leaves unchanged $\mathbb{Y}_+^n(x,0)$ for $n \ge k+1$. For simplicity, we denote

$$\mathbb{Y}^n(x,\lambda) := \mathbb{M}\, \mathbb{Y}_+^n(x,\lambda)$$

for all $n \in M_+$. We thus have

$$D(\lambda) = \det(\langle\, \mathbb{Y}_+^n(0,\lambda), \mathbb{W}_-^m(0,\lambda)\,\rangle) = \det M^{-1} \det(\langle\, \mathbb{Y}^n(0,\lambda), \mathbb{W}_-^m(0,\lambda)\,\rangle),$$

and, in view of (3.54) and (4.32),

$$\beta_{k,m}(\lambda) = \frac{\det \check{B}^m(\lambda)}{\det \widetilde{B}(\lambda)},$$

where

$$\begin{aligned}
&\widetilde{B}_{n,l}(\lambda) = \langle\, \mathbb{Y}^n(0,\lambda), \mathbb{W}_-^l(0,\lambda)\,\rangle \quad \forall n, l \in \{1, \dots, N\},\\
&\check{B}_{n,l}^m(\lambda) = \widetilde{B}_{n,l}(\lambda), \quad l \ne k,\\
&\check{B}_{n,k}^m(\lambda) = \langle\, \mathbb{Y}^n(0,\lambda), \mathbb{W}_+^m(0,\lambda)\,\rangle \quad \forall n \in \{1, \dots, N\}.
\end{aligned}$$

Using (3.52), (4.45), and (4.48), for all $n \leq k-1$ we have

$$\langle\, \mathbb{Y}^k(0,0), \mathbb{W}^n_-(0,0)\,\rangle = \lim_{x\to-\infty} \langle\, \mathbb{Y}^k(x,0), \mathbb{W}^n_-(x,0)\,\rangle = \left\langle\, (h^k A_-, h^k), \begin{pmatrix} r^n_- \\ r^n_- \end{pmatrix} \right\rangle = 0$$

because of (4.47). Furthermore, thanks to (4.45) and (4.46),

$$\mathbb{Y}^k(x,0) \in \text{Span}\,\{\mathbb{Y}^1_+(x,0), \dots, \mathbb{Y}^k_+(x,0)\}$$

is bounded on the whole real line, and for $k \leq n \leq N$, $\mathbb{W}^n_-(-\infty,0) = 0$. So we also have

$$\langle\, \mathbb{Y}^k(0,0), \mathbb{W}^n_-(0,0)\,\rangle = 0.$$

This means that the kth row of $\widetilde{B}(\lambda)$ vanishes at $\lambda = 0$. This shows once more that D vanishes at $\lambda = 0$. And (H6) says that $\det\widetilde{B}(\lambda)^{-1}$ has a simple pole at that point. But the kth row of $\check{B}^m(\lambda)$ also vanishes at $\lambda = 0$ for $m \geq k+1$, since $\mathbb{Y}^k(x,0)$ is bounded on the whole real line and $\mathbb{W}^m_+(+\infty,\lambda) = 0$ for $m \geq k+1$. So the pole due to the denominator is compensated by the cancellation of the numerator $\det\check{B}^m(\lambda)$ at $\lambda = 0$. Therefore, $\beta_{k,m}$ is analytic on C_r for $m \geq k+1$.

We now turn to the study of the coefficients $\gamma_{l,m}(\lambda)$ for $m \geq k+1$. We recall from (4.25) that

$$\gamma_{l,m} = \sum_{n=1}^{N} \beta_{n,m}(\lambda)\, \langle\, \mathbb{Y}^l_+(x,\lambda)\,,\, \mathbb{W}^n_-(x,\lambda)\,\rangle.$$

- By Theorem 4.7 and the previous argument, $\beta_{n,m}(\lambda)$ is analytic on C_r for all $n \in \{k, \cdot, N\}$ and $m \in \{k+1, \dots, N\}$. And $\langle\, \mathbb{Y}^l_+(0,0), \mathbb{W}^n_-(0,0)\,\rangle = 0$ for $N+k+1 \leq l \leq 2N$ since $\mathbb{Y}^l_+(x,0)$ is bounded and $\mathbb{W}^n_-(-\infty,0) = 0$. Consequently, we have

$$\beta_{n,m}(\lambda)\langle\, \mathbb{Y}^l_+(0,\lambda), \mathbb{W}^n_-(0,\lambda)\,\rangle = \mathcal{O}(\lambda), \quad n \in \{k, \dots, N\}. \tag{4.49}$$

- For $n \leq k-1$, we have to study more carefully the coefficients $\beta_{n,m}(\lambda)$ to get additional cancellation. Denoting

$$E(\lambda) = \det(B_1(\lambda), \dots, B_{n-1}(\lambda), e_m, B_{n+1}(\lambda), \dots, B_N(\lambda),$$

we have $E(0) = 0$ because $B_k(0) = 0$ (as already noted in the proof of Theorem 4.7). Furthermore,

$$E'(0) = \det(B_1(0), \dots, B_{n-1}(0), e_m, B_{n+1}(0), \dots, B'_k(0), \dots, B_N(0)).$$

And

$$\langle\, \mathbb{Y}^p_+(0,0), \mathbb{W}^l_-(0,0)\,\rangle = 0, \qquad p \leq k, \quad k \leq l \leq N,$$

since $\mathbb{Y}^p_+(x,0)$ is bounded and $\mathbb{W}^l_-(-\infty,0) = 0$. This means that the first k components of $B_l(0)$ are null for $l \geq k$. This is also the case for the vector e_m ($m \geq k+1$). Hence the $N \times N$ determinant involved in $E'(0)$ has at least a $k \times (N-k+1)$ null block. So this determinant must be equal to 0, i.e., $E'(0) = 0$. By (4.32) and (H6), this shows that

$$\beta_{n,m}(\lambda) = \mathcal{O}(\lambda), \quad n \in \{1, \cdot, k-1\}, \in \{k+1, \dots, N\}. \tag{4.50}$$

Collecting (4.49) and (4.50), we complete the proof of Lemma 4.9. ∎

For notational convenience, we set

$$L_+^n(x) = \mathbb{Y}_+^n(x,0) \,=\, \mathcal{J}\,\ell_+^n \quad \text{and} \quad R_+^m \,=\, \mathcal{I}\,r_+^m \,=\, \begin{pmatrix} r_+^m \\ r_+^m \end{pmatrix} \,=\, \Phi_+^m(0).$$

We observe that

$$L_+^n(x)\,\Upsilon \,=\, \ell_+^n, \tag{4.51}$$

where we have made again an abuse of notation; i.e., we mean that

$$\langle\, L_+^n(x)\,,\, \Upsilon\cdot c\,\rangle \,=\, \ell_+^n\cdot c \quad \forall c\in\mathbb{C}^{\mathbb{N}},$$

and $\Gamma\cdot R_+^m \,=\, r_+^m$. Hence, because of (4.11), the undifferentiated expansions of G_λ (i.e., with $\alpha=0$) of Theorem 4.8 follow from similar expansions of $\mathbb{G}_\lambda$, with r_+^n and ℓ_+^n replaced by R_+^n and L_+^n. The improved differentiated bounds for the residual term $\partial_y^\alpha R_\lambda$ will also follow from the expansion of $\mathbb{G}_\lambda$.

The next step in the proof of Theorem 4.8 is to describe as precisely as in Lemma 5.6 of [18] the slow modes $\mathbb{Y}^n$ for the adjoint dynamical system.

LEMMA 4.10. *We have*

$$\overline{\mathbb{Y}}_+^n(y,\lambda) = \mathrm{e}^{-\mu_+^n(\lambda)y}L_+^n(y) + \lambda\,\Theta_+^n(y,\lambda), \quad n\in\{1,\dots,k\}\cup\{N+k+1,\dots,2N\}, \tag{4.52}$$

where Θ_+^n *verifies*

$$|\Theta_+^n(y,\lambda)| \,\le\, C\,|\mathrm{e}^{-\mu_+^n(\lambda)y}| \quad \forall y\ge 0, \tag{4.53}$$

$$|\partial_y\Theta_+^n(y,\lambda)| \,\le C\,|\mathrm{e}^{-\mu_+^n(\lambda)y}|\Big(\,|\lambda| \,+\, \mathrm{e}^{-\omega y}\Big) \quad \forall y\ge 0. \tag{4.54}$$

Proof of Lemma 4.10. The same analysis as in [18, Lemma 5.6] can be done, since the gap lemma is still valid in our framework (see [2]). The idea is to apply the gap lemma in the augmented unknown

$$\widetilde{\mathbb{Y}}^n = \begin{pmatrix} \overline{\mathbb{Y}^n} \\ \partial_y\overline{\mathbb{Y}^n} \end{pmatrix},$$

where we omit writing the index $\pm$ for simplicity. As usual, we set

$$\widetilde{\mathbb{Y}}^n = e^{-\mu^n(\lambda)y}\widetilde{\mathbb{V}}^n(y,\lambda),\ \widetilde{\mathbb{V}}^n(y,\lambda) = \begin{pmatrix} \mathbb{V}^n \\ -\mu^n(\lambda)\mathbb{V}^n + \partial_y\mathbb{V}^n \end{pmatrix}.$$

Using the gap lemma, we get the estimate

$$\partial_\lambda\widetilde{\mathbb{V}}^n(y,\lambda) = \partial_\lambda\widetilde{\Psi}^n(\lambda) + \mathcal{O}(e^{-\alpha|y|}) \tag{4.55}$$

for some $\alpha>0$, where $\widetilde{\Psi}^n(\lambda) = \left(\begin{smallmatrix} \Psi^n(\lambda) \\ \mu^n(\lambda)\Psi^n(\lambda) \end{smallmatrix}\right)$. Using the Taylor formula, we get

$$\widetilde{\mathbb{Y}}^n(y,\lambda) = e^{-\mu^n(\lambda)y}\left(\widetilde{\mathbb{V}}^n(y,0) + \lambda\frac{\partial}{\partial\lambda}\widetilde{\mathbb{V}}^n(y,0) + \frac{\lambda^2}{2}\int_0^1 \partial_\lambda^2\widetilde{\mathbb{V}}(y,s\lambda)\,ds\right),$$

and it suffices to take the first component of this equation to get the first estimate in (4.53). To get the second one, we use the second component. On the one hand, this yields

$$\partial_y \overline{\mathbb{Y}^n}(y,\lambda) = e^{-\mu^n(\lambda)y}\Big(-\mu^n(0)\mathbb{V}(y,0) + \partial_\lambda \mathbb{V}^n(y,0) + \lambda\partial_\lambda(-\mu^n(0)\mathbb{V}^n(y,0) + \partial_y V^n(y,0)) + \mathcal{O}(\lambda^2)\Big). \tag{4.56}$$

On the other hand, we have

$$\begin{aligned}\partial_y(e^{-\mu^n(\lambda)y}L^n(y)) &= e^{-\mu^n(\lambda)y}\Big(-\mu^n(\lambda)L^n(y) + \partial_y L^n(y)\Big)\\ &= e^{-\mu^n(\lambda)y}\Big(-\mu^n(0)L^n(y) - \lambda\partial_\lambda\mu^n(0)L^n(y) + \partial_y L^n(y) + \mathcal{O}(\lambda^2)\Big).\end{aligned} \tag{4.57}$$

Finally, since $\mu^n(0) = 0$, $L^n(y) = \mathbb{V}^n(y,0)$, and $|\partial_{\lambda y}\mathbb{V}^n(y,0)| = \mathcal{O})(e^{-\alpha|y|})$, we can substract (4.56) and (4.57) to get

$$\partial_y \Theta^n = \mathcal{O}(e^{-\mu^n(\lambda)y}(\lambda + e^{-\omega y})). \qquad \square$$

Because of (4.51), equation (4.52) implies that

$$\overline{\mathbb{Y}}^n_+(y,\lambda)\,\Upsilon \;=\; \mathrm{e}^{-\mu^n_+(\lambda)y}\,\ell^n_+ \;+\; \lambda\Theta^n_+(y,\lambda)\Upsilon. \tag{4.58}$$

As pointed out in [18], the main interest of this lemma is that, since ℓ^n_+ is a constant and $\mu^n_+(\lambda) = \mathcal{O}(\lambda)$ for $n \in \{1,\dots,k\} \cup \{N+k+1,\dots,2N\}$, we get from (4.58) the estimate

$$|\,\partial_y(\,\overline{\mathbb{Y}^n}_+(y,\lambda)\,\Upsilon\,)\,| \;=\; \mathcal{O}(\lambda\,\mathrm{e}^{-\mu^n_+(\lambda)y}), \quad n \in \{1,\dots,k\} \cup \{N+k+1,\dots,2N\}, \tag{4.59}$$

which is better than the basic estimate given by (3.53),

$$|\,\partial_y(\,\overline{\mathbb{Y}^n_+}(y,\lambda)\,\Upsilon\,)\,| \;=\; \mathcal{O}(\mathrm{e}^{-\mu^n_+(\lambda)y}).$$

As it is well explained in [23, 18], this is a crucial part in the analysis. It yields additional cancellations in the expansion of $\partial_y G_\lambda$ at $\lambda = 0$, which in turn will improve the time asymptotic behavior of the "discrete derivative" $G^{j_0}_j(t,t_0) - G^{j_0-1}_j(t,t_0)$, in a way that is necessary to close the fixed point argument in the nonlinear stability analysis. Note that the improved estimate (4.59) is *not* true for $\partial_y \mathbb{Y}^+_n$ itself, since the first component of $L^n_+(y)$ does depend on y.

Proof of Theorem 4.8. We use the expansions given by Theorem 4.7 and Lemmas 4.9 and 4.10. This is a tedious but not difficult job. For example, let us consider the case $x > y$ and look for a refined expansion of (4.21).

- Thanks to (3.55), $T_\lambda(x,y)P_+(y,\lambda)$ is fastly decreasing; hence it is a part of the residual term R_λ.
- Next, we have to study $\mathbb{W}^l_+(x,\lambda)\mathbb{Y}^l_+(y,\lambda)$ for $N+k+1 \le l \le 2N$. Using (3.52) and (4.52) from Lemma 4.10, we have

$$\mathbb{W}^l_+(x,\lambda)\,\mathbb{Y}^l_+(y,\lambda) \;=\mathrm{e}^{\mu^l_+ x}\,\Big(\,R^l_+ \;+\; \mathcal{O}(\mathrm{e}^{-\omega x}\,)\Big)\Big(\,\mathrm{e}^{-\mu^l_+ y}\,L^l_+ \;+\; \lambda\,\Theta^l_+(y,\lambda)\,\Big). \tag{4.60}$$

Hence the leading term

$$\mathrm{e}^{\nu_+^l(x-y)} R_+^l L_+^l$$

is part of the scattering term S_λ, and an easy computation shows that the remaining terms are part of R_λ^S, thanks to the estimate (4.53). We get an additional power of λ in the estimate of $\partial_y G_\lambda$ using (4.11), (4.58) and that $\mu_+^l(\lambda) = \mathcal{O}(\lambda)$.

- Next, we look at

$$g_\lambda^{n,m}(x,y) := \beta_{n,m}(\lambda) T_\lambda(x,0) P_+(0,\lambda) \mathbb{W}_-^n(0,\lambda) \mathbb{Y}_+^m(y,\lambda)$$

for $1 \le n, m \le N$. If $n \neq k$, $\beta_{n,m}$ is analytic, as proved in Theorem 4.7, and thanks to (3.53), (3.55), we get the estimate[7]

$$|\, g_\lambda^{n,m}(x,y)\,| \lesssim \mathrm{e}^{-\omega x} \left| \mathrm{e}^{-\mu_+^m y} \right| \lesssim \mathrm{e}^{-(\omega/2)x} \left| \mathrm{e}^{\mu_+^m (x-y)} \right|. \tag{4.61}$$

This is a part of the residual term R_λ. The additional power of λ in the y-derivative also comes from Lemma 4.10. If $n = k$ and $m \ge k+1$, then $\beta_{n,m}$ by Lemma 4.9; hence using again (3.55), (3.53), we have an estimate similar to (4.61), and this term is a part of the residual term. The critical case is $n = k$, $m \le k$, when $\beta_{n,m}$ has a pole of order 1 at $\lambda = 0$ and $\overline{Y}_+^m$ is a slow mode. We can write

$$\beta_{k,m}(\lambda) = \frac{c_m}{\lambda} + g_m(\lambda),$$

where g_m is analytic in C_r. Then, thanks to Lemma 4.10, we have

$$g_\lambda^{k,m}(x,y) = \Big(\frac{c_m}{\lambda} + g_m(\lambda)\Big) \Big(\mathbb{U}'(x) + \lambda\, \mathcal{O}(\mathrm{e}^{-\omega x})\Big) \Big(\mathrm{e}^{-\mu_+^m y} L_+^m + \lambda\, \Theta_m^+(y,\lambda)\Big).$$

The leading term $\lambda^{-1}\, c_m\, \mathbb{U}'(x)\, \mathrm{e}^{-\nu_+^m y}\, L_+^m$ is a part of the excited term E_λ. We easily show, thanks to Lemma 4.10, that the remaining part verifies the estimates of R_λ^E.

- We finally turn to the terms

$$g_\lambda^{l,m}(x,y) := \gamma_{ji}(\lambda) \mathbb{W}_+^j(x,\lambda) \mathbb{Y}_+^i(y,\lambda)$$

for $1 \le m \le N$, $N + k + 1 \le l \le 2N$. As stated in Theorem 4.7, $\gamma_{l,m}$ is analytic. Consequently, for $m \le k$, using (3.52) and Lemma 4.10, we have

$$\begin{aligned} g_\lambda^{l,m}(x,y) = {} & \Big(\widetilde{\gamma}_{l,m} + \mathcal{O}(\lambda)\Big) \mathrm{e}^{\mu_+^l(\lambda)x} \Big(R_+^l + \mathcal{O}(\mathrm{e}^{-\omega x})\Big) \\ & \times \Big(\mathrm{e}^{-\mu_+^m(\lambda)y} L_+^m + \lambda \Theta_m^+(y,\lambda)\Big). \end{aligned}$$

As previously, the leading term

$$\widetilde{\gamma}_{l,m} \mathrm{e}^{\nu_+^l(\lambda)x - \nu_+^m(\lambda)y} R_+^l L_+^m$$

[7] Here and below, the notation $\lesssim$ stands for $\le$ up a to a harmless multiplicative constant.

is a part of the scattering term S_λ, and the remaining terms are part of R^S_λ. For $m \geq k+1$, thanks to Lemma 4.9, we have $\gamma_{l,m}(\lambda) = \mathcal{O}(\lambda)$, and thus, thanks to (3.52), (3.53), we get

$$|g^{l,m}_\lambda(x,y)| \lesssim |\lambda| \left| \mathrm{e}^{\mu^l_+(\lambda)x} \right| \mathrm{e}^{-\omega y} \lesssim |\lambda| \left| \mathrm{e}^{\mu^l_+(\lambda)(x-y)} \right| \mathrm{e}^{-(\omega/2)y}.$$

Hence this term is part of R^S_λ. Note that the additional $|\lambda|$ in this term due to Lemma 4.9 is crucial. ∎

4.4. Pointwise estimates on the Green's function. We now come to our main theorem, which describes very precisely the behavior of the Green's function for the LDS.

THEOREM 4.11. *Assuming* (H4)–(H6), *the Green's function* $G^{j_0}_j(t,t_0)$ *can be decomposed into*

$$G^{j_0}_j(t,t_0) = E(\tau,x,y) + S(\tau,x,y) + R(\tau,x,y), \tag{4.62}$$

where we have set $x = j - st$, $y = j_0 - st_0$, $\tau = t - t_0$. *For* $y \geq 0$ ($j_0 \geq st_0$) *we have the expansions*

$$E(\tau,x,y) = \sum_{\lambda^m_+ < 0} c^+_m \left(errfnc\left(\frac{y + |\lambda^m_+|\tau}{\sqrt{4d^m_+\tau}} \right) - errfnc\left(\frac{y - |\lambda^m_+|\tau}{\sqrt{4d^m_+\tau}} \right) \right) U'(x) \ell^m_+, \tag{4.63}$$

where

$$errfnc(X) := -\pi^{-\frac{1}{2}} \int_X^{+\infty} \mathrm{e}^{-z^2}\, dz, \tag{4.64}$$

$$\begin{aligned} S(\tau,x,y) = {} & \chi_{\tau\geq 1} \sum_{\lambda^m_+ > 0} \frac{|\lambda^m_+|}{\sqrt{4\pi d^m_+ \tau}} \mathrm{e}^{-\frac{(x-y-\lambda^m_+\tau)^2}{4d^m_+\tau}} r^m_+ \ell^m_+ \\ & + \chi_{\tau\geq 1}\chi_{x\geq 0} \sum_{\lambda^m_+ < 0} \frac{|\lambda^m_+|}{\sqrt{4\pi d^m_+ \tau}} \mathrm{e}^{-\frac{(x-y-\lambda^m_+\tau)^2}{4d^m_+\tau}} r^m_+ \ell^m_+ \\ & + \chi_{\tau\geq 1}\chi_{x\geq 0} \sum_{\lambda^m_+ < 0,\, \lambda^n_+ > 0} \widetilde{\gamma}^+_{n,m} \frac{1}{\sqrt{4\pi d^{n,m}_+ \tau}} \mathrm{e}^{-\frac{(x-z^{n,m}_+)^2}{4d^{n,m}_+\tau}} \\ & + \chi_{\tau\geq 1}\chi_{x\leq 0} \sum_{\lambda^m_+ < 0,\, \lambda^n_- < 0} \widetilde{\gamma}^+_{n,m} \frac{1}{\sqrt{4\pi d^{n,m}_- \tau}} \mathrm{e}^{-\frac{(x-z^{n,m}_-)^2}{4d^{n,m}_-\tau}} \end{aligned} \tag{4.65}$$

and

$$R = R_1 + R^{exp},$$

where

$$R^{exp}(\tau,x,y) = \mathcal{O}(\mathrm{e}^{-\omega(\tau+|x-y|)}) \tag{4.66}$$

and for $\alpha \le 1$, there exists $M > 0$ such that

$$\partial_y^\alpha R_1 = \mathcal{O}\left\{ \sum_{m=1}^{N} \frac{\mathrm{e}^{-\frac{(x-y-\lambda_+^m \tau)^2}{M\tau}}}{(1+\tau)^{\frac{\alpha+1}{2}}} \left(\mathrm{e}^{-\omega|x|} + \frac{1}{(1+\tau)^{\frac{1}{2}}} \right) \right. \tag{4.67}$$
$$+ \sum_{\lambda_+^m<0,\, \lambda_+^n>0} \frac{\mathrm{e}^{-\frac{(x-z_+^{n,m})^2}{M\tau}}}{(1+\tau)^{\frac{\alpha+1}{2}}} \left(\mathrm{e}^{-\omega|x|} + \frac{1}{(1+\tau)^{\frac{1}{2}}} \right)$$
$$\left. + \sum_{\lambda_+^m<0,\, \lambda_-^n<0} \frac{\mathrm{e}^{-\frac{(x-z_-^{n,m})^2}{M\tau}}}{(1+\tau)^{\frac{\alpha+1}{2}}} \left(\mathrm{e}^{-\omega|x|} + \frac{1}{(1+\tau)^{\frac{1}{2}}} \right) \right\},$$

all these estimates being uniform in j, j_0, t, t_0, $t \ge t_0$. The characteristics $z_\pm^{n,m}$ and the diffusion coefficients d_+^m, $d_\pm^{n,m}$ are defined by

$$z_\pm^{n,m} = \lambda_\pm^n \left(\tau - \frac{|y|}{|\lambda_+^m|} \right), \qquad d_+^m = \frac{a_+^m}{2}, \quad \textit{and} \qquad d_\pm^{n,m} = \frac{a_\pm^n\,|x|}{2|\lambda_\pm^l|\tau} + \frac{a_+^m\,|\lambda_\pm^n|^2\,|y|}{2|\lambda_+^m|^3\tau}.$$

Symmetric estimates hold for $y \le 0$.

Remark 4.12. We see that the scattering term S involves some heat kernels propagating along the characteristic paths ($x = y + \lambda_+^m \tau$) of the underlying continuous system of conservation laws (2.4). Unsurprisingly, the diffusion coefficients $d_+^m = \frac{a_+^m}{2}$ of these heat kernels correspond to the viscosity of the upwind scheme. As a matter of fact, it is well known that the semidiscrete system (2.1) is equivalent to the viscous system of conservation laws

$$\partial_t u + \partial_x(f(u)) = \partial_x(\, B(u)\, \partial_x u\,), \quad \text{where} \quad B(u) = \tfrac{\Delta x}{2}\, df(u), \tag{4.68}$$

up to higher order terms in Δx. The decomposition of the Green's function obtained in Theorem 4.11 is actually very similar to the one obtained in [23] (for viscous shocks) and [18] (for relaxation shocks; in this case, there are additional terms due to a singular short time behavior). This theorem confirms a posteriori that the qualitative behavior of the solutions of the semidiscrete scheme (2.1) is analogous to the behavior of the solutions of the viscous system of conservation laws (4.68), *regardless* of the order of magnitude of Δx.

Despite its apparent technicality, this theorem sheds light on several important phenomena. It means that an elementary signal originated at (j_0, t_0), $j_0 > st_0$, behaves, up to a neglectible term, as moving heat kernels along characteristic paths. Additionally, there are interaction terms involving both incoming and outgoing waves (with respect to the shock) in the scattering (and remainder) terms. When an incoming wave hits the shock (which happens after a time $\tau = \frac{|y|}{|\lambda_+^m|}$), there is an excited part described by the term E, due to accumulation of mass in the shock layer and reemission of signals on both sides of the shock, which propagate along the paths $z_\pm^{n,m}$ at an averaged diffusion rate $d_\pm^{n,m}$.

Proof. Our short frequency estimates given by Theorem 4.8 are completely similar to the estimates of Proposition 5.2 in [18]. Thanks to (4.3) we can use the same paths as in [18] to get the estimates in Theorem 4.11 in the long time regime $\frac{|x-y|}{\tau} \lesssim 1$. The short time regime will come from (4.3) and the crucial property (4.5), which allows us to use Cauchy's formula. For clarity, we explain below the method on various significant examples.

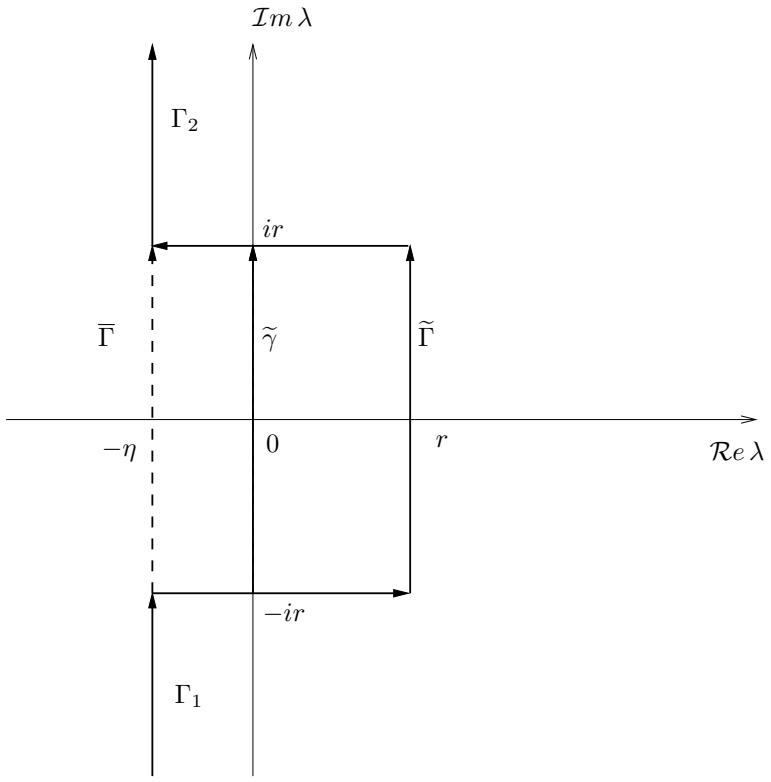


FIG. 4.1. *Examples of contours.*

We first use as a basic path $\Gamma = \Gamma_1 \cup \widetilde{\Gamma} \cup \Gamma_2$, where

$$\Gamma_1 = [-\eta - i\pi\sigma, -\eta - ir], \quad \Gamma_2 = [-\eta + ir, -\eta + i\pi\sigma],$$

$$\widetilde{\Gamma} = [-\eta - ir, r - ir] \cup [r - ir, r + ir] \cup [r + ir, -\eta + ir].$$

Note that all our contours will have the same shape, the main difference being the point where they cross the real axis. Some examples are given in Figure 4.1.

By the analyticity of G_λ (see Proposition 4.4), property (4.5), and Cauchy's theorem we have

$$G_j^{j_0}(t, t_0) = I + \widetilde{I},$$

where

$$I = \frac{1}{2i\pi\sigma} \int_{\Gamma_1 \cup \Gamma_2} \mathrm{e}^{\lambda\tau} G_\lambda(x, y), d\lambda, \quad \widetilde{I} = \frac{1}{2i\pi\sigma} \int_{\widetilde{\Gamma}} \mathrm{e}^{\lambda\tau} G_\lambda(x, y) d\lambda.$$

The integral I is very easy to handle. By (4.20) (see Corollary 4.6), we have

$$I \lesssim \mathrm{e}^{-\eta\tau}\, \mathrm{e}^{-\alpha|x-y|},$$

which contributes only to the fastly decreasing residual term R^{exp}.

The most difficult part is to estimate the integral $\widetilde{I}$. We must distinguish between several space-time regimes.

- Case $\frac{|x-y|}{\tau} \gg 1$.
  If $|x-y| \geq V\tau$, every term in the expansion (4.62) is bounded by $\mathrm{e}^{-c|x-y|}$, and hence by

$$\mathrm{e}^{-\frac{c}{2}|x-y|}\,\mathrm{e}^{-\frac{c}{2}V\tau}$$

  for some $c > 0$. We just have to show that $\widetilde{I}$ enjoys a similar estimate. Using again (4.20), we have

$$\widetilde{I} \lesssim \mathrm{e}^{r\tau}\,\mathrm{e}^{-\omega|x-y|} \leq \mathrm{e}^{-\frac{\omega}{2}|x-y|}\,\mathrm{e}^{(r-V\frac{\omega}{2})\tau}.$$

  Choosing V so large that $r - V\omega/4 < 0$, we have the requested estimate with $c = \omega/2$.
- Case $\frac{|x-y|}{\tau} \leq V$. This is the more difficult one. We need the expansion of G_λ given by Theorem 4.8, which yields

$$\widetilde{I} = J_1 + J_2 + J_3,$$

  where

$$J_1 = \frac{1}{2i\pi\sigma}\int_{\widetilde{\Gamma}} \mathrm{e}^{\lambda\tau}E_\lambda\,d\lambda,\ J_2 = \frac{1}{2i\pi\sigma}\int_{\widetilde{\Gamma}} \mathrm{e}^{\lambda\tau}S_\lambda\,d\lambda,\ J_3 = \frac{1}{2i\pi\sigma}\int_{\widetilde{\Gamma}} \mathrm{e}^{\lambda\tau}R_\lambda\,d\lambda.$$

  We are going to estimate each term separately. Since there are many cases, we concentrate on the case $x > y$, the other ones being left to the reader (they can be handled by similar techniques).

Estimate of J_2. This term will appear to contribute to S and R. Using again Theorem 4.8 (case $x > y$), we have two types of terms to deal with, namely

$$\alpha_l = \frac{1}{2i\pi\sigma}\int_{\widetilde{\Gamma}} \mathrm{e}^{\lambda\tau}\mathrm{e}^{\nu_+^l(\lambda)(x-y)}\,d\lambda,\ n+k+1 \leq l \leq 2N$$

and

$$\alpha_{l,m} = \frac{1}{2i\pi\sigma}\int_{\widetilde{\Gamma}} \mathrm{e}^{\lambda\tau}\mathrm{e}^{\nu_+^l(\lambda)x}\mathrm{e}^{-\nu_+^m(\lambda)(y)}\,d\lambda,\ m \leq k,\ N+k+1 \leq l \leq 2N.$$

* Estimate of α_l. A preliminary remark concerns the case

$$\frac{|x-y|}{\tau} \ll 1.$$

  Using again Cauchy's formula, we can change the path $\widetilde{\Gamma}$ into $\overline{\Gamma} := [-\eta - ir, -\eta + ir]$ and thus show that

$$|\alpha_l| \lesssim \mathrm{e}^{-\eta\tau}\mathrm{e}^{C|x-y|} \leq \mathrm{e}^{-(\eta - C\varepsilon)\tau}$$

  for $|x-y| \leq \varepsilon\tau$. Hence choosing ε so small that $\eta - C\varepsilon/2 > 0$ we have the bound

$$|\alpha_l| \lesssim \mathrm{e}^{-\frac{C\varepsilon}{2}\tau} \leq \mathrm{e}^{-\frac{C\varepsilon}{4}\tau}\mathrm{e}^{-\frac{C}{4}|x-y|}. \tag{4.69}$$

  Therefore, α_l can be absorbed by the residual term R^{exp}. Since the excited term E and the scattering term S are also bounded by the right-hand side of

(4.69) this estimate agrees with (4.62) when $|x-y| \le \varepsilon\tau$. Consequently, we can restrict our study to

$$\varepsilon \le \frac{|x-y|}{\tau} \le V. \tag{4.70}$$

We use once more Cauchy's formula to change $\widetilde{\Gamma}$ into $\gamma = \gamma_1 \cup \widetilde{\gamma} \cup \gamma_2$, where

$$\gamma_1 = [-\eta - ir, -ir], \quad \gamma_2 = [ir, -\eta + ir],\ \widetilde{\gamma} = [-ir, ir].$$

We easily show that the integrals on γ_1 and γ_2 are bounded by residual terms similarly as in (4.69), since on these paths we have

$$\operatorname{Re} \nu_+^l(\lambda) \le -\beta \tag{4.71}$$

for some $\beta > 0$ (thanks to (3.49) and (4.35)). It remains to estimate

$$\widetilde{\alpha}_l := \frac{1}{2i\pi\sigma} \int_{\widetilde{\gamma}} \mathrm{e}^{\lambda\tau} \mathrm{e}^{\nu_+^l(\lambda)(x-y)}\, d\lambda = \frac{1}{2\pi} \int_{\mathbb{R}} \mathrm{e}^{i\xi\tau} \mathrm{e}^{\nu_+^l(i\xi)(x-y)}\, d\xi$$

$$-\frac{1}{2\pi} \int_{-\infty}^{r} \mathrm{e}^{i\xi\tau} \mathrm{e}^{\nu_+^l(i\xi)(x-y)}\, d\xi - \frac{1}{2\pi} \int_{r}^{+\infty} \mathrm{e}^{i\xi\tau} \mathrm{e}^{\nu_+^l(i\xi)(x-y)}\, d\xi.$$

We easily show that the last two integrals hereabove can be bounded by (4.69), and hence are part of the residual term R^{exp}. Indeed, using (4.71), we have

$$\left| \int_r^{+\infty} \mathrm{e}^{i\xi\tau} \mathrm{e}^{\nu_+^l(i\xi)(x-y)}\, d\xi \right| \lesssim \int_r^{+\infty} \mathrm{e}^{-\beta\xi^2(x-y)}\, d\xi \le \int_r^{+\infty} \frac{\xi}{r} \mathrm{e}^{-\beta\xi^2(x-y)}\, d\xi$$
$$\le \frac{\mathrm{e}^{-\beta r^2(x-y)}}{2r\beta(x-y)} \lesssim \frac{1}{\tau}\, \mathrm{e}^{-\frac{\varepsilon\beta r^2}{2}\tau}\, \mathrm{e}^{-\frac{\beta r^2}{2}(x-y)}$$

because of (4.70). So the estimate of $\widetilde{\alpha}_l$ amounts to computing

$$\underline{\alpha}_l := \frac{1}{2\pi\sigma} \int_{\mathbb{R}} \mathrm{e}^{i\xi\tau} \mathrm{e}^{\nu_+^m(i\xi)(x-y)}\, d\xi,$$

which can be done by Fourier transform. As a matter of fact, using (4.35), $\underline{\alpha}_l$ is just the Fourier transform of a Gaussian. This yields

$$\underline{\alpha}_l = \frac{\lambda_+^l}{\sqrt{4\pi \frac{a_+^l(x-y)}{2(\lambda_+^l)}}} \exp\left(-\frac{(x-y-\lambda_+^l\tau)^2}{4\frac{a_+^l}{2\lambda_+^l}(x-y)} \right) = \lambda_+^l k\left(X, \frac{x-y}{\lambda_+^m} \right),$$

where $X = x - y - \lambda_+^l\tau$ and

$$k(X,z) = \frac{1}{\sqrt{4\pi d_+^l z}} \mathrm{e}^{-\frac{X^2}{4d_+^l z}}, \quad d_+^l = \frac{a_+^l}{2} > 0$$

is the heat kernel. (Recall that by convention $a^l = a^{l-N}$.) To get the final form of the expansion (4.62), we write

$$k\left(X, \frac{x-y}{\lambda_+^l} \right) = k(X,\tau) + k\left(X, \frac{x-y}{\lambda_+^l} \right) - k(X,\tau). \tag{4.72}$$

The first term $k(X,\tau)$ is precisely a part of the scattering term S in (4.62). Recalling that $X = x - y - \lambda_+^l \tau$ and using (4.70), we have by the mean value theorem

$$\left| k\left(X, \frac{x-y}{\lambda_+^l}\right) - k(X,\tau)\right| \lesssim \frac{X}{\tau^{\frac{3}{2}}} \mathrm{e}^{-\frac{X^2}{M\tau}} \lesssim \frac{1}{\tau} \mathrm{e}^{-\frac{X^2}{M\tau}}$$

for some positive M and

$$\left| \partial_y \left(k\left(X, \frac{x-y}{\lambda_+^m}\right) - k(X,\tau)\right)\right| \lesssim \frac{1}{\tau^{\frac{3}{2}}} \mathrm{e}^{\frac{-X^2}{M\tau}}.$$

We have used hereabove the obvious estimate

$$|u|\mathrm{e}^{-u^2} \lesssim \mathrm{e}^{-u^2/2}.$$

This will be done repeatedly, without explicit mention. Consequently, the terms $\underline{\alpha}_l$ contribute only to the residual term R in (4.62).

* Estimate of $\alpha_{l,m}$, $m \le k$, $N + k + 1 \le l \le 2N$. Using the same technique as previously, we see that, up to a residual decreasing term like in (4.69), we can restrict the study to

$$\varepsilon \le \frac{x+y}{\tau} \le V. \tag{4.73}$$

Furthermore, using the same change of paths, we see that $\alpha_{l,m} = \underline{\alpha}_{l,m}$+ h.o.t , where

$$\underline{\alpha}_{l,m} = \frac{1}{2\pi\sigma} \int_{\mathbb{R}} \mathrm{e}^{i\xi\tau} \mathrm{e}^{\nu_+^l(i\xi)x} \mathrm{e}^{-\nu_+^m(i\xi)y}\, d\xi.$$

Using (4.35), we can still compute explicitly this Fourier transform. We find

$$\underline{\alpha}_{l,m} = \frac{1}{\sqrt{4\pi d_+^{l,m}}} \exp\left(- \frac{\left(x - \lambda_+^l \left(\tau - \frac{y}{|\lambda_+^m|}\right)\right)^2}{4 d_+^{l,m}}\right),$$

where the diffusion coefficient $d_+^{l,m}$ is

$$d_+^{l,m} = \frac{a_+^l}{2\lambda_+^l}|x| + \frac{a_+^m}{2(\lambda_+^m)^3}(\lambda_+^l)^2|y|.$$

This ends the expansion of J_2. Note that we do not find all the terms in the expansion of S as they are given in Theorem 4.11; we find only the terms corresponding to $\lambda_+^m > 0$ and $\lambda_+^l < 0$, $\lambda_+^m > 0$. Nevertheless, the formula of Theorem 4.11 is true, since the other terms in S are exponentially decreasing as in (4.69), and hence can be considered as a part of R^{exp}.

Estimate of J_1. Using (4.37) in Theorem 4.11 we have to estimate terms like

$$b_m(t,y) = \frac{1}{2i\pi\sigma} \int_{\widetilde{\Gamma}} \frac{\mathrm{e}^{\lambda\tau}}{\lambda} \mathrm{e}^{-\nu_+^m(\lambda)y}\, d\lambda,\ m \le k.$$

Using the residues theorem, we change the path of integration $\widetilde{\gamma}$ into $[-\eta - ir, -\eta + ir]$ and obtain

$$b_m(\tau,y) = 1 + \widetilde{b}_m(\tau,y), \quad \widetilde{b}_m(\tau,y) = \frac{1}{2i\pi} \int_{[-\eta-ir,-\eta+ir]} \frac{\mathrm{e}^{\lambda\tau}}{\lambda} \mathrm{e}^{-\nu_+^m(\lambda)y}\, d\lambda.$$

For fixed y,

$$|\widetilde{b}_m(\tau,y)| \le C(y)\mathrm{e}^{-\eta\tau} \to 0 \text{ when } t \to +\infty; \tag{4.74}$$

hence

$$\widetilde{b}_m(\tau,y) = -\int_\tau^{+\infty} \partial_\tau \widetilde{b}_m(\theta,y)\,ds. \tag{4.75}$$

Since

$$\partial_\tau \widetilde{b}_m(\theta,y) = \frac{1}{2i\pi\sigma}\int_{[-\eta-ir,-\eta+ir]} \mathrm{e}^{\lambda\theta}\mathrm{e}^{-\nu_+^m(\lambda)y}\,d\lambda$$

this term is very similar to $\widetilde{\alpha}_l$. The same computation yields

$$\partial_\tau \widetilde{b}_m(\theta,y) = \frac{1}{\sqrt{4\pi\frac{a_+^m y}{2|\lambda_+^m|}}}\exp\left(-\frac{(y-|\lambda_+^m|\theta)^2}{4\frac{a_+^m}{2|\lambda_+^m|}y}\right) + \mathcal{O}(\mathrm{e}^{-\omega(\theta+y)})$$

for some $\omega > 0$. Thanks to (4.74), (4.75), we get

$$b_m(\tau,y) = \mathrm{errfnc}\left(\frac{y-|\lambda_+^m|\tau}{4\frac{a_+^m}{2|\lambda_+^m|}y}\right) + \mathcal{O}(\mathrm{e}^{-\omega(\theta+y)}),$$

where the errfnc function is defined as in (4.64). Finally, using the same trick as in (4.72), we write

$$b_m(\tau,y) = \mathrm{errfnc}\left(\frac{y-|\lambda_+^m|\tau}{4\frac{a_+^m}{2}\tau}\right) + b^r(\tau,y) + \mathcal{O}(\mathrm{e}^{-\omega(\theta+y)}),$$

$$b^r(\tau,y) = \mathrm{errfnc}\left(\frac{y-|\lambda_+^m|\tau}{4\frac{a_+^m}{2|\lambda_+^m|}y}\right) - \mathrm{errfnc}\left(\frac{y-|\lambda_+^m|\tau}{4\frac{a_+^m}{2}\tau}\right).$$

The first term is a part of the excited term E in (4.62) and, since we easily find that

$$|\partial_y^p b^r(\tau,y)| \le C\frac{1}{\tau^{1+p/2}}\mathrm{e}^{-\frac{(y-|\lambda_+^m\tau)^2}{M\tau}}, \quad p \ge 1,$$

for some $M > 0$, the other term is a part of the residual term R in the expansion (4.62). Note that in the final expansion (4.62) these terms are multiplied by U', which is bounded by some $\mathrm{e}^{-\omega x}$.

Estimate of J_3. Using (4.41), (4.42), $\partial_y^\alpha J_3$ is given by a sum of many terms which can be handled by similar techniques. We shall just give an example, concerning the term

$$K_m := \int_{\widetilde{\Gamma}} \mathrm{e}^{\lambda\tau}\mathrm{e}^{-\nu_+^m(\lambda)y}\mathrm{e}^{-\omega x}\mathcal{O}\Big(\lambda^{\alpha-1}(\mathrm{e}^{\mathcal{O}(\lambda^3)y}-1) + \lambda^{\alpha-1}(\mathrm{e}^{\mathcal{O}(\lambda^3)x}-1+\lambda^\alpha\Big)\,d\lambda$$

for $1 \le m \le k$, $x > 0$, and $y > 0$, which can be rewritten as

$$K_m = \int_{\widetilde{\Gamma}} \mathrm{e}^{\varphi(\lambda;\tau,x,y)}\,\mathcal{O}\Big(|\lambda|^{\alpha+2}y\mathrm{e}^{\mathcal{O}(\lambda^3)y} + |\lambda|^{\alpha+2}x\mathrm{e}^{\mathcal{O}(\lambda^3)x} + |\lambda|^\alpha\Big)\,d\lambda,$$

where the phase φ is defined by

$$\varphi(\lambda;\tau,x,y) \,=\, \lambda\tau - \nu_+^m(\lambda)y - \omega x \,=\, \lambda\tau - \frac{\lambda}{|\lambda_+^m|}y + \frac{a_+^m y}{2|\lambda_+^m|^3}\lambda^2 - \omega x.$$

(Recall that $\lambda_+^m = a_+^m - \sigma < 0$ for $m \le k$.) As in [24, 18], we may use the stationary phase method, or more precisely the method of steepest descent, to estimate K_m. The idea is to change the path $\widetilde{\Gamma}$ into another one that passes through the stationary point of the phase and still lies in the part of C_r where the estimates are valid. The stationary point of the phase is given by

$$\lambda = \lambda_s = \frac{y - |\lambda_+^m|\tau}{a_+^m y}\,|\lambda_+^m|^2.$$

When $|\lambda_s| \le \varepsilon$, $\varepsilon > 0$ sufficiently small, we change $\widetilde{\Gamma}$ into $\Gamma_s = S_s \cup \gamma_s$, where

$$S_s = [-\eta - ir, \lambda_s - ir] \cup [\lambda_s + ir, -\eta + ir], \quad \gamma_s = [\lambda_s - ir, \lambda_s + ir].$$

We have $\mathrm{Re}\,(-\nu_+^m) \le -\beta$ for some $\beta > 0$ on S_s. And $|\lambda_s| \le \varepsilon$ implies that

$$\frac{|\lambda_+^m|\,\tau}{1 \,+\, \varepsilon/\varepsilon_m} \,\le\, y \,\le\, \frac{|\lambda_+^m|\,\tau}{1 \,-\, \varepsilon/\varepsilon_m}, \tag{4.76}$$

where $\varepsilon_m := |\lambda_+^m|^2/a_+^m$. Therefore, the phase φ enjoys the estimate

$$\mathrm{Re}\varphi(\lambda;\tau,x,y) \,\le\, \lambda_s\tau - \beta y - \omega x \,\le\, -(\,\beta \,-\, \varepsilon\,(1 \,+\, \varepsilon/\varepsilon_m)/|\lambda_+^m|\,)\,y \,-\, \omega x,$$

which yields an estimate of the integral

$$\left|\int_{S_s} \mathrm{e}^{\varphi}\,\mathcal{O}\Big(\dots\Big)\,d\lambda\right| \lesssim \mathrm{e}^{-\omega x/2}\mathrm{e}^{-\beta y/2}$$

for ε small enough. Thus this part of K_m contributes to R^{exp} only. The main contribution of K_m comes from the integral on γ_s. After some computations, we get

$$\left|\int_{\gamma_s} \mathrm{e}^{\varphi}\,\mathcal{O}\Big(\dots\Big)\,d\lambda\right| \,\lesssim\, \mathrm{e}^{-\omega x}\mathrm{e}^{-\frac{(y-|\lambda_+^m|\tau)^2}{My}} \int_{-r}^{r} \mathrm{e}^{-\frac{\xi^2 y}{M}}\Big(|\lambda_s|^{\alpha+2}y + |\lambda_s|^{\alpha} + |\xi|^{\alpha}\Big)\,d\xi$$

for some $M > 0$ sufficiently large, which yields

$$\left|\int_{\gamma_s} \mathrm{e}^{\varphi}\,\mathcal{O}\Big(\dots\Big)\,d\lambda\right| \,\lesssim\, \frac{1}{\sqrt{y}^{\alpha+1}}\mathrm{e}^{-\omega x}\mathrm{e}^{-\frac{(y-|\lambda_+^m|\tau)^2}{My}}.$$

Moreover, because of (4.76), we can replace y by τ. This part of K_m thus contributes to the residual term $\partial_y^{\alpha} R_1$.

It remains to study the case $|\lambda_s| \ge \varepsilon$. We consider $\Gamma_{\pm\varepsilon} = S_{\pm\varepsilon} \cup \gamma_{\pm\varepsilon}$, where

$$S_{\pm\varepsilon} = [-\eta - ir, \pm\varepsilon - ir] \cup [\pm\varepsilon + ir, -\eta + ir], \quad \gamma_{\pm\varepsilon} = [\pm\varepsilon - ir, \pm\varepsilon + ir],$$

and take as an approximate optimal path Γ_ε if $\lambda_s \ge \varepsilon$ and $\Gamma_{-\varepsilon}$ if $\lambda_s \le -\varepsilon$.

We just briefly explain how to handle the case $\lambda_s \ge \varepsilon$. As previously, the integrals on S_ε are part of R^{exp}. To estimate the integral on γ_ε, we notice that when $\mathrm{Re}\,\lambda = \varepsilon$,

$$\operatorname{Re}\varphi(\lambda,\tau,x,y)=\varphi(\varepsilon,\tau,x,y)-\frac{a_+^m y}{2|\lambda_+^m|^3}(\operatorname{Im}\lambda)^2$$

$$=\varepsilon\left(\tau-\frac{y}{|\lambda_+^m|}\right)-\frac{a_+^m y}{2|\lambda_+^m|^3}(\operatorname{Im}\lambda)^2-\omega x.$$

Consequently, using that $\lambda_s\geq\varepsilon$, we get

$$\operatorname{Re}\varphi(\lambda,\tau,x,y)\leq-\frac{\varepsilon^2 y}{\varepsilon_m|\lambda_+^m|}\;-\;\frac{a_+^m y}{2|\lambda_+^m|^3}(\operatorname{Im}\lambda)^2-\omega x,$$

where $\varepsilon_m=|\lambda_+^m|^2/a_+^m$ as before. But $\lambda_s\geq\varepsilon$ implies that

$$y\;\geq\;\frac{|\lambda_+^m|\,\tau}{1\,-\,\varepsilon/\varepsilon_m}.$$

Hence

$$\operatorname{Re}\varphi(\lambda,\tau,x,y)\leq\;-\frac{\varepsilon^2}{2}y-\frac{\varepsilon^2}{2(\varepsilon_m-\varepsilon)}\tau-\omega x$$

on γ_ε and therefore

$$\left|\int_{\gamma_\varepsilon}\mathrm{e}^{\varphi}\;\mathcal{O}\Big(\dots\Big)\,d\lambda\right|\;\lesssim\;\mathrm{e}^{-\varepsilon^2 y/4}\,\mathrm{e}^{-\varepsilon^2\tau/(4\varepsilon_m)}\,\mathrm{e}^{-\omega x/2}$$

falls into R^{exp}. ∎

5. Nonlinear stability of semidiscrete shocks. We now have all the ingredients to show the nonlinear stability of a semidiscrete shock profile $u_j(t)=U(j-st)$, solution of the upwind scheme

$$\frac{dv_j}{dt}+f(v_j)-f(v_{j-1})=0. \tag{5.1}$$

Of course, there is a one-dimensional manifold of profiles, since for all $\delta\in\mathbb{R}$, $u_j^\delta(t)=u_j(t+\delta)$ is another semidiscrete shock profile. So we cannot expect genuine asymptotic stability. Similarly as for other kinds of shock profiles [24, 18], the relevant notion is orbital stability.

THEOREM 5.1. *Assuming* (H5)–(H6), *there exist* $\delta>0$ *and* $C>0$ *such that any solution of* (5.1) *with initial data* $v_j(0)=u_j(0)+w_j^0$, *where* $w^0\in\mathcal{L}^1(\mathbb{Z})$ *and* $||w^0||_1\leq\delta$, *verifies*

$$||v(t)-u(t+\widetilde{p}(t))||_{\mathcal{L}^\alpha}\leq\frac{C\delta}{(1+t)^{\frac{1}{2}\left(1-\frac{1}{\alpha}\right)}}\quad\forall t\geq0,\quad\forall\alpha\geq1, \tag{5.2}$$

where $\widetilde{p}(t)$, *the perturbation of the shock position, verifies*

$$|\widetilde{p}(t)|\leq C\delta\quad\forall t\geq0, \tag{5.3}$$

$$|\widetilde{p}'(t)|\leq\frac{C\delta}{(1+t)^{\frac{1}{2}}}\quad\forall t\geq0. \tag{5.4}$$

Proof. The proof is close to the proof of stability of viscous and relaxation shock profiles of [23, 18]. It uses the Green's function bounds obtained in the previous section. By an obvious interpolation argument, it suffices to prove the theorem in $\mathcal{L}^1$ and $\mathcal{L}^\infty$. Also note that in our discrete setting we have the embedding $\mathcal{L}^1 \subset \mathcal{L}^\infty$.

Let us consider the solution of (5.1) with initial data $u(0) + w^0$. We set

$$w(t) = v(t + p(t)) - u(t), \tag{5.5}$$

where $p(t)$ is such that $p(0) = 0$. Using that both u and v are solutions of (5.1), we can rewrite the system satisfied by w as

$$\frac{dw_j}{dt} + A_j(t)w_j(t) - A_{j-1}(t)w_{j-1}(t) = S_j(t, w), \tag{5.6}$$

where $A_j(t) = df(u_j(t))$,

$$S(t, w) = (I-\mathcal{T})Q(w) - p'(t)\,(I-\mathcal{T})f(u+w), \quad Q(w) = -f(u+w) + f(u) + df(u)w.$$

Another way of writing the source term in (5.6) uses that

$$(I - \mathcal{T})f(u + w) = (I - \mathcal{T})\, f(u) + (I - \mathcal{T})\, N(w) = -\frac{du}{dt} + (I - \mathcal{T})\, N(w),$$

where $N(w) = f(u + w) - f(u)$, and thus

$$S(t, w) = (I - \mathcal{T})\,(\,Q(w) - p'(t)\, N(w)\,) + p'(t)\frac{du}{dt}. \tag{5.7}$$

Using the Green's function $G_j^{j_0}(t, t_0)$ and (5.7), the solution of (5.6) is given by

$$\begin{aligned} w_j(t) = G_j(t, 0) * w(0) &+ \int_0^t G_j(t, t_0) * \Big((I - \mathcal{T})(Q(w) - p'N(w))(t_0)\Big)\, dt_0 \\ &+ \int_0^t G_j(t, t_0) * \frac{du}{dt}(t_0)p'(t_0)\, dt_0, \end{aligned}$$

where we have used the standard notation

$$G_j(t, t_0) * h = \sum_{j_0 \in \mathbb{Z}} G_j^{j_0}(t, t_0)h_{j_0} \quad \forall h = (h_j)_{j\in\mathbb{Z}} \in \mathcal{L}^1.$$

Note that, differentiating (5.1), we have

$$\frac{d}{dt}\frac{du_j}{dt} + A_j(t)\frac{du_j}{dt} - A_{j-1}(t)\frac{du_{j-1}}{dt} = 0$$

and therefore

$$\int_0^t G_j(t, t_0) * \frac{du}{dt}(t_0)\, p'(t_0)\, dt_0 = \int_0^t \frac{du_j}{dt}(t)\, p'(t_0)\, dt_0 = \frac{du_j}{dt}(t)\, p(t).$$

Consequently, after a "discrete integration by part," we can write w under the form

$$\begin{aligned} w_j(t) = \; & G_j(t, 0) * w^0 + \int_0^t (I - \mathcal{T}^{-1})G_j(t, t_0) * (Q(w) - p'N(w))(t_0)\, dt_0 \\ & + \frac{du_j}{dt}(t)\, p(t). \end{aligned} \tag{5.8}$$

Using the decomposition of the Green's function given by Theorem 4.11, we set

$$\widetilde{G}(\tau,x,y) = S(\tau,x,y) + R(\tau,x,y)$$

and

$$E(\tau,x,y) = U'(x)e(\tau,y). \tag{5.9}$$

Since

$$\frac{du_j}{dt}(t) \;=\; -\,s\,U'(j-st),$$

we split the problem of finding (w,p) into

$$s\,p(t) \;=\; e(t,\cdot)*w^0 \;+\; \int_0^t (I-\mathcal{T}^{-1})e(t-t_0,\cdot-st_0)*(Q(w)-p'N(w))(t_0)\,dt_0, \tag{5.10}$$

which describes the nonlinear evolution of the shock position, and

$$\begin{aligned} w_j(t) &= \widetilde{G}(t,j-st,\cdot-st_0)*w^0 \\ &\quad + \int_0^t (I-\mathcal{T}^{-1})\widetilde{G}(t-t_0,j-st,\cdot-st_0)*(Q(w)-p'N(w))(t_0)\,dt_0. \end{aligned} \tag{5.11}$$

To prove Theorem 5.1, it suffices to prove the existence of solutions of the fixed point problem (5.11), (5.10) in the Banach space X, where

$$\begin{aligned} X = \Big\{ &(w,p) \in \mathcal{C}^0([0,+\infty),\mathcal{L}^1)\times\mathcal{C}^1([0,+\infty),\mathbb{R}) \\ &||(w,p)||_X = \sup_{t\ge 0}\Big((1+t)||w(t)||_\infty + ||w(t)||_1 + |p(t)| + (1+t)^{\frac12}|p'(t)|\Big) \le \delta R\Big\} \end{aligned}$$

for some $R>0$ to be chosen. Let us define on $X\times\{w^0\in\mathcal{L}^1(\mathbb{Z}),\ ||w^0||_1\le\delta\}$

$$\mathcal{N}(w,p,w^0) = \begin{pmatrix} W(w,p,w^0) \\ \frac1s P(w,p,w^0) \end{pmatrix},$$

where W is equal to the right member of (5.11) and P is equal to the right member of (5.10). We want to use the classical Banach fixed point theorem to solve $\mathcal{N}(w,p,w^0) = (p,w)$.

The proof will rely on $\mathcal{L}^\alpha$ estimates of the Green's function coming from the expansions of Theorem 4.11. Thanks to classical comparisons between series and integrals we get for all $j_0\in\mathbb{Z}$ and for all t, t_0, $t>t_0$,

$$||\widetilde{G}(t-t_0,\cdot-st,j_0-st_0)||_{\mathcal{L}^\alpha} \le \frac{C}{(t-t_0)^{\frac12(1-\frac1\alpha)}}, \tag{5.12}$$

$$||(I-\mathcal{T}^{-1})\widetilde{G}(t-t_0,\cdot-st,j_0-st_0)||_{\mathcal{L}^\alpha} \le \frac{C}{(t-t_0)^{\frac12(2-\frac1\alpha)}}. \tag{5.13}$$

Note that our definition of e is similar to the one in [18, 23]; hence we can use the following key lemma from [23] (also see [18]).

LEMMA 5.2 (see [23]). *Let E be defined as in Theorem* 4.11 *and e as in* (5.9); *then*

$$||\partial_y e(\tau,\cdot)||_{L^\alpha(\mathbb{R})} + ||\partial_\tau e(\tau,\cdot)||_{L^\alpha(\mathbb{R})} \le \frac{C}{\tau^{\frac{1}{2}(1-\frac{1}{\alpha})}}, \tag{5.14}$$

$$||\partial_{\tau y} e(\tau,\cdot)||_{L^\alpha(\mathbb{R})} \le \frac{C}{\tau^{\frac{1}{2}(2-\frac{1}{\alpha})}}. \tag{5.15}$$

Moreover,

$$\lim_{\tau\to 0} ||e(\tau,\cdot)||_{L^1} = 0. \tag{5.16}$$

Note that using elementary comparisons between series and integrals, (5.14), (5.15), (5.16) are still true if we consider the discrete norm in $\mathcal{L}^\alpha(\mathbb{Z})$. Moreover, (5.14), (5.15) also implies that

$$||(I-\mathcal{T}^{-1})e(t-t_0,\cdot - st_0)||_{\mathcal{L}^\alpha} \le \frac{C}{(t-t_0)^{\frac{1}{2}(1-\frac{1}{\alpha})}}, \tag{5.17}$$

$$||\partial_\tau (I-\mathcal{T}^{-1})e(t-t_0,\cdot - st_0)||_m \le \frac{C}{(t-t_0)^{\frac{1}{2}(2-\frac{1}{\alpha})}}. \tag{5.18}$$

Using (5.16) as in [23, 18], we can take the derivative of (5.10) to get

$$\begin{aligned} s\frac{d}{dt}P(w,p,w^0)(t) &= \partial_\tau e(t-t_0,\cdot - st_0) * w^0 \\ &+ \int_0^t (I-\mathcal{T}^{-1})\partial_\tau e(t-t_0,\cdot - st_0) * (Q(w) - p'N(w))(t_0)\,dt_0. \end{aligned} \tag{5.19}$$

The most difficult step is to show that

$$\mathcal{N} : X \times \{w^0 \in \mathcal{L}^1(\mathbb{Z}),\ ||w^0||_1 \le \delta\} \to X. \tag{5.20}$$

Note that for $(w,p)\in X$, the "quadratic term" verifies the estimates

$$||Q(w)(t_0)||_1 + |p'(t)|\,||N(w)||_1 \le \frac{C}{(1+t_0)^{\frac{1}{2}}} R^2\delta^2, \tag{5.21}$$

$$||Q(w)(t_0)||_\infty + |p'(t)|\,||N(w)||_\infty \le \frac{C}{(1+t_0)} R^2\delta^2. \tag{5.22}$$

Using (5.11), basic convolution estimates, and (5.12), (5.13), (5.21), (5.22) we get

$$\begin{aligned} &||W(w,p,w^0)(t)||_\infty \le \frac{C\delta}{\sqrt{t}} \\ &+ \int_0^{\frac{t}{2}} \sup_l ||(I-\mathcal{T}^{-1})\widetilde{G}(t-t_0,\cdot - st, l - st_0)||_\infty ||(Q(w) - p'N(w))(t_0)||_1\,dt_0 \\ &+ \int_{\frac{t}{2}}^t \sup_l ||(I-\mathcal{T}^{-1})\widetilde{G}(t-t_0,\cdot - st, l - st_0)||_1 ||(Q(w) - p'N(w))(t_0)||_\infty\,dt_0 \\ &\le \frac{C\delta}{\sqrt{t}} + CR^2\delta^2 \int_0^{\frac{t}{2}} \frac{1}{t-t_0}\frac{1}{\sqrt{t_0+1}}\,dt_0 + CR^2\delta^2 \int_{\frac{t}{2}}^t \frac{1}{\sqrt{t-t_0}}\frac{1}{t_0}\,dt_0 \\ &\le \frac{C(\delta + R^2\delta^2)}{\sqrt{t}} \end{aligned} \tag{5.23}$$

and

$$
\begin{aligned}
&||W(w,p,w^0)(t)||_1 \le C\delta \\
&+ \int_0^t \sup_l ||(I-\mathcal{T}^{-1})\widetilde{G}(t-t_0,\cdot - st, l-st_0)||_1 ||(Q(w)-p'N(w))(t_0)||_1 \, dt_0 \\
&\le C\delta + CR^2\delta^2 \int_0^t \frac{1}{\sqrt{t-t_0}} \frac{1}{\sqrt{t_0+1}} \, dt_0 \\
&\le C(\delta + R^2\delta^2). \qquad (5.24)
\end{aligned}
$$

The same kind of computations with, moreover, the use of (5.17), (5.18) yields

$$|P(w,p,w^0)(t)| \le C(\delta + R^2\delta^2), \qquad (5.25)$$

$$\Big|\frac{d}{dt}P(w,p,w^0)(t)\Big| \le \frac{C(\delta + R^2\delta^2)}{\sqrt{t}}. \qquad (5.26)$$

Consequently, we get (5.20) by collecting (5.24), (5.23), (5.25), (5.26) and by choosing δ sufficiently small and R sufficiently large such that

$$C(\delta + \delta^2 R^2) \le \delta R.$$

A little variation in the previous computations gives

$$||\mathcal{N}(w_1,p_1,w^0) - \mathcal{N}(w_2,p_2,w^0)||_X \le CR\delta ||(w_1,p_1)-(w_2,p_2)||_X.$$

Hence it suffices to choose δ, R such that $C\delta R < 1$ to get the existence of a fixed point in X for $\mathcal{N}$. Consequently, we have shown (5.3), (5.4) and coming back to (5.5) we also have

$$||v(t+p(t)) - u(t)||_{\mathcal{L}^\alpha} \le \frac{C\delta}{(1+t)^{\frac{1}{2}\left(1-\frac{1}{\alpha}\right)}} \quad \forall t \ge 0, \quad \forall m \ge 1. \qquad (5.27)$$

To get the final form (5.2) in Theorem 5.1, we notice that $t \mapsto \theta = t + p(t)$ is a diffeomorphism from $\mathbb{R}_+$ to $\mathbb{R}_+$, thanks to the choice $C\delta R < 1$ since $|p'(t)| \le CR\delta$ and $p(0) = 0$. Consequently, let us define

$$\widetilde{p}(\theta) = t(\theta) - \theta;$$

we get

$$|\theta - t| \le CR\delta,$$

$$|\widetilde{p}(\theta)| \le CR\delta, \; |\widetilde{p}'(\theta)| \le \frac{CR\delta}{(1-CR\delta)^{\frac{3}{2}}\sqrt{1+t}} \quad \forall \theta \ge 0,$$

and

$$||v(\theta) - u(\theta + \widetilde{p}(\theta))||_{\mathcal{L}^\alpha} \le \frac{\widetilde{C}\delta}{(1+\theta)^{\frac{1}{2}\left(1-\frac{1}{\alpha}\right)}} \quad \forall \theta \ge 0, \quad \forall \alpha \ge 1. \qquad (5.28)$$

This ends the proof of Theorem 5.1. ∎

REFERENCES

[1] S. Benzoni-Gavage, *Semi-discrete shock profiles for hyperbolic systems of conservation laws*, Phys. D, 115 (1998), pp. 109–123.

[2] S. Benzoni-Gavage, *Stability of semi-discrete shock profiles by means of an Evans function in infinite dimensions*, J. Dynam. Differential Equations, 14 (2002), pp. 613–674.

[3] S. Benzoni-Gavage and P. Huot, *Existence of semi-discrete shocks*, Discrete Contin. Dynam. Systems, 8 (2002), pp. 163–190.

[4] S. Benzoni-Gavage, D. Serre, and K. Zumbrun, *Alternate Evans functions and viscous shock waves*, SIAM J. Math. Anal., 32 (2001), pp. 929–962.

[5] S. Bianchini, *BV solutions of the semidiscrete upwind scheme*, Arch. Ration. Mech. Anal., 167 (2003), pp. 1–81.

[6] S.-N. Chow, J. Mallet-Paret, and W. Shen, *Traveling waves in lattice dynamical systems*, J. Differential Equations, 149 (1998), pp. 248–291.

[7] R. A. Gardner and C. K. R. T. Jones, *Traveling waves of a perturbed diffusion equation arising in a phase field model*, Indiana Univ. Math. J., 39 (1990), pp. 1197–1222.

[8] R. A. Gardner and K. Zumbrun, *The gap lemma and geometric criteria for instability of viscous shock profiles*, Comm. Pure Appl. Math., 51 (1998), pp. 797–855.

[9] J. Goodman, *Nonlinear asymptotic stability of viscous shock profiles for conservation laws*, Arch. Rational Mech. Anal., 95 (1986), pp. 325–344.

[10] J. Härterich, B. Sandstede, and A. Scheel, *Exponential dichotomies for linear non-autonomous functional differential equations of mixed type*, Indiana Univ. Math. J., 51 (2002), pp. 1081–1109.

[11] P. D. Lax, *Hyperbolic systems of conservation laws.* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[12] X.-B. Lin, *Exponential dichotomies and homoclinic orbits in functional-differential equations*, J. Differential Equations, 63 (1986), pp. 227–254.

[13] J.-G. Liu and Z. P. Xin, *Nonlinear stability of discrete shocks for systems of conservation laws*, Arch. Rational Mech. Anal., 125 (1993), pp. 217–256.

[14] T.-P. Liu and S.-H. Yu, *Continuum shock profiles for discrete conservation laws.* I. *Construction*, Comm. Pure Appl. Math., 52 (1999), pp. 85–127.

[15] A. Majda and J. Ralston, *Discrete shock profiles for systems of conservation laws*, Comm. Pure Appl. Math., 32 (1979), pp. 445–482.

[16] J. Mallet-Paret, *The Fredholm alternative for functional differential equations of mixed type*, J. Dynam. Differential Equations, 11 (1999), pp. 1–47.

[17] J. Mallet-Paret and S. Verduyn Lunel, *Exponential dichotomies and Wiener-Hopf factorizations for mixed-type functional differential equations*, J. Differential Equations, to appear.

[18] C. Mascia and K. Zumbrun, *Pointwise Green's function bounds and stability of relaxation shocks*, Indiana Univ. Math. J., 51 (2002), pp. 773–904.

[19] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[20] F. Rousset, *Viscous approximation of strong shocks of systems of conservation laws*, SIAM J. Math. Anal., 35 (2003), pp. 492–519.

[21] A. Rustichini, *Functional-differential equations of mixed type: The linear autonomous case*, J. Dynam. Differential Equations, 1 (1989), pp. 121–143.

[22] M. Schechter, *Spectra of Partial Differential Operators*, North-Holland, Amsterdam, 1971.

[23] K. Zumbrun, *Refined wave-tracking and nonlinear stability of viscous Lax shocks*, Methods Appl. Anal., 7 (2000), pp. 747–768.

[24] K. Zumbrun and P. Howard, *Pointwise semigroup methods and stability of viscous shock waves*, Indiana Univ. Math. J., 47 (1998), pp. 741–871.

SIAM J. MATH. ANAL.
Vol. 35, No. 3, pp. 708–758

NONLINEAR APPROXIMATION FROM DIFFERENTIABLE PIECEWISE POLYNOMIALS*

OLEG DAVYDOV† AND PENCHO PETRUSHEV‡

Abstract. We study nonlinear n-term approximation in $L_p(\mathbb{R}^2)$ $(0 < p \le \infty)$ from hierarchical sequences of stable local bases consisting of differentiable (i.e., C^r with $r \ge 1$) piecewise polynomials (splines). We construct such sequences of bases over multilevel nested triangulations of $\mathbb{R}^2$, which allow arbitrarily sharp angles. To quantize nonlinear n-term spline approximation, we introduce and explore a collection of smoothness spaces (B-spaces). We utilize the B-spaces to prove companion Jackson and Bernstein estimates and then characterize the rates of approximation by interpolation. Even when applied on uniform triangulations with well-known families of basis functions such as box splines, these results give a more complete characterization of the approximation rates than the existing ones involving Besov spaces. Our results can easily be extended to properly defined multilevel triangulations in $\mathbb{R}^d$, $d > 2$.

Key words. nonlinear approximation, Jackson and Bernstein estimates, multivariate splines, multilevel nested triangulations, multilevel bases, stable local spline bases

AMS subject classifications. 41A15, 41A25, 41A63, 65D07, 65D17

DOI. 10.1137/S0036141002409374

1. Introduction.

Nonlinear approximation of functions in dimensions $d > 1$ is a challenging area, especially if one moves away from tensor product-type approaches in order to more adequately approximate functions with singularities along curves and with other anisotropies. One of the most natural tools for approximation is piecewise polynomials over triangulations, and a fundamental problem is to characterize the rate of nonlinear approximation in L_p $(0 < p \le \infty)$ in terms of properly defined global smoothness conditions. This problem is disheartening if one allows the nonlinear approximation to be from any piecewise polynomial over an arbitrary triangulation. The difficulty stems from the highly nonlinear nature of piecewise polynomials in dimensions $d > 1$. For instance, if s_1 and s_2 are two piecewise polynomials over n triangles in $\mathbb{R}^2$ each, then $s_1 + s_2$ is in general a piecewise polynomial over many more than n (even $> n^2$) pieces. Therefore, the well-known recipe of proving Jackson and Bernstein estimates and then applying interpolation is useless.

The problem becomes even harder when *differentiable* piecewise polynomials are needed, which, for instance, is the case for numerous practical applications of surface modeling and for the conforming finite element methods for higher order PDEs. Moreover, there is an intrinsic demand for differentiability of the approximating tools from the point of view of the nonlinear approximation theory itself. Indeed, this property, together with local reproduction of higher degree polynomials, is crucial for the ability to represent higher order smoothness spaces, such as classical Sobolev or Besov spaces in regular settings (see Theorem 2.15). The desirable differentiability of the approximating piecewise polynomials, however, leads to additional difficulties

*Received by the editors June 9, 2002; accepted for publication (in revised form) March 14, 2003; published electronically October 2, 2003.
http://www.siam.org/journals/sima/35-3/40937.html
†Mathematisches Institut, Justus-Liebig-Universität Giessen, D-35392 Giessen, Germany (oleg.davydov@math.uni-giessen.de, http://www.uni-giessen.de/∼gcn5/davydov/).
‡Department of Mathematics, University of South Carolina, Columbia, SC 29208, (pencho@math.sc.edu, http://www.math.sc.edu/∼pencho). This author was supported by National Science Foundation grant DMS-0200665.

because of the complicated structure of spaces of multivariate splines. For example, the dimension is not known and stable local bases are impossible in general already for the space of all piecewise polynomials of degree $< k$ and smoothness $r \geq 1$ with respect to a finite triangulation of a polygonal domain in $\mathbb{R}^2$ if $k \leq 3r + 2$ [10].

A reasonable alternative to "spline approximation with free triangulations" is to consider nonlinear n-term approximation from hierarchical sequences of spline bases over *multilevel nested triangulations* of $\mathbb{R}^d$. (For the sake of simplicity, we shall restrict ourselves in this article to the case $d = 2$.) To explain this concept more precisely, consider a sequence $(\mathcal{T}_m)_{m\in\mathbb{Z}}$ of partitions of $\mathbb{R}^2$ into triangles with disjoint interiors such that each *level* $\mathcal{T}_m$ is a refinement of the previous one $\mathcal{T}_{m-1}$. Let $\mathcal{T} := \bigcup_{m\in\mathbb{Z}} \mathcal{T}_m$. We impose certain mild (and natural) conditions on the triangulations which prevent them from deterioration but still allow the triangles to change in size, shape, and orientation quickly when moving around at a given level or through the levels. In particular, triangles with arbitrarily sharp angles may occur at any location. We denote by $\mathcal{S}^{k,r}(\mathcal{T}_m)$ the set of all r-times differentiable piecewise polynomials with respect to $\mathcal{T}_m$ of degree $< k$. Given a ladder of spaces

$$\cdots \subset \mathcal{S}_{-1} \subset \mathcal{S}_0 \subset \mathcal{S}_1 \subset \cdots, \qquad \mathcal{S}_m \subset \mathcal{S}^{k,r}(\mathcal{T}_m), \tag{1.1}$$

and bases Φ_m of $\mathcal{S}_m$, $m \in \mathbb{Z}$, we set $\Phi := \Phi_{\mathcal{T}} := \bigcup_{m\in\mathbb{Z}} \Phi_m$. Using the standard wavelet terminology, we can describe such a nested sequence of spaces with bases as *"spline multiresolution"* (or *"multiresolution analysis"*).

Consider now the problem for nonlinear (n-term) approximation from the set Σ_n of all piecewise polynomials of the form $s = \sum_{j=1}^{n} c_j\varphi_j$, where $\varphi_j \in \Phi$ may come from different levels and locations. Once a particular multilevel triangulation has been selected, the variety of piecewise polynomial approximations significantly reduces. However, a great deal of flexibility is retained, and the problem remains highly nonlinear. For instance, thin and elongated basis functions are allowed. On the other hand, the advantages of multilevel approximation methods can be exploited in full.

Our program consists of the following basic steps:

1. We construct hierarchical sequences of bases $(\Phi_m)_{m\in\mathbb{Z}}$ on multilevel triangulations satisfying certain requirements of *local regularity* allowing anisotropically shaped triangles.

2. To quantify the approximation process, we introduce and develop a family (library) of smoothness spaces $B^\alpha_\tau(\Phi_{\mathcal{T}})$ depending on $\Phi_{\mathcal{T}}$ and as a consequence on the triangulation $\mathcal{T}$. We call them B-spaces since they have some resemblance to Besov spaces. So, the idea is to measure the smoothness of the functions using a family of space scales $B^\alpha_\tau(\Phi_{\mathcal{T}})$ (which vary with $\Phi_{\mathcal{T}}$) instead of a single scale of smoothness spaces like the scale of Besov spaces.

3. We develop a coherent theory of nonlinear n-term approximation from $\Phi_{\mathcal{T}}$ based on the idea of proving Jackson and Bernstein estimates and interpolation.

4. We utilize this theory in the development of algorithms for nonlinear piecewise polynomial (spline) approximation which capture the rate of the best approximation.

The logic of the resulting *approximation scheme* is the following: Suppose $\{\Phi_{\mathcal{T}}\}_{\mathcal{T}}$ is a collection of multilevel sequences of (spline) bases as above.

(i) For a given function f, find the "right" triangulation $\mathcal{T} := \mathcal{T}_f$ such that f exhibits the most smoothness (sparsity of its representation) when measured via the scale $B^\alpha(\Phi_{\mathcal{T}})$.

(ii) Find an optimal or near optimal representation of f using $\Phi_{\mathcal{T}}$. (Note that $\Phi_{\mathcal{T}}$ is redundant, i.e., linearly dependent.)

(iii) Using this representation, run an algorithm for n-term L_p-approximation which achieves the rate of the best n-term approximation.

Naturally, the first step presents the most challenging problem in this scheme. We do not have a completely satisfactory algorithm for this step. (Note that this problem has a complete and efficient solution in the simpler case of nonlinear approximation from piecewise polynomials over dyadic partitions; see [54].) As it will be shown in this article, the other steps are now well understood and have complete solutions.

The above program has been suggested in [38] and implemented in [38, 39] in the cases of approximation from discontinuous piecewise polynomials and continuous piecewise linear functions ($r = -1$, $k \geq 1$, and $r = 0$, $k = 2$, where $r = -1$ corresponds to the discontinuous case). The simplest example of a hierarchical family of continuous basis functions is the set of all Courant elements generated by a multilevel nested triangulation $\mathcal{T}$, that is, the set of all piecewise linear and continuous functions $\Phi_{\mathcal{T}} = \{\varphi_\theta\}$ supported on the cells $\{\theta\}$ (each θ is the union of all triangles of a particular level $\mathcal{T}_m$ attached to a vertex); see [38].

In the present article, we develop the theory of nonlinear n-term approximation for basis families consisting of differentiable piecewise polynomials ($r \geq 1$). The construction of such basis functions suitable for application is hampered by the fact that both the classical differentiable finite elements [14] and the earlier polynomial spline basis constructions on arbitrary triangulations [1, 8, 16, 17, 18, 35, 36, 44, 48, 57] are difficult to use for our purposes; see Remark 4.12 and the discussion in section 5.3. The stable local spline bases of [27] can in principle be used in two variables. However, all other arguments of our article are basically "dimension independent," and we refrain here from treating the case $d > 2$ only for the sake of simplicity and clarity. Therefore, we build upon the nodal spline bases of [22], which is the only known approach that produces stable local bases for nested spline spaces on general triangulations in all dimensions.

However, these bases are stable only for triangulations satisfying (in $\mathbb{R}^2$) the minimal angle condition. We extend the construction of [22] to a wider class of *strong locally regular* triangulations; see section 2 for a definition. Note that the new basis functions are invariant under affine transforms (see Remark 4.9). In the case $r = 0$ our construction reduces to the classical continuous Lagrange finite elements and is valid for any *locally regular* triangulation; see Remark 4.13.

A focal point of our development is the characterization of nonlinear n-term approximation from families of differentiable spline basis functions, including the development of B-spaces, proof of Jackson and Bernstein estimates, and characterization of the approximation spaces by interpolation (see sections 2–3). In [39], there are three algorithms developed for nonlinear n-term approximation in L_p ($0 < p \leq \infty$) from Courant elements. These can be immediately implemented for n-term approximation from differentiable spline bases, and it can be shown similarly as in [39] that they achieve the rate of the best approximation. We do not pursue this goal here.

The B-spaces from the present article can be viewed as a generalization of the "approximation spaces" from section 3.4 of [51] (see also the references therein). More precisely, in the specific setting of "quasi-uniform partitions" and the basis functions used in [51], our B-spaces coincide with the approximation spaces of [51].

The theory of nonlinear n-term approximation from box splines (on uniform triangulations) has been developed in [29] ($p < \infty$) and [30] ($p = \infty$) (for nonlinear

spline approximation in dimension $d = 1$, see [53]). In these articles, direct, inverse, and characterization theorems have been proved utilizing certain Besov spaces. Even in this case, our results, which utilize B-spaces (in place of Besov spaces), are more complete since they characterize nonlinear n-term box spline approximation for all rates of approximation, while in the above-mentioned articles the rate is restricted by the Besov smoothness of the box splines.

There is an apparent connection between our developments here and multilevel finite element methods for PDEs; see, e.g., [51]. Therefore, it seems an interesting task to develop finite element algorithms for solving PDEs which achieve the rate of the best n-term approximation of the solution.

The outline of the article is the following. In section 2, we introduce and develop the B-spaces needed for the characterization of nonlinear approximation for any family of basis functions with certain properties. In section 3, we develop the general theory of nonlinear n-term approximation from piecewise polynomials, where the global smoothness of functions is measured by means of our B-spaces. In section 4, we construct hierarchical sequences of bases consisting of differentiable piecewise polynomials. In section 5, we review a number of alternative constructions fitting into our scheme, based on box splines and some other spline bases on special triangulations. The final section A is an appendix containing some of the proofs.

Throughout the article, we use the following notation: $L_\infty^{\rm loc}(\mathbb{R}^2) := C(\mathbb{R}^2)$ and $L_\infty(\mathbb{R}^2) := C_0(\mathbb{R}^2) := \{f \in C(\mathbb{R}^2) : \lim_{x\to\infty} f(x) = 0\}$, $L_q^{\rm loc} := L_q^{\rm loc}(\mathbb{R}^2)$, $0 < q \le \infty$, $C := C(\mathbb{R}^2)$, $\|\cdot\|_q := \|\cdot\|_{L_q(\mathbb{R}^2)}$, $0 < q \le \infty$; Π_k denotes the set of all algebraic polynomials in two variables of total degree $< k$. For any $\Omega \subset \mathbb{R}^2$, $\mathbb{1}_\Omega$ denotes the characteristic function of Ω and $|\Omega|$ denotes the Lebesgue measure of Ω. Positive constants are denoted by $c, c_1, \dots$ (they may vary at every occurrence), $\alpha \approx \beta$ means $c_1\alpha \le \beta \le c_2\alpha$, and $\alpha := \beta$ or $\beta =: \alpha$ stands for "α is by definition equal to β."

2. B-spaces generated by spline multiresolution. In the present section, we introduce and explore the smoothness spaces we need for the characterization of nonlinear n-term spline approximation generated by families of differentiable basis functions over multilevel nested triangulations.

2.1. Triangulations. In our development, we utilize three types of multilevel nested triangulations. We shall call each of them simply a *triangulation*, although such a triangulation does not form a single partition of $\mathbb{R}^2$ but rather an infinite nested family of partitions (each of them is a triangulation of $\mathbb{R}^2$ in the more commonly used sense).

Let $\mathcal{T} = \bigcup_{m\in\mathbb{Z}} \mathcal{T}_m$ be a set of closed triangles in $\mathbb{R}^2$ with *levels* $\mathcal{T}_m$, $m \in \mathbb{Z}$. Denote by $\mathcal{V}_m$ the set of all vertices (nodal points) of triangles from $\mathcal{T}_m$ and set $\mathcal{V} := \bigcup_{m\in\mathbb{Z}} \mathcal{V}_m$. We say that $\mathcal{T}$ is a *triangulation* of $\mathbb{R}^2$ if the following conditions are fulfilled:

(a) Every level $\mathcal{T}_m$ is a set of triangles with disjoint interiors which cover $\mathbb{R}^2$: $\mathbb{R}^2 = \bigcup_{\triangle\in\mathcal{T}_m} \triangle$.

(b) The levels $(\mathcal{T}_m)_{m\in\mathbb{Z}}$ of $\mathcal{T}$ are *nested*; i.e., $\mathcal{T}_{m+1}$ is a refinement of $\mathcal{T}_m$ obtained by splitting each $\triangle \in \mathcal{T}_m$ into subtriangles with disjoint interiors called *children* of $\triangle$.

(c) Each triangle $\triangle \in \mathcal{T}_m$ has at least two and at most M_0 children in $\mathcal{T}_{m+1}$, where $M_0 \ge 2$ is a constant independent of m.

(d) *No hanging vertices condition:* No vertex of any triangle $\triangle \in \mathcal{T}_m$ lies in the interior of an edge of another triangle from $\mathcal{T}_m$.

(e) The *valence* N_v of each vertex $v \in \mathcal{V}_m$ (the number of triangles $\triangle \in \mathcal{T}_m$ which share v as a vertex) is $\le N_0$, where N_0 is a constant.

(f) For any compact $K \subset \mathbb{R}^2$ and any fixed $m \in \mathbb{Z}$, there is a finite collection of triangles from $\mathcal{T}_m$ which cover K.

Note that any two triangles in $\mathcal{T}$ either have disjoint interiors or one of them contains the other. In particular, $\triangle' \in \mathcal{T}_{m+1}$ is a child of $\triangle \in \mathcal{T}_m$ ($m \in \mathbb{Z}$) if and only if $\triangle' \subset \triangle$. If $\triangle$ and $\triangle'$ are two different triangles in $\mathcal{T}$ and $\triangle' \subset \triangle$, then we say that $\triangle$ is an *ancestor* of $\triangle'$, while $\triangle'$ is a *descendant* of $\triangle$.

Locally regular triangulations. We call a triangulation $\mathcal{T} = \bigcup_{m\in\mathbb{Z}} \mathcal{T}_m$ a *locally regular triangulation* of $\mathbb{R}^2$, or briefly an *LR-triangulation*, if $\mathcal{T}$ satisfies the following additional conditions:

(g) There exists a constant $1/2 \le \rho < 1$ such that for each $\triangle \in \mathcal{T}$ and any child $\triangle' \in \mathcal{T}$ of $\triangle$,

$$(1-\rho)|\triangle| \le |\triangle'| \le \rho|\triangle|. \tag{2.1}$$

(h1) There exists a constant $0 < \delta_1 \le 1$ independent of m such that for any $\triangle', \triangle'' \in \mathcal{T}_m$ ($m \in \mathbb{Z}$) with a common edge,

$$\delta_1 \le |\triangle'|/|\triangle''| \le \delta_1^{-1}. \tag{2.2}$$

By (e), it follows that for any $\triangle', \triangle'' \in \mathcal{T}_m$ with at least one common vertex, (2.2) holds with δ_1 replaced by $\delta_1^{N_0/2}$.

Strong locally regular triangulations. We call a triangulation $\mathcal{T} = \bigcup_{m\in\mathbb{Z}} \mathcal{T}_m$ a *strong locally regular triangulation* of $\mathbb{R}^2$, or briefly an *SLR-triangulation*, if $\mathcal{T}$ satisfies (2.1) and the following condition that replaces (2.2):

(h2) There exists a constant $0 < \delta_2 \le 1/2$ such that for any $\triangle', \triangle'' \in \mathcal{T}_m$ ($m \in \mathbb{Z}$) sharing an edge,

$$|\mathrm{conv}\,(\triangle' \cup \triangle'')|/|\triangle'| \le \delta_2^{-1}, \tag{2.3}$$

where $\mathrm{conv}\,(G)$ denotes the convex hull of $G \subset \mathbb{R}^2$.

Obviously, (2.3) implies (2.2) with $\delta_1 = \delta_2$. Therefore, each SLR-triangulation is an LR-triangulation.

Regular triangulations. By definition, a triangulation $\mathcal{T} = \bigcup_{m\in\mathbb{Z}} \mathcal{T}_m$ is called a *regular triangulation* if $\mathcal{T}$ satisfies the following condition:

(h3) There exists a constant $\beta = \beta(\mathcal{T}) > 0$ such that the minimal angle of each triangle $\triangle \in \mathcal{T}$ is $\ge \beta$.

Next, we make a few remarks which will help understand better the nature of the triangulations that we utilize.

(i) For each of the three types of triangulations there is a number of constants that are assumed fixed. In what follows we refer to them as *parameters*. Thus the parameters of an SLR-triangulation are M_0, N_0, ρ, and δ_2. Notice that because of (2.1), we can set $M_0 := 1/(1-\rho)$ and remove M_0 from the list of parameters. However, this would tend to obscure the actual role of ρ and M_0.

(ii) It is a key observation that the collection of all SLR-triangulations with given (fixed) parameters is invariant under affine transforms. The same is true for LR-triangulations.

(iii) It is easy to see that (2.3) is equivalent to the following condition introduced in [38]. *Affine transform angle condition:* There exists a constant $\beta = \beta(\mathcal{T})$, $0 < \beta \le \pi/3$, such that if $\triangle_0 \in \mathcal{T}_m$, $m \in \mathbb{Z}$, and $A : \mathbb{R}^2 \to \mathbb{R}^2$ is an affine transform that maps

$\triangle_0$ one-to-one onto an equilateral reference triangle, then for every $\triangle \in \mathcal{T}_m$ which has at least one common vertex with $\triangle_0$, we have

$$\min \text{angle}\,(A(\triangle)) \geq \beta, \tag{2.4}$$

where $A(\triangle)$ is the image of $\triangle$ by the affine transform A. (The equivalence of the two conditions follows easily from the obvious but important fact that both conditions are invariant under affine transforms. Note that we prefer to use (2.3) rather than (2.4) in the definition of SLR-triangulations in this article since the constant δ_2 appears naturally when estimating norms of the basis functions constructed in section 4 (see (4.8)) and also (2.3) is easier to verify in practical situations.)

(iv) As we have already mentioned, every SLR-triangulation is an LR-triangulation, but the converse statement is not true. Also, every regular triangulation is an SLR-triangulation but not the other way around. Counterexamples are given in [38].

(v) The *maximal angle (MA) condition*

$$\pi - \max \text{angle}\,(\triangle) \geq \beta > 0, \qquad \triangle \in \mathcal{T}, \tag{2.5}$$

known from the finite element method [2] is totally different from our conditions of regularity. It is easy to see that there are SLR-triangulations that do not satisfy MA, and there are triangulations that satisfy the MA and fail to be locally regular. As we shall see below (Example 4.7), our construction of stable differentiable basis functions does not extend to triangulations satisfying the MA condition but failing to be SLR.

(vi) The rate of change of the size of the elements ($|\triangle|$, $\min \text{angle}\,(\triangle)$, and $\text{diam}(\triangle)$) of a triangle $\triangle \in \mathcal{T}$ as $\triangle$ moves away from a fixed triangle $\triangle^\diamond \in \mathcal{T}$ for different types of triangulations $\mathcal{T}$ is explored in [38]. We shall briefly discuss this issue for SLR-triangulations, which are the most important type of triangulations for the present article. An SLR-triangulation $\mathcal{T}$ may have an equilateral (or close to such) triangle $\triangle^\diamond$ at any level T_m with descendants $\triangle_1 \supset \triangle_2 \supset \cdots$ such that $\min \text{angle}\,(\triangle_j) \to 0$ as $j \to \infty$, and also a sequence $(\triangle'_j)_{j=0}^{\infty} \subset \mathcal{T}_m$ with $\triangle'_0 = \triangle$ and $\triangle'_j \cap \triangle'_{j+1} \neq \emptyset$ $(j = 0, 1, \ldots)$ such that $\min \text{angle}\,(\triangle'_j) \to 0$. Conditions (2.1) and (2.3) suggest geometric rates of change of $|\triangle|$, $\min \text{angle}\,(\triangle)$, and $\text{diam}(\triangle)$ as $\triangle \in \mathcal{T}_m$ moves away from a fixed $\triangle^\diamond \in \mathcal{T}_m$. In fact, the rate of change is a power of the minimal number of edges connecting $\triangle$ and $\triangle^\diamond$; see [38].

(vii) We shall need to know what happens with the levels $\mathcal{T}_m$ of a triangulation $\mathcal{T}$ as $m \to -\infty$. By Lemma 2.1 from [38], for each LR-triangulation $\mathcal{T}$ there exists a finite cover $\mathcal{T}_{-\infty}$ of $\mathbb{R}^2$ such that either $\mathcal{T}_{-\infty} = \{\mathbb{R}^2\}$ or $\mathcal{T}_{-\infty} = (\triangle^j_\infty)_{j=1}^{N_\infty}$, $N_\infty \leq N_0$, where each $\triangle^j_\infty$ is an *infinite triangle*, i.e., the set of all points on and between two rays which are not collinear and have a common beginning. Moreover, in the second case, the infinite triangles $(\triangle^j_\infty)_{j=1}^{N_\infty}$ have a single common vertex and disjoint interiors, and also each triangle $\triangle \in \mathcal{T}$ and all its ancestors are contained in an infinite triangle $\triangle^j_\infty \in \mathcal{T}_{-\infty}$.

For more details about multilevel triangulations, see [38].

Some additional notation and preliminaries. We denote by $[v_1, v_2]$ the interval (straight line segment) with endpoints v_1, v_2 and by $|e|$ the length of $e = [v_1, v_2]$. Furthermore, we let $[v_1, v_2, v_3]$ denote the triangle with vertices v_1, v_2, v_3, and let $|\triangle|$ denote the area of $\triangle = [v_1, v_2, v_3]$. Throughout the article, we assume that the vertices v_1, v_2, v_3 of any triangle $[v_1, v_2, v_3]$ are ordered counterclockwise.

For a triangle $\triangle \in \mathcal{T}_m$ $(m \in \mathbb{Z})$, we define $\text{level}(\triangle) := m$.

For any vertex $v \in \mathcal{V}_m$, we let $\text{star}\,(v) = \text{star}^{\,1}(v)$ denote the *star* of v, i.e., the union of all triangles $\triangle \in \mathcal{T}_m$ attached to v. Moreover, for each $\ell \geq 2$, we denote

by $\operatorname{star}^\ell(v)$ the union of $\operatorname{star}^{\ell-1}(v)$ and the stars of the vertices of $\operatorname{star}^{\ell-1}(v)$. (Note that $\operatorname{star}^\ell(v)$ depends also on the level m, but we do not indicate this in the notation since it is always clear from the context what level is meant.) We also set

$$\Omega^\ell_\triangle := \cup\{\operatorname{star}^\ell(v) : v \in \mathcal{V}_m,\ \triangle \subset \operatorname{star}^\ell(v)\}, \quad \triangle \in \mathcal{T}_m. \tag{2.6}$$

It is easy to check that $\Omega^\ell_\triangle := \cup\{\operatorname{star}^{2\ell-1}(v) : v \text{ is a vertex of } \triangle\}$, $\triangle \in \mathcal{T}_m$.

It is readily seen that there exists a constant $c^\star = c^\star(N_0, \ell) \le N_0^\ell$ such that

$$\#\{\triangle \in \mathcal{T}_m : \triangle \subset \operatorname{star}^\ell(v)\} \le c^\star, \quad v \in \mathcal{V}_m, \tag{2.7}$$

and hence there exists a constant $c^{\star\star} = c^{\star\star}(N_0, \ell) \le 3c^\star(N_0, 2\ell-1) - 5 \le 3N_0^{2\ell-1}$ such that

$$\#\{\triangle \in \mathcal{T}_m : \triangle \subset \Omega^\ell_{\triangle'}\} \le c^{\star\star}, \quad \triangle' \in \mathcal{T}_m. \tag{2.8}$$

We denote by $\mathcal{E}_m$ the set of all edges of triangles of $\mathcal{T}_m$ and set $\mathcal{E} := \bigcup_{m\in\mathbb{Z}} \mathcal{E}_m$. We let $\operatorname{star}(e)$ denote the union of the two triangles attached to $e \in \mathcal{E}_m$.

For future use, we state the following inequality:

$$\sum_{\triangle\in\mathcal{T},\, \triangle\supset\triangle'} (|\triangle'|/|\triangle|)^\gamma \le \sum_{j=0}^\infty \rho^{j\gamma} = c(\rho,\gamma) < \infty, \quad \triangle' \in \mathcal{T},\ \gamma > 0, \tag{2.9}$$

which is immediate from the properties of LR-triangulations ($|\triangle'| \le \rho|\triangle|$ if $\triangle'$ is a child of $\triangle$).

2.2. Basis functions: The general setting. Let $\mathcal{T} = \bigcup_{m\in\mathbb{Z}} \mathcal{T}_m$ be a locally regular (or better) triangulation. For $m \in \mathbb{Z}$, $r \ge 0$, and $k \ge 1$, we denote by $\mathcal{S}_m^{k,r} = \mathcal{S}^{k,r}(\mathcal{T}_m)$ the set of all r times differentiable piecewise polynomial functions of degree $< k$ over $\mathcal{T}_m$; i.e., $s \in \mathcal{S}_m^{k,r}$ if and only if $s \in C^r(\mathbb{R}^2)$ and $s = \sum_{\triangle\in\mathcal{T}_m} \mathbb{1}_\triangle \cdot P_\triangle$ with $P_\triangle \in \Pi_k$. Naturally, $\mathcal{S}_m^{k,-1}$ will denote the set of all piecewise polynomials of degree $< k$ over $\mathcal{T}_m$ which are, in general, discontinuous across the edges from $\mathcal{E}_m$.

We assume that for each $m \in \mathbb{Z}$ there is a subspace $\mathcal{S}_m$ of $\mathcal{S}_m^{k,r}$ ($r \ge 0$, $k \ge 2$) and a family $\Phi_m = \{\varphi_\theta : \theta \in \Theta_m\} \subset \mathcal{S}_m$ of *basis functions* satisfying the following conditions:

1. $\Pi_{\tilde{k}} \subset \mathcal{S}_m$ for some $1 \le \tilde{k} \le k$ ($\tilde{k}$ independent of m).
2. $\mathcal{S}_m \subset \mathcal{S}_{m+1}$ ($m \in \mathbb{Z}$).
3. For any $s \in \mathcal{S}_m$ there exists a unique sequence of real coefficients $a(s) = (a_\theta(s))_{\theta\in\Theta_m}$ such that

$$s = \sum_{\theta\in\Theta_m} a_\theta(s)\varphi_\theta.$$

(Thus, Φ_m is a basis for $\mathcal{S}_m$ and $(a_\theta(\cdot))_{\theta\in\Theta_m}$ are the dual functionals.)

4. For each $\theta \in \Theta_m$ there is a vertex $v = v_\theta \in \mathcal{V}_m$ such that

$$\operatorname{supp}\varphi_\theta \subset \operatorname{star}^\ell(v) =: E_\theta, \tag{2.10}$$

$$\|\varphi_\theta\|_{L_\infty(\mathbb{R}^2)} = \|\varphi_\theta\|_{L_\infty(E_\theta)} \le M_1, \tag{2.11}$$

$$|a_\theta(s)| \le M_2\|s\|_{L_\infty(E_\theta)}, \qquad s \in \mathcal{S}_m, \tag{2.12}$$

where $\ell \ge 1$ and M_1, M_2 are positive constants, all independent of θ and m.

Let

$$\Phi := \bigcup_{m\in\mathbb{Z}} \Phi_m \quad \text{and} \quad \Theta := \bigcup_{m\in\mathbb{Z}} \Theta_m.$$

We shall refer to r, k, $\tilde{k}$, ℓ, M_1, and M_2 as *parameters* of Φ.

A simple example of a family of basis functions satisfying the above conditions is the set of well-known Courant elements (continuous piecewise linear basis functions, $r = 0$, $k = 2$) associated with $\mathcal{T}$ (see [38]). Concrete constructions of differentiable basis functions ($r \geq 1$) will be discussed below in sections 4–5.

Although Θ and Θ_m ($m \in \mathbb{Z}$) are simply index sets, in the case of Courant elements, Θ can be identified as the set of all cells (supports of basis functions). As we shall see in sections 4–5, in general, several basis functions of Φ_m may have the same support. However, the supports of only $\leq$ *constant* of them may overlap.

LEMMA 2.1. *There is a constant L depending only on k, ℓ, and N_0 such that for any $\triangle \in \mathcal{T}_m$ ($m \in \mathbb{Z}$),*

$$\#\{\theta \in \Theta_m \,:\, E_\theta \supset \triangle\} \leq L, \tag{2.13}$$

where E_θ is defined in (2.10).

Proof. We have by (2.10) and (2.8)

$$\begin{aligned}\#\{\theta \in \Theta_m \,:\, \triangle \subset E_\theta\} &\leq \dim \mathcal{S}_m^{k,r}|_{\Omega_\triangle^\ell} \leq \dim \mathcal{S}^{k,-1}(\mathcal{T}_m)|_{\Omega_\triangle^\ell} \\ &= \binom{k+1}{2} \#\{\triangle' \in \mathcal{T}_m \,:\, \triangle' \subset \Omega_\triangle^\ell\} \\ &\leq \binom{k+1}{2} c^{\star\star}. \qquad \square\end{aligned}$$

We shall frequently use the equivalence of different norms of polynomials as stated in the following lemma (see also [38]).

LEMMA 2.2. *Let $P \in \Pi_k$, $k \geq 1$, and $0 < p, q \leq \infty$.*

(a) *For any triangle $\triangle \subset \mathbb{R}^2$, $\|P\|_{L_p(\triangle)} \approx |\triangle|^{1/p-1/q}\|P\|_{L_q(\triangle)}$ with constants of equivalence depending only on p, q, and k.*

(b) *If $\triangle$ and $\triangle'$ are two triangles such that $\triangle' \subset \triangle$ and $|\triangle| \leq c_1|\triangle'|$, then $\|P\|_{L_p(\triangle)} \leq c\|P\|_{L_p(\triangle')}$ with $c = c(p, k, c_1)$.*

(c) *If $\triangle'$ and $\triangle$ are two triangles such that $\triangle' \subset \triangle$ and $|\triangle'| \leq c_2|\triangle|$ with $0 < c_2 < 1$, then $\|P\|_{L_p(\triangle)} \leq c\|P\|_{L_p(\triangle\setminus\triangle')} \approx |\triangle|^{1/p-1/q}\|P\|_{L_q(\triangle\setminus\triangle')}$ with constants depending only on p, q, k, and c_2.*

By (2.2) and (2.7), $|E_\theta| \approx |\triangle|$ if $\triangle \subset E_\theta$, $\triangle \in \mathcal{T}_m$, and $\theta \in \Theta_m$. Using this and Lemma 2.2, we obtain that, for $0 < p, q \leq \infty$,

$$\|s\|_{L_p(E_\theta)} \approx |E_\theta|^{1/p-1/q}\|s\|_{L_q(E_\theta)}, \qquad s \in \mathcal{S}_m, \quad \theta \in \Theta_m, \tag{2.14}$$

where the constants of equivalence depend on p, q, k, and δ_1. In particular, we shall need (2.14) with $s = \varphi_\theta$, when it takes the form $\|\varphi_\theta\|_p \approx |E_\theta|^{1/p-1/q}\|\varphi_\theta\|_q$, in view of (2.10).

LEMMA 2.3. *The bases Φ_m are L_q-*stable *for all $0 < q \leq \infty$. That is, if $g := \sum_{\theta\in\Theta_m} b_\theta\varphi_\theta$, where $(b_\theta)_{\theta\in\Theta_m}$ is an arbitrary sequence of real numbers, then*

$$\|g\|_q \approx \left(\sum_{\theta\in\Theta_m} \|b_\theta\varphi_\theta\|_q^q\right)^{1/q}.$$

Moreover, for any $\gamma \in \mathbb{R}$ and $0 < \tau \le \infty$,

$$\left(\sum_{\triangle \in \mathcal{T}_m} (|\triangle|^\gamma \|g\|_{L_q(\triangle)})^\tau \right)^{1/\tau} \approx \left(\sum_{\theta \in \Theta_m} (|E_\theta|^\gamma \|b_\theta \varphi_\theta\|_q)^\tau \right)^{1/\tau}, \tag{2.15}$$

where the constants of equivalence are independent of m and g. In the case $q = \infty$ (or $\tau = \infty$) the ℓ_q-norm (ℓ_τ-norm) above is replaced by the sup-norm as usual.

Proof. We have to prove only (2.15), since the first statement of the lemma then follows with $\gamma = 0$ and $\tau = q$. For each $\triangle \in \mathcal{T}_m$, we have by (2.10)

$$\|g\|_{L_q(\triangle)} = \| \sum_{\theta \in \Theta_m,\, E_\theta \supset \triangle} b_\theta \varphi_\theta \|_q \le c \sum_{\theta \in \Theta_m,\, E_\theta \supset \triangle} \|b_\theta \varphi_\theta\|_q.$$

Therefore, by Lemma 2.1 and (2.7),

$$\begin{aligned} \sum_{\triangle \in \mathcal{T}_m} (|\triangle|^\gamma \|g\|_{L_q(\triangle)})^\tau &\le c \sum_{\triangle \in \mathcal{T}_m} \sum_{\theta \in \Theta_m,\, E_\theta \supset \triangle} (|E_\theta|^\gamma \|b_\theta \varphi_\theta\|_q)^\tau \\ &\le c \sum_{\theta \in \Theta_m} (|E_\theta|^\gamma \|b_\theta \varphi_\theta\|_q)^\tau. \end{aligned}$$

In the other direction, since Φ is a basis of $\mathcal{S}_m$ and $g \in \mathcal{S}_m$, we have $b_\theta = a_\theta(g)$, $\theta \in \Theta_m$, and hence, by (2.12), (2.14), and (2.11),

$$\begin{aligned} \|b_\theta \varphi_\theta\|_q = \|a_\theta(g) \varphi_\theta\|_q &\le c \|g\|_{L_\infty(E_\theta)} \|\varphi_\theta\|_q \le c \|g\|_{L_\infty(E_\theta)} |E_\theta|^{1/q} \\ &\le c \|g\|_{L_q(E_\theta)} \le c \sum_{\triangle \in \mathcal{T}_m,\, \triangle \subset E_\theta} \|g\|_{L_q(\triangle)}. \end{aligned}$$

Since $|E_\theta| \approx |\triangle|$ if $\triangle \in \mathcal{T}_m$ and $\triangle \subset E_\theta$, we have, by (2.7) and Lemma 2.1,

$$\begin{aligned} \sum_{\theta \in \Theta_m} (|E_\theta|^\gamma \|b_\theta \varphi_\theta\|_q)^\tau &\le c \sum_{\theta \in \Theta_m} \sum_{\triangle \in \mathcal{T}_m,\, \triangle \subset E_\theta} (|\triangle|^\gamma \|g\|_{L_q(\triangle)})^\tau \\ &\le c \sum_{\triangle \in \mathcal{T}_m} (|\triangle|^\gamma \|g\|_{L_q(\triangle)})^\tau. \end{aligned}$$

□

Local polynomial approximation is an important tool in spline approximation. For a function $f \in L_q(G)$, $G \subset \mathbb{R}^2$, we denote by $E_k(f, G)_q$ the error of the best L_q-approximation to f on G from Π_k and by $\omega_k(f, G)_q$ the kth *local modulus of smoothness* of f on G:

$$E_k(f, G)_q := \inf_{P \in \Pi_k} \|f - P\|_{L_q(G)}, \qquad \omega_k(f, G)_q := \sup_{h \in \mathbb{R}^2} \|\Delta_h^k(f, \cdot)\|_{L_q(G)}.$$

Whitney's theorem gives an important relation between these two quantities: If $f \in L_q(G)$, $0 < q \le \infty$, where $G = \triangle$ is an arbitrary triangle or $G = \Omega_\triangle$ with $\triangle \in \mathcal{T}$, $\mathcal{T}$ is an SLR-triangulation, then

$$E_k(f, G)_q \le c \omega_k(f, G)_q, \tag{2.16}$$

where $c = c(q, k)$ if $G = \triangle$ and $c = c(q, k, \delta_2)$ if $G = \Omega_\triangle$ (δ_2 is from (2.3)). For a proof of this estimate, see, e.g., the appendix of [38]. Note that this estimate holds

for much more general regions G, but then the constant $c = c(G)$ may become hard to control.

For $0 < q \le \infty$ and a triangle $\triangle$, we let $P_{\triangle,q} : L_q(\triangle) \to \Pi_k$ be a projector such that

$$\|f - P_{\triangle,q}(f)\|_{L_q(\triangle)} \le cE_k(f,\triangle)_q \quad \text{for } f \in L_q(\triangle). \tag{2.17}$$

Note that $P_{\triangle,q}$ can be realized as a linear projector if $q \ge 1$. For instance, one can utilize the averaged Taylor polynomial. Namely, suppose $\triangle_0$ is an equilateral reference triangle and A is an affine transform mapping $\triangle$ onto $\triangle_0$. Let now $P(g) \in \Pi_k$ be the averaged Taylor polynomial of the function $g := f \circ A^{-1}$ (the composition of f with A^{-1}) over the disc B inscribed in $\triangle_0$ (see, e.g., section 4.1 of [12]). Clearly, $P : L_q(B) \to \Pi_k$ is a linear operator, $\|P(g)\|_{L_q(B)} \le c\|g\|_{L_q(B)}$ $(q \ge 1)$, and P is a projector, i.e., $P(Q) = Q$ for $Q \in \Pi_k$. From these properties of P, it follows that for an arbitrary $Q \in \Pi_k$,

$$\begin{aligned}\|g - P(g)\|_{L_q(\triangle_0)} &\le \|g - Q\|_{L_q(\triangle_0)} + \|Q - P(g)\|_{L_q(\triangle_0)} \\ &\le \|g - Q\|_{L_q(\triangle_0)} + c\|P(g - Q)\|_{L_q(B)} \le c\|g - Q\|_{L_q(\triangle_0)},\end{aligned}$$

which implies $\|g - P(g)\|_{L_q(\triangle_0)} \le cE_k(g,\triangle_0)_q$. Substituting back, one easily obtains $\|f - (P \circ A)(f)\|_{L_q(\triangle)} \le cE_k(f,\triangle)_q$. Finally, we set $P_{\triangle,q} := P \circ A$, which is the desired linear projector of $L_q(\triangle)$ into Π_k.

Note that $P_{\triangle,q}$ cannot be realized as a linear operator if $0 < q < 1$ (otherwise, we would be able to construct a nonzero bounded linear functional on L_q).

We define a linear operator $Q_m : \mathcal{S}^{k,-1}(\mathcal{T}_m) \to \mathcal{S}_m$ as follows. For each $\theta \in \Theta_m$, let $\lambda_\theta : \mathcal{S}^{k,-1}(\mathcal{T}_m)|_{E_\theta} \to \mathbb{R}$ be a linear functional such that

$$\lambda_\theta(s|_{E_\theta}) = a_\theta(s), \quad s \in \mathcal{S}_m, \quad \text{and}$$

$$|\lambda_\theta(f)| \le M_2\|f\|_{L_\infty(E_\theta)}, \quad f \in \mathcal{S}^{k,-1}(\mathcal{T}_m)|_{E_\theta}.$$

Such linear functional always exists by the Hahn–Banach theorem. We set

$$Q_m(s) := \sum_{\theta\in\Theta_m} \lambda_\theta(s|_{E_\theta})\varphi_\theta, \qquad s \in \mathcal{S}^{k,-1}(\mathcal{T}_m). \tag{2.18}$$

Clearly, $Q_m(s) = s$ if $s \in \mathcal{S}_m$, and thus Q_m is a linear projector of $\mathcal{S}^{k,-1}(\mathcal{T}_m)$ into $\mathcal{S}_m$.

LEMMA 2.4. *For any $s \in \mathcal{S}^{k,-1}(\mathcal{T}_m)$, $0 < q \le \infty$, and $\triangle \in \mathcal{T}_m$,*

$$\|Q_m(s)\|_{L_q(\triangle)} \le c\|s\|_{L_q(\Omega^\ell_\triangle)}, \tag{2.19}$$

with a constant c independent of m, $\triangle$, and s.

Proof. By Lemma 2.2 and (2.14), we have

$$\begin{aligned}\|\varphi_\theta\|_{L_q(\triangle)} \le c_1|\triangle|^{1/q}\|\varphi_\theta\|_{L_\infty(\triangle)} \le c_1M_1|\triangle|^{1/q}, \\ \|s\|_{L_\infty(E_\theta)} \le c_2|\triangle|^{-1/q}\|s\|_{L_q(E_\theta)},\end{aligned}$$

where c_1 and c_2 depend only on q and k. Therefore,

$$\begin{aligned}\|Q_m(s)\|_{L_q(\triangle)} &= \Big\| \sum_{\substack{\theta\in\Theta_m \\ \triangle\subset E_\theta}} \lambda_\theta(s|_{E_\theta})\varphi_\theta \Big\|_{L_q(\triangle)} \le c \sum_{\substack{\theta\in\Theta_m \\ \triangle\subset E_\theta}} |\lambda_\theta(s|_{E_\theta})|\,\|\varphi_\theta\|_{L_q(\triangle)} \\ &\le c \sum_{\substack{\theta\in\Theta_m \\ \triangle\subset E_\theta}} \|s\|_{L_\infty(E_\theta)}|\triangle|^{1/q} \le c \sum_{\substack{\theta\in\Theta_m \\ \triangle\subset E_\theta}} \|s\|_{L_q(E_\theta)} \le c\|s\|_{L_q(\Omega^\ell_\triangle)}. \qquad \square\end{aligned}$$

We now extend Q_m to $L_q^{\rm loc}(\mathbb{R}^2)$, $0 < q \le \infty$. Let $P_{\triangle,q} : L_q(\triangle) \to \Pi_k$ be a projector satisfying (2.17) (linear if $q \ge 1$). We define

$$p_{m,q}(f) := \sum_{\triangle\in\mathcal{T}_m} \mathbb{1}_\triangle \cdot P_{\triangle,q}(f) \quad \text{for } f \in L_q^{\rm loc}, \tag{2.20}$$

which is a projector of $L_q^{\rm loc}$ into $\mathcal{S}_m^{k,-1}$.

We put

$$Q_{m,q}(f) := Q_m(p_{m,q}(f)) \quad \text{for } f \in L_q^{\rm loc}, \tag{2.21}$$

which is evidently a projector of $L_q^{\rm loc}$ into $\mathcal{S}_m$ (linear if $q \ge 1$ and all $P_{\triangle,q}$ are linear).

We next show that $Q_{m,q}$ provides a good local L_q-approximation from $\mathcal{S}_m$. We let $\mathbb{S}_\triangle(f)_q$ denote the error of $L_q(\Omega_\triangle^\ell)$-approximation from $\mathcal{S}_m$, i.e.,

$$\mathbb{S}_\triangle(f)_q := \inf_{s\in\mathcal{S}_m} \|f - s\|_{L_q(\Omega_\triangle^\ell)}, \quad \triangle \in \mathcal{T}_m. \tag{2.22}$$

Thus, $\mathbb{S}_\triangle(f)_q$ is the error of approximation to f from *restrictions* to $\Omega_\triangle^\ell$ of functions from $\mathcal{S}_m$, which is not necessarily the same as the approximation by all r times differentiable piecewise polynomials of degree $< k$ *defined only* on $\Omega_\triangle^\ell$, even if $\mathcal{S}_m$ coincides with $\mathcal{S}_m^{k,r}$. However, since $\Pi_{\tilde k} \subset \mathcal{S}_m$, $\mathbb{S}_\triangle(f)_q$ does not exceed the error of $L_q(\Omega_\triangle^\ell)$-approximation to f from polynomials of degree $< \tilde k$.

LEMMA 2.5. *If $f \in L_q^{\rm loc}(\mathbb{R}^2)$, $0 < q \le \infty$ ($f \in C$ if $q = \infty$), then*

$$\|f - Q_{m,q}(f)\|_{L_q(\triangle)} \le c\mathbb{S}_\triangle(f)_q, \quad \triangle \in \mathcal{T}_m \ (m \in \mathbb{Z}),$$

with c independent of f, m, and $\triangle$.

Proof. Let $s_\triangle \in \mathcal{S}_m$ be such that $\|f - s_\triangle\|_{L_q(\Omega_\triangle^\ell)} \le c\,\mathbb{S}_\triangle(f)_q$. Using the properties of Q_m (see Lemma 2.4), we find

$$\begin{aligned}
\|f - Q_{m,q}(f)\|_{L_q(\triangle)} &= \|f - Q_m(p_{m,q}(f))\|_{L_q(\triangle)} \\
&\le c\|f - s_\triangle\|_{L_q(\triangle)} + c\|s_\triangle - Q_m(p_{m,q}(f))\|_{L_q(\triangle)} \\
&\le c\mathbb{S}_\triangle(f)_q + c\|Q_m(s_\triangle - p_{m,q}(f))\|_{L_q(\triangle)} \\
&\le c\mathbb{S}_\triangle(f)_q + c\|s_\triangle - p_{m,q}(f)\|_{L_q(\Omega_\triangle^\ell)} \\
&\le c\mathbb{S}_\triangle(f)_q + c\|f - s_\triangle\|_{L_q(\Omega_\triangle^\ell)} + c\|f - p_{m,q}(f)\|_{L_q(\Omega_\triangle^\ell)} \\
&\le c\mathbb{S}_\triangle(f)_q. \qquad \square
\end{aligned}$$

LEMMA 2.6. (a) *If $f \in L_q^{\rm loc}(\mathbb{R}^2)$, $0 < q \le \infty$, then for every compact $K \subset \mathbb{R}^2$,*

$$\|f - Q_{m,q}(f)\|_{L_q(K)} \to 0 \quad \textit{as} \quad m \to \infty. \tag{2.23}$$

(b) *If $f \in L_q(\mathbb{R}^2)$, $0 < q \le \infty$, then*

$$\|f - Q_{m,q}(f)\|_{L_q(\mathbb{R}^2)} \to 0 \quad \textit{as} \quad m \to \infty. \tag{2.24}$$

For the proof of this lemma, we need the following result.

LEMMA 2.7. *If $\mathcal{T}$ is an LR-triangulation, then for each triangle $\triangle^\diamond \in \mathcal{T}$*

$$\max\{\operatorname{diam}(\triangle) : \triangle \in \mathcal{T}_m, \triangle \subset \triangle^\diamond\} \to 0 \quad \textit{as} \quad m \to \infty. \tag{2.25}$$

Proof. Let $m_0 := \text{level}(\triangle^\diamond)$. We set $d_m := \max\{\text{diam}(\triangle) : \triangle \in \mathcal{T}_m, \triangle \subset \triangle^\diamond\}$. Since $(d_m)_{m=m_0}^\infty$ is nonincreasing, it suffices to show the existence of a subsequence tending to zero. Let e be an edge of a triangle $\triangle \in \mathcal{T}_m$, $\triangle \subset \triangle^\diamond$. If it is also an edge of a child of $\triangle$, then the valence of at least one of the two endpoints of e will increase by one at level $m+1$. (Recall that there are always at least two children, so that a child and a parent cannot be the same triangle.) Therefore, e will be subdivided at least once after at most $S := 2(N_0 - 3) + 1$ steps of refinement. By (2.1), it readily follows that any edge e' obtained by subdividing e satisfies $|e'| \le \rho|e| \le \rho d_m$.

We call an edge of a descendant of $\triangle^\diamond$ a *cutting edge* for $\triangle^\diamond$ if one of its endpoints is a vertex of $\triangle^\diamond$ and the other lies in the interior of the opposite edge of $\triangle^\diamond$. Since all cutting edges must emanate from the same vertex of $\triangle^\diamond$, there are totally no more than $M := N_0 - 3$ such edges for $\triangle^\diamond$. Therefore, no *new* cutting edges for $\triangle^\diamond$ will be created at levels $m > m_0 + M$. (It is easy to see that as soon as no new cutting edges are created at a level m, they cannot be created on any further level.) Using this and the above observation, we conclude that there will be no cutting edges at levels $m > m_0 + M + S$ since they all will be subdivided. Therefore, each edge e inside $\triangle^\diamond$ at these levels is either a proper part of an edge of $\triangle^\diamond$ or has both of its endpoints in the interiors of two different edges of $\triangle^\diamond$, or else it has at least one endpoint in the interior of $\triangle^\diamond$. In all cases, condition (2.1) ensures that $|e| \le \rho d_{m_0}$, which implies $d_{m_1} \le \rho d_{m_0}$, where $m_1 = m_0 + M + S + 1$. It is clear now that there is an increasing sequence $\{m_k\}_{k=1}^\infty$ such that

$$d_{m_k} \le \rho^k d_{m_0} \to 0 \qquad \text{as} \quad k \to \infty,$$

which completes the proof. □

Proof of Lemma 2.6. (a) By condition (f) on triangulations, it suffices to prove the lemma for $K = \triangle^\diamond$, an arbitrary triangle from $\mathcal{T}$. By Lemma 2.7,

$$\max\{\text{diam}(\Omega^\ell_\triangle) : \triangle \in \mathcal{T}_m, \triangle \subset \Omega^\ell_{\triangle^\diamond}\} \to 0 \quad \text{as} \quad m \to \infty. \tag{2.26}$$

Case 1. $q < \infty$. Fix $\varepsilon > 0$. In view of (2.26), there exists a piecewise constant function S_ε of the form

$$S_\varepsilon = \sum_{\triangle \in \mathcal{T}_{m_\varepsilon},\, \triangle \subset \Omega^\ell_{\triangle^\diamond}} c_\triangle \mathbb{1}_\triangle, \quad m_\varepsilon \ge \text{level}(\triangle^\diamond),$$

such that

$$\|f - S_\varepsilon\|_{L_q(\Omega^\ell_{\triangle^\diamond})} < \varepsilon \tag{2.27}$$

(choose first $g \in C(\Omega^\ell_{\triangle^\diamond})$ so that $\|f - g\|_{L_q(\Omega^\ell_{\triangle^\diamond})} < \varepsilon/2$ and then choose S_ε so that $\|g - S_\varepsilon\|_{L_\infty(\Omega^\ell_{\triangle^\diamond})} < \frac{\varepsilon}{2}|\Omega^\ell_{\triangle^\diamond}|^{-1/q}$). Then $Q_{m,q}(S_\varepsilon) = Q_m(S_\varepsilon)$.

We have, for $m \ge m_\varepsilon$,

$$\begin{aligned} \|f - Q_{m,q}(f)\|_{L_q(\triangle^\diamond)} &\le c\|f - S_\varepsilon\|_{L_q(\triangle^\diamond)} + c\|S_\varepsilon - Q_{m,q}(S_\varepsilon)\|_{L_q(\triangle^\diamond)} \\ &\quad + c\|Q_m(S_\varepsilon - p_{m,q}(f))\|_{L_q(\triangle^\diamond)}. \end{aligned} \tag{2.28}$$

For the third term above, we have

$$\begin{aligned} \|Q_m(S_\varepsilon - p_{m,q}(f))\|_{L_q(\triangle^\diamond)} &\le c\|S_\varepsilon - p_{m,q}(f)\|_{L_q(\Omega^\ell_{\triangle^\diamond})} \\ &\le c\|f - S_\varepsilon\|_{L_q(\Omega^\ell_{\triangle^\diamond})} + c\|f - p_{m,q}(f)\|_{L_q(\Omega^\ell_{\triangle^\diamond})} \\ &\le c\|f - S_\varepsilon\|_{L_q(\Omega^\ell_{\triangle^\diamond})} \le c\varepsilon, \end{aligned} \tag{2.29}$$

where we used Lemma 2.4 and that $\|f - p_{m,q}(f)\|_{L_q(\Omega^\ell_{\triangle^\diamond})} \le c\|f - S_\varepsilon\|_{L_q(\Omega^\ell_{\triangle^\diamond})}$ ($m \ge m_\varepsilon$), by (2.17).

It remains to show that $\|S_\varepsilon - Q_{m,q}(S_\varepsilon)\|_{L_q(\triangle^\diamond)} \le c\varepsilon$ for sufficiently large m. Denote by G the union of the edges of all triangles $\triangle \in \mathcal{T}_{m_\varepsilon}$ such that $\triangle \subset \triangle^\diamond$ and by $G_\delta := \{x \in \mathbb{R}^2 : \operatorname{dist}(x, G) \le \delta\}$ the δ-neighborhood of G. Clearly, there exists $\delta > 0$ such that $\|S_\varepsilon\|_{L_q(G_\delta)} < \varepsilon$.

By (2.26), there exists $m_1 \ge m_\varepsilon$ such that $\operatorname{diam}(\Omega^\ell_\triangle) < \delta$ for all triangles $\triangle \in T_m$ ($m \ge m_1$) such that $\triangle \subset \triangle^\diamond$ and $\Omega^\ell_\triangle \cap G \ne \emptyset$. Since $\Pi_1 \subset \mathcal{S}_m$, $Q_m(S_\varepsilon)|_\triangle = S_\varepsilon|_\triangle$ if $S_\varepsilon|_{\Omega^\ell_\triangle} = \textit{constant}$. Using this, we obtain by Lemma 2.5

$$\|S_\varepsilon - Q_{m,q}(S_\varepsilon)\|_{L_q(\triangle^\diamond)} \le c\left(\sum_{\triangle\in T_m,\, \Omega^\ell_\triangle\cap G\ne\emptyset} \mathbb{S}_\triangle(S_\varepsilon)_q^q\right)^{1/q} \le c\|S_\varepsilon\|_{L_q(G_\delta)} \le c\varepsilon.$$

We substitute this estimate together with (2.27) and (2.29) in (2.28) to obtain

$$\|f - Q_{m,q}(f)\|_{L_q(\triangle^\diamond)} \le c\varepsilon \quad \text{for } m \ge m_1.$$

This implies (2.23) if $q < \infty$.

Case 2. $q = \infty$. We have, by Lemma 2.5 and the fact that $\Pi_1 \subset \mathcal{S}_m$,

$$\|f - Q_{m,q}(f)\|_{L_\infty(\triangle^\diamond)} \le c \max_{\triangle\in\mathcal{T}_m,\, \triangle\subset\triangle^\diamond} \inf_{C\in\Pi_1} \|f - C\|_{L_\infty(\Omega^\ell_\triangle)}.$$

Now the result follows, using (2.26) and the fact that f is uniformly continuous on $\Omega^\ell_{\triangle^\diamond}$.

Part (b) of the lemma is immediate from part (a). □

We denote $\mathcal{S}_{-\infty} := \bigcap_{m\in\mathbb{Z}} \mathcal{S}_m$. As we already mentioned, there are only two possibilities for $\mathcal{T}_{-\infty}$: $\mathcal{T}_{-\infty} = \{\mathbb{R}^2\}$ or $\mathcal{T}_{-\infty} = (\triangle^j_\infty)_{j=1}^{N_\infty}$, $N_\infty \le N_0$, where $\{\triangle^j_\infty\}$ are infinite triangles with disjoint interiors and a common vertex which cover $\mathbb{R}^2$. If $\mathcal{T}_{-\infty} = \{\mathbb{R}^2\}$, then obviously $\mathbb{R}^2$ is the union of a sequence of nested triangles, and hence each $s \in \mathcal{S}_{-\infty}$ is a polynomial of degree $< k$ on $\mathbb{R}^2$. Therefore, if $\mathcal{T}_{-\infty} = \{\mathbb{R}^2\}$, then $\mathcal{S}_{-\infty}$ a subspace of Π_k.

Suppose $\mathcal{T}_{-\infty} = (\triangle^j_\infty)_{j=1}^{N_\infty}$ and $s \in \mathcal{S}_{-\infty}$. Then each triangle $\triangle^j_\infty$ can be represented as the union of a sequence of nested triangles, and hence s is a polynomial of degree $< k$ on $\triangle^j_\infty$. Therefore, in this case, $s \in \mathcal{S}_{-\infty}$ implies $s \in C^r(\mathbb{R}^2)$ and $s|_{\triangle^j_\infty} = P_j|_{\triangle^j_\infty}$ for some $P_j \in \Pi_k$, $j = 1, \dots, N_\infty$.

Furthermore, if $s \in \mathcal{S}_{-\infty}$ and $|\{x \in \mathbb{R}^2 : |s(x)| > t\}| < \infty$ for some $t > 0$, then $s = \text{const}$. In particular, if $s \in \mathcal{S}_{-\infty} \cap L_p$ ($p < \infty$), then $s \equiv 0$.

2.3. Definition of B-spaces. Equivalent norms. Interpolation. Suppose $\mathcal{T}$ is an LR(or better)-triangulation and $\Phi = \Phi_\mathcal{T}$ is a family of differentiable piecewise polynomial basis functions over $\mathcal{T}$ as described in sections 2.1–2.2. For the characterization of nonlinear n-term L_p-approximation from Φ, we need the *B-spaces* $B^\alpha_\tau(\Phi)$ which we shall introduce and explore in this subsection. In fact, the spaces $B^\alpha_\tau(\Phi)$ depend only on the underlying ladder of spaces $\cdots \subset \mathcal{S}_{-1} \subset \mathcal{S}_0 \subset \mathcal{S}_1 \subset \cdots$ associated with the bases $(\Phi_m)_{m\in\mathbb{Z}}$, but as it will be shown below these spaces have atomic representations using Φ, which justifies our notation.

We shall need the B-spaces $B^\alpha_\tau(\Phi)$ in two cases: (a) $0 < p < \infty$ and $\alpha > 0$, or (b) $p = \infty$ and $\alpha \ge 1$ (see Remark 2.14). In both cases, we define τ from the identity $1/\tau = \alpha + 1/p$ ($1/\infty := 0$).

Definition of $B^\alpha_\tau(\Phi)$ via local approximation. We define the B-space $B^\alpha_\tau(\Phi)$ as the set of all functions $f \in L_\tau(\mathbb{R}^2)$ such that

$$\|f\|_{B^\alpha_\tau(\Phi)} := \left(\sum_{\triangle\in\mathcal{T}} (|\triangle|^{-\alpha}\mathbb{S}_\triangle(f)_\tau)^\tau \right)^{1/\tau} < \infty, \tag{2.30}$$

where $\mathbb{S}_\triangle(f)_\tau$ is the error of L_τ-approximation of f on $\Omega^\ell_\triangle$ from $\mathcal{S}_m$ if $\triangle \in \mathcal{T}_m$ (see (2.22)).

It is readily seen that $B^\alpha_\tau(\Phi)$ is a linear space, $\|cf\|_{B^\alpha_\tau} = |c|\|f\|_{B^\alpha_\tau}$, and $\|f + g\|^\lambda_{B^\alpha_\tau} \le \|f\|^\lambda_{B^\alpha_\tau} + \|g\|^\lambda_{B^\alpha_\tau}$, with $\lambda := \min\{\tau, 1\}$. Clearly, see Theorem 2.8, if $\|f\|_{B^\alpha_\tau} = 0$, then $f = 0$ a.e. Therefore, $\|\cdot\|_{B^\alpha_\tau}$ is a norm if $\tau \ge 1$ and a quasi norm if $\tau < 1$.

We next define other equivalent norms in $B^\alpha_\tau(\Phi)$. We define

$$N_{\Phi,\mathbb{S},\eta}(f) := \left(\sum_{\triangle\in\mathcal{T}} (|\triangle|^{1/p-1/\eta}\mathbb{S}_\triangle(f)_\eta)^\tau \right)^{1/\tau}, \tag{2.31}$$

where we have taken into account that $1/\tau := \alpha + 1/p$. Thus, $N_{\Phi,\mathbb{S},\tau}(f) = \|f\|_{B^\alpha_\tau(\Phi)}$. Moreover, we shall show that $N_{\Phi,\mathbb{S},\eta}(f) \approx \|f\|_{B^\alpha_\tau(\Phi)}$ if $0 < \eta < p$ (see Theorem 2.10).

Definition of norms in $B^\alpha_\tau(\Phi)$ via basis functions (atomic decomposition). For $f \in L_\tau(\mathbb{R}^2)$, we define

$$N_\Phi(f) := \inf_{f=\sum_{\theta\in\Theta} c_\theta\varphi_\theta} \left(\sum_{\theta\in\Theta} (|E_\theta|^{-\alpha}\|c_\theta\varphi_\theta\|_\tau)^\tau \right)^{1/\tau}, \tag{2.32}$$

where the infimum is over all representations of f in the form $f = \sum_{\theta\in\Theta} c_\theta\varphi_\theta$ in L_τ. (Note that the existence of such representations for each $f \in L_\tau$ follows by Lemma 2.6.) By Theorem 2.9,

$$\sum_{\theta\in\Theta} (|E_\theta|^{-\alpha}\|c_\theta\varphi_\theta\|_\tau)^\tau < \infty \quad \text{implies} \quad \Big\| \sum_{\theta\in\Theta} |c_\theta\varphi_\theta(\cdot)| \Big\|_p < \infty,$$

and hence $\sum_{\theta\in\Theta} c_\theta\varphi_\theta(x)$ converges absolutely a.e. Therefore, the specific type of convergence that we use in the definition of $N_\Phi(f)$ above is not essential. Using (2.14), we have

$$\begin{aligned} N_\Phi(f) &\approx \inf_{f=\sum_{\theta\in\Theta} c_\theta\varphi_\theta} \left(\sum_{\theta\in\Theta} (|E_\theta|^{1/p-1/\eta}\|c_\theta\varphi_\theta\|_\eta)^\tau \right)^{1/\tau} \\ &\approx \inf_{f=\sum_{\theta\in\Theta} c_\theta\varphi_\theta} \left(\sum_{\theta\in\Theta} \|c_\theta\varphi_\theta\|_p^\tau \right)^{1/\tau}. \end{aligned} \tag{2.33}$$

Definition of norms in $B^\alpha_\tau(\Phi)$ via projections. For $f \in L^{\mathrm{loc}}_\eta$, we set

$$q_{m,\eta}(f) := Q_{m,\eta}(f) - Q_{m-1,\eta}(f) \in \mathcal{S}_m, \tag{2.34}$$

where $Q_{m,\eta}$ is from (2.21), and let $(b_{\theta,\eta}(f))_{\theta\in\Theta_m}$ be defined by the identity

$$q_{m,\eta}(f) = \sum_{\theta\in\Theta_m} b_{\theta,\eta}(f)\varphi_\theta, \text{ i.e., } \quad b_{\theta,\eta}(f) := a_\theta(q_{m,\eta}(f)), \quad \theta \in \Theta_m. \tag{2.35}$$

We define

$$N_{\Phi,Q,\tau}(f) := \left(\sum_{\theta\in\Theta} (|E_\theta|^{-\alpha} \|b_{\theta,\tau}(f)\varphi_\theta\|_\tau)^\tau \right)^{1/\tau} \tag{2.36}$$

and, more generally (see (2.31)),

$$N_{\Phi,Q,\eta}(f) := \left(\sum_{\theta\in\Theta} (|E_\theta|^{1/p-1/\eta} \|b_{\theta,\eta}(f)\varphi_\theta\|_\eta)^\tau \right)^{1/\tau}, \quad 0<\eta<p. \tag{2.37}$$

By Lemmas 2.2–2.3, it follows that

$$N_{\Phi,Q,\eta}(f) \approx \left(\sum_{m\in\mathbb{Z}} \sum_{\triangle\in\mathcal{T}_m} (|\triangle|^{1/p-1/\eta} \|q_{m,\eta}(f)\|_{L_\eta(\triangle)})^\tau \right)^{1/\tau} \tag{2.38}$$

and, for $0<\mu\le\infty$,

$$N_{\Phi,Q,\eta}(f) \approx \left(\sum_{\theta\in\Theta} (|E_\theta|^{1/p-1/\mu} \|b_{\theta,\eta}(f)\varphi_\theta\|_\mu)^\tau \right)^{1/\tau} \approx \left(\sum_{\theta\in\Theta} \|b_{\theta,\eta}(f)\varphi_\theta\|_p^\tau \right)^{1/\tau}. \tag{2.39}$$

We shall show (see Theorem 2.10) that all of the above norms are equivalent. To this end, we need the following embedding theorem.

THEOREM 2.8. *If $f\in L_\tau(\mathbb{R}^2)$ and $N_{\Phi,Q,\eta}(f)<\infty, 0<\eta<p$, then*

$$f = \sum_{m\in\mathbb{Z}} q_{m,\eta}(f) = \sum_{\theta\in\Theta} b_{\theta,\eta}(f)\varphi_\theta, \tag{2.40}$$

with the series converging absolutely a.e. and in L_p, and

$$\|f\|_p \le c\Big\| \sum_{m\in\mathbb{Z}} |q_{m,\eta}(f)(\cdot)| \Big\|_p \le c\Big\| \sum_{\theta\in\Theta} |b_{\theta,\eta}(f)\varphi_\theta(\cdot)| \Big\|_p \le cN_{\Phi,Q,\eta}(f), \tag{2.41}$$

with c independent of f.

The proof of Theorem 2.8 hinges on the following more general embedding theorem, which is a special case of Theorem 2.5 from [54].

THEOREM 2.9. *If $0<\tau<p<\infty$, or $p=\infty$, and $0<\tau\le 1$, then for any sequence of real numbers $(c_\theta)_{\theta\in\Theta}$ we have*

$$\Big\| \sum_{\theta\in\Theta} |c_\theta\varphi_\theta(\cdot)| \Big\|_p \le c \left(\sum_{\theta\in\Theta} \|c_\theta\varphi_\theta\|_p^\tau \right)^{1/\tau}, \tag{2.42}$$

with c independent of $(c_\theta)_{\theta\in\Theta}$.

For completeness, we give the simple proof of this theorem in the appendix.

Proof of Theorem 2.8. We introduce the following abbreviated notation: $Q_m := Q_{m,\eta}(f)$, $q_m := q_{m,\eta}(f)$, $b_\theta := b_{\theta,\eta}(f)$, and $N(f) := N_{\Phi,Q,\eta}(f)$. By (2.35), (2.39), and Theorem 2.9, we have

$$\Big\| \sum_{m\in\mathbb{Z}} |q_m(\cdot)| \Big\|_p \le c\Big\| \sum_{\theta\in\Theta} |b_\theta\varphi_\theta(\cdot)| \Big\|_p \le cN(f) < \infty, \tag{2.43}$$

and hence $\sum_{m\in\mathbb{Z}}|q_m(x)|<\infty$ a.e. On the other hand, by Lemma 2.6, we have $\|f-Q_m\|_{L_\eta(\triangle)}\to 0$ as $m\to\infty$ for each $\triangle\in\mathcal{T}$. The above two facts imply

$$f-Q_0=\sum_{m=1}^{\infty}q_m \quad \text{absolutely a.e. on } \mathbb{R}^2. \tag{2.44}$$

We use Lemmas 2.1 and 2.2 to obtain, for $\triangle\in\mathcal{T}_m$ $(m\in\mathbb{Z})$,

$$\|q_m\|_{L_\infty(\triangle)}\le c|\triangle|^{-\frac{1}{p}}\|q_m\|_{L_p(\triangle)}\le c|\triangle|^{-\frac{1}{p}}\sum_{\theta\in\Theta_m,\ E_\theta\supset\triangle}\|b_\theta\varphi_\theta\|_p\le c|\triangle|^{-\frac{1}{p}}N(f).$$

Therefore, for a fixed $\triangle'\in\mathcal{T}_\nu$ $(\nu\in\mathbb{Z})$,

$$\begin{aligned}\sum_{m=-\infty}^{\nu}\|q_m\|_{L_\infty(\triangle')}&\le cN(f)\sum_{\triangle\in\mathcal{T},\,\triangle\supset\triangle'}|\triangle|^{-1/p}\\ &=cN(f)|\triangle'|^{-1/p}\sum_{\triangle\in\mathcal{T},\,\triangle\supset\triangle'}(|\triangle'|/|\triangle|)^{1/p}\\ &\le c|\triangle'|^{-1/p}N(f)<\infty,\end{aligned} \tag{2.45}$$

where we used (2.9). We set

$$s_\infty:=Q_0-\sum_{m=-\infty}^{0}q_m \quad \text{pointwise in } \mathbb{R}^2. \tag{2.46}$$

From (2.45), it follows that s_∞ is well defined and the series in (2.46) converges uniformly on every compact in $\mathbb{R}^2$. Evidently, (2.46) yields $s_\infty=Q_\nu-\sum_{m=-\infty}^{\nu}q_m$ for each $\nu\in\mathbb{Z}$.

Fix $n\in\mathbb{Z}$. Using Theorem 2.9, we obtain, for $\nu\le n$,

$$\begin{aligned}\inf_{s\in\mathcal{S}_n}\|s_\infty-s\|_p\le\|s_\infty-Q_\nu\|_p&=\Big\|\sum_{m=-\infty}^{\nu}q_m\Big\|_p\\ &\le c\Bigg(\sum_{\theta\in\bigcup_{m=-\infty}^{\nu}\Theta_m}\|b_\theta\varphi_\theta\|_p^\tau\Bigg)^{1/\tau}\to 0 \quad \text{as } \nu\to-\infty,\end{aligned}$$

where we used that $(\sum_{\theta\in\Theta}\|b_\theta\varphi_\theta\|_p^\tau)^{1/\tau}\approx N(f)<\infty$. Therefore, $s_\infty\in\mathcal{S}_n$ for every $n\in\mathbb{Z}$, and hence $s_\infty\in\bigcap_{n\in\mathbb{Z}}\mathcal{S}_n=\mathcal{S}_{-\infty}$.

Identities (2.44) and (2.46) yield

$$f-s_\infty=\sum_{m\in\mathbb{Z}}q_{m,\eta}(f)=\sum_{\theta\in\Theta}b_{\theta,\eta}(f)\varphi_\theta \quad \text{absolutely a.e.,} \tag{2.47}$$

and hence, using (2.43),

$$\begin{aligned}\|f-s_\infty\|_p&\le c\Big\|\sum_{m\in\mathbb{Z}}|q_{m,\eta}(f)(\cdot)|\Big\|_p\\ &\le c\Big\|\sum_{\theta\in\Theta}|b_{\theta,\eta}(f)\varphi_\theta(\cdot)|\Big\|_p\le cN_{\Phi,Q,\eta}(f)<\infty.\end{aligned} \tag{2.48}$$

Since $f \in L_\tau$ and $f - s_\infty \in L_p$, it readily follows that, for $t > 0$,

$$\begin{aligned}|\{x : |s_\infty(x)| > t\}| &\le |\{x : |f(x)| > t/2\}| + |\{x : |f(x) - s_\infty(x)| > t/2\}| \\ &\le (t/2)^{-\tau}\|f\|_\tau^\tau + (t/2)^{-p}\|f - s_\infty\|_p^p < \infty,\end{aligned}$$

which implies $s_\infty \equiv 0$ (see the end of section 2.2). From this, (2.47), and (2.48), we infer (2.40) and (2.41). The proof is complete. □

THEOREM 2.10. *The norms $\|\cdot\|_{B_\tau^\alpha(\Phi)}$, $N_{\Phi,\mathbb{S},\eta}(\cdot)$ $(0 < \eta < p)$, $N_\Phi(\cdot)$, and $N_{\Phi,Q,\eta}(\cdot)$ $(0 < \eta < p)$, defined in* (2.30)–(2.32) *and* (2.37), *are equivalent with constants of equivalence depending only on p, α, η, and the parameters of $\mathcal{T}$ and Φ.*

Proof. Theorem 2.8 readily implies

$$N_\Phi(f) \le N_{\Phi,Q,\eta}(f), \quad 0 < \eta < p, \tag{2.49}$$

if $N_{\Phi,Q,\eta}(f) < \infty$.

Suppose $N_{\Phi,\mathbb{S},\eta}(f) < \infty$. For each $\triangle \in \mathcal{T}_m$ $(m \in \mathbb{Z})$, we have, by (2.34) and Lemma 2.5,

$$\|q_{m,\eta}(f)\|_{L_\eta(\triangle)} \le c\|f - Q_{m,\eta}\|_{L_\eta(\triangle)} + c\|f - Q_{m-1,\eta}\|_{L_\eta(\triangle)} \le c\mathbb{S}_\triangle(f)_\eta + c\mathbb{S}_{\triangle^\diamond}(f)_\eta,$$

where $\triangle^\diamond \supset \triangle$, $\triangle^\diamond \in \mathcal{T}_{m-1}$, is the only parent of $\triangle$. These estimates readily imply

$$N_{\Phi,Q,\eta}(f) \le N_{\Phi,\mathbb{S},\eta}(f), \quad 0 < \eta < p. \tag{2.50}$$

It remains to prove that

$$N_{\Phi,\mathbb{S},\eta}(f) \le N_\Phi(f), \quad 0 < \eta < p, \tag{2.51}$$

provided $N_\Phi(f) < \infty$. Evidently, (2.49)–(2.51) imply the desired equivalence of norms.

Notice first that, by Hölder's inequality, $N_{\Phi,\mathbb{S},\mu}(f) \le N_{\Phi,\mathbb{S},\eta}(f)$ if $0 < \mu \le \eta$, and hence it suffices to prove (2.51) only for $\tau < \eta < p$.

Suppose $f \in L_\tau$ and $0 < N_\Phi(f) < \infty$. Then it follows by the definition of $N_\Phi(f)$ that there exists a sequence $(c_\theta)_{\theta\in\Theta}$ such that

$$f = \sum_{\theta\in\Theta} c_\theta\varphi_\theta \quad \text{in } L_\tau \tag{2.52}$$

and $(\sum_{\theta\in\Theta}(|E_\theta|^{-\alpha}\|c_\theta\varphi_\theta\|_\tau)^\tau)^{1/\tau} \le 2N_\Phi(f)$. Theorem 2.9 implies that in (2.52) we have absolute convergence a.e. We next estimate

$$N_{\Phi,\mathbb{S},\eta}(f) := \left(\sum_{\triangle\in\mathcal{T}} [|\triangle|^{1/p-1/\eta}\mathbb{S}_\triangle(f)_\eta]^\tau\right)^{1/\tau}, \tag{2.53}$$

using that $\mathbb{S}_\triangle(g)_\eta = 0$ if $g \in \mathcal{S}_m$ and $\triangle \in \mathcal{T}_m$, and $\mathbb{S}_\triangle(g)_\eta \le \|g\|_{L_\eta(\Omega_\triangle^\ell)}$, in general. We denote $f_j := \sum_{\theta\in\Theta_j} c_\theta\varphi_\theta$. Fix $\triangle' \in \mathcal{T}$ and assume that $\triangle' \in \mathcal{T}_m$ $(m \in \mathbb{Z})$. We have, using Theorem 2.9 $(\tau < \eta < \infty)$ and (2.14),

$$\begin{aligned}\mathbb{S}_{\triangle'}(f)_\eta^\tau = \mathbb{S}_{\triangle'}\left(\sum_{j=m+1}^\infty f_j\right)_\eta^\tau &\le \Big\|\sum_{j=m+1}^\infty f_j\Big\|_{L_\eta(\Omega_{\triangle'}^\ell)}^\tau \\ &\le \Big\|\sum_{j=m+1}^\infty \sum_{\theta\in\Theta_j,\, E_\theta\subset\Omega_{\triangle'}^{2\ell}} c_\theta\varphi_\theta\Big\|_{L_\eta(\Omega_{\triangle'}^\ell)}^\tau \le c\sum_{\theta\in\Theta,\, E_\theta\subset\Omega_{\triangle'}^{2\ell}} \|c_\theta\varphi_\theta\|_\eta^\tau \\ &\le c\sum_{\theta\in\Theta,\, E_\theta\subset\Omega_{\triangle'}^{2\ell}} |E_\theta|^{\tau(1/\eta-1/\tau)}\|c_\theta\varphi_\theta\|_\tau^\tau.\end{aligned}$$

Substituting this in (2.53), we obtain

$$\begin{aligned} N_{\Phi,\mathbb{S},\eta}(f)^\tau &\le c \sum_{\triangle'\in\mathcal{T}} |\triangle'|^{\tau(1/p-1/\eta)} \sum_{\theta\in\Theta,\, E_\theta\subset\Omega^{2\ell}_{\triangle'}} |E_\theta|^{\tau(1/\eta-1/\tau)} \|c_\theta\varphi_\theta\|_\tau^\tau \\ &= c \sum_{\triangle'\in\mathcal{T}} \sum_{\theta\in\Theta,\, E_\theta\subset\Omega^{2\ell}_{\triangle'}} (|E_\theta|/|\triangle'|)^{\tau(1/\eta-1/p)} (|E_\theta|^{-\alpha}\|c_\theta\varphi_\theta\|_\tau)^\tau \\ &\le c \sum_{\theta\in\Theta} (|E_\theta|^{-\alpha}\|c_\theta\varphi_\theta\|_\tau)^\tau \sum_{\triangle'\in\mathcal{T},\, \Omega^{2\ell}_{\triangle'}\supset E_\theta} (|E_\theta|/|\triangle'|)^{\tau(1/\eta-1/p)}, \end{aligned}$$

where we once switched the order of summation. By (2.1)–(2.2),

$$\#\{\triangle'\in\mathcal{T}_\nu : \Omega^{2\ell}_{\triangle'}\supset E_\theta\} \le c(N_0,\ell), \qquad \nu\in\mathbb{Z}, \quad \theta\in\Theta,$$

and $|E_\theta| \le c\rho^j|\triangle'|$ if $E_\theta\subset\Omega^{2\ell}_{\triangle'}$ with $\triangle'\in\mathcal{T}_m$ and $\theta\in\Theta_{m+j}$ $(m\in\mathbb{Z},\, j\ge 0)$. Using these, we obtain

$$\sum_{\triangle'\in\mathcal{T},\, \Omega^{2\ell}_{\triangle'}\supset E_\theta} (|E_\theta|/|\triangle'|)^{\tau(1/\eta-1/p)} \le c\sum_{j=0}^{\infty} \rho^{j\tau(1/\eta-1/p)} \le c < \infty.$$

Therefore, $N_{\Phi,\mathbb{S},\eta}(f)^\tau \le c\sum_{\theta\in\Theta}(|E_\theta|^{-\alpha}\|c_\theta\varphi_\theta\|_\tau)^\tau \le cN_\Phi(f)^\tau$ which yields (2.51). □

The following embedding result is quite obvious.

THEOREM 2.11. *For $0<\alpha_0<\alpha_1$ and $\tau_j := (\alpha_j+1/p)^{-1}$, $j=0,1$, we have the continuous embedding*

$$B^{\alpha_1}_{\tau_1}(\Phi) \subset B^{\alpha_0}_{\tau_0}(\Phi); \tag{2.54}$$

i.e., if $f\in B^{\alpha_1}_{\tau_1}(\Phi)$, then $f\in B^{\alpha_0}_{\tau_0}(\Phi)$ and $\|f\|_{B^{\alpha_0}_{\tau_0}(\Phi)} \le c\|f\|_{B^{\alpha_1}_{\tau_1}(\Phi)}$.

Proof. By Theorem 2.8, if $f\in B^{\alpha_1}_{\tau_1}(\Phi)$, then $f\in L_{\tau_1}\cap L_p\subset L_{\tau_0}$. Fix $0<\eta<p$. Then by (2.39), we have

$$\|f\|_{B^{\alpha_j}_{\tau_j}(\Phi)} \approx \left(\sum_{\theta\in\Theta} \|b_{\theta,\eta}(f)\varphi_\theta\|_p^{\tau_j}\right)^{1/\tau_j}, \quad j=0,1,$$

and the theorem follows since $\tau_1<\tau_0$. □

Interpolation of B-spaces. We first recall some basic definitions from the real interpolation method. We refer the reader to [3] and [4] as general references for interpolation theory. For a pair of quasi-normed spaces X_0, X_1, embedded in a Hausdorff space, the space X_0+X_1 is defined as the collection of all functions f that can be represented as f_0+f_1 with $f_0\in X_0$ and $f_1\in X_1$. The quasi norm in X_0+X_1 is defined by

$$\|f\|_{X_0+X_1} := \inf_{f=f_0+f_1} \|f_0\|_{X_0} + \|f_1\|_{X_1}.$$

Peetre's K-functional is defined for each $f\in X_0+X_1$ and $t>0$ by

$$K(f,t) := K(f,t;X_0,X_1) := \inf_{f=f_0+f_1} \|f_0\|_{X_0} + t\|f_1\|_{X_1}. \tag{2.55}$$

The real interpolation space $(X_0, X_1)_{\lambda,q}$ with $0 < \lambda < 1$ and $0 < q \le \infty$ is defined as the set of all $f \in X_0 + X_1$ such that

$$\|f\|_{(X_0,X_1)_{\lambda,q}} := \|f\|_{X_0+X_1} + \left(\int_0^\infty (t^{-\lambda} K(f,t))^q \frac{dt}{t} \right)^{1/q} < \infty$$

with the L_q-norm replaced by the sup-norm if $q = \infty$.

It is easily seen that if $X_1 \subset X_0$ (X_1 continuously embedded in X_0), then $K(f,t) \approx \|f\|_{X_0}$ for $f \in X_0$ and $t \ge 1$ and, consequently,

$$\|f\|_{(X_0,X_1)_{\lambda,q}} \approx \|f\|_{X_0} + \left(\sum_{\nu=0}^\infty [2^{\nu\lambda} K(f, 2^{-\nu})]^q \right)^{1/q}. \tag{2.56}$$

THEOREM 2.12. *Suppose $0 < p < \infty$ and $\alpha_0, \alpha_1 > 0$, or $p = \infty$ and $\alpha_0, \alpha_1 \ge 1$. Let $\tau_j := (\alpha_j + 1/p)^{-1}$, $j = 0, 1$. Then*

$$(B^{\alpha_0}_{\tau_0}(\Phi), B^{\alpha_1}_{\tau_1}(\Phi))_{\lambda,\tau} = B^\alpha_\tau(\Phi) \tag{2.57}$$

with equivalent norms, provided $\alpha = (1-\lambda)\alpha_0 + \lambda\alpha_1$ with $0 < \lambda < 1$ and $\tau := (\alpha + 1/p)^{-1}$.

Proof. We shall use some ideas from [32]. We may assume that $\alpha_0 < \alpha_1$. We denote briefly $B^\alpha := B^\alpha_\tau(\Phi)$ and $B^{\alpha_j} := B^{\alpha_j}_{\tau_j}(\Phi)$, $j = 0, 1$. Furthermore, we denote by ℓ_q the space of all sequences $\mathbf{a} = (a_\theta)_{\theta\in\Theta}$ of real numbers such that

$$\|\mathbf{a}\|_{\ell_q} := \left(\sum_{\theta\in\Theta} |a_\theta|^q \right)^{1/q} < \infty.$$

We shall utilize the following well-known interpolation result (see, e.g., [3]):

$$(\ell_{\tau_0}, \ell_{\tau_1})_{\lambda,\tau} = \ell_\tau, \quad \text{where } \tfrac{1}{\tau} = \tfrac{1-\lambda}{\tau_0} + \tfrac{\lambda}{\tau_1} \text{ with } 0 < \lambda < 1. \tag{2.58}$$

We fix $0 < \eta < p$. Then we normalize the basis functions from Φ in L_p, that is, $\|\varphi_\theta\|_p = 1$ (we use the same notation for the normalized basis functions). We also renormalize the dual functionals λ_θ in the definition of Q_m in (2.18) accordingly.

We denote by $\mathbf{b}(f) = (b_\theta(f))_{\theta\in\Theta}$ the sequence of numbers defined by (see (2.34)–(2.35))

$$q_{m,\eta}(f) =: \sum_{\theta\in\Theta_m} b_\theta(f)\varphi_\theta, \quad m \in \mathbb{Z} \quad (\|\varphi_\theta\|_p = 1).$$

By Theorem 2.8, Theorem 2.10, and (2.39), if $f \in B^{\alpha_j}$ ($j = 0, 1$), then

$$f \overset{L_p}{=} \sum_{\theta\in\Theta} b_\theta(f)\varphi_\theta \quad \text{and} \quad \|f\|_{B^{\alpha_j}} \approx \|\mathbf{b}(f)\|_{\ell_{\tau_j}}, \tag{2.59}$$

and similarly for $f \in B^\alpha$.

The theorem will follow by (2.58) and the following lemma.

LEMMA 2.13. *For $f \in B^{\alpha_0} + B^{\alpha_1} = B^{\alpha_0}$ ($\alpha_0 < \alpha_1$), we have*

$$K(f,t; B^{\alpha_0}, B^{\alpha_1}) \approx K(\mathbf{b}(f), t; \ell_{\tau_0}, \ell_{\tau_1}), \quad t > 0. \tag{2.60}$$

Proof. We first prove that

$$K(f,t;B^{\alpha_0},B^{\alpha_1}) \le cK(\mathbf{b}(f),t;\ell_{\tau_0},\ell_{\tau_1}), \quad t>0. \tag{2.61}$$

Indeed, let $\mathbf{a}=(a_\theta)_{\theta\in\Theta}\in\ell_{\tau_1}$. Then $\mathbf{a}\in\ell_{\tau_0}$ $(\tau_0>\tau_1)$ and since $\mathbf{b}(f)\in\ell_{\tau_0}$ $(f\in B^{\alpha_0})$, we have $\mathbf{b}(f)-\mathbf{a}\in\ell_{\tau_0}$. We define $g\overset{L_p}{:=}\sum_{\theta\in\Theta}a_\theta\varphi_\theta$. Then by Theorem 2.9, g is well defined, and hence

$$f-g\overset{L_p}{=}\sum_{\theta\in\Theta}(b_\theta(f)-a_\theta)\varphi_\theta.$$

By (2.33) and Theorem 2.10, we infer

$$\|g\|_{B^{\alpha_1}}\le c\|\mathbf{a}\|_{\ell_{\tau_1}} \quad\text{and}\quad \|f-g\|_{B^{\alpha_0}}\le c\|\mathbf{b}(f)-\mathbf{a}\|_{\ell_{\tau_0}}.$$

Since $\mathbf{a}\in\ell_{\tau_1}$ is arbitrary, the last two estimates give (2.61).

We next prove that

$$K(\mathbf{b}(f),t;\ell_{\tau_0},\ell_{\tau_1}) \le cK(f,t;B^{\alpha_0},B^{\alpha_1}), \quad t>0. \tag{2.62}$$

Suppose $g\in B^{\alpha_1}$; then by Theorem 2.11, $g\in B^{\alpha_0}$ $(\alpha_0<\alpha_1)$, and hence $f-g\in B^{\alpha_0}$. We shall show that there exists a sequence $\mathbf{b}(g)=(b_\theta(g))_{\theta\in\Theta}\in\ell_{\tau_1}$ such that

$$g\overset{L_p}{=}\sum_{\theta\in\Theta}b_\theta(g)\varphi_\theta \quad\text{with}\quad \|g\|_{B^{\alpha_1}}\approx\|\mathbf{b}(f)\|_{\ell_{\tau_1}} \tag{2.63}$$

and

$$f-g\overset{L_p}{=}\sum_{\theta\in\Theta}(b_\theta(f)-b_\theta(g))\varphi_\theta \quad\text{with}\quad \|f-g\|_{B^{\alpha_0}}\approx\|\mathbf{b}(f)-\mathbf{b}(g)\|_{\ell_{\tau_0}}. \tag{2.64}$$

Clearly, estimate (2.62) follows by (2.63)–(2.64).

Notice that if $\eta\ge 1$, then $\mathbf{b}(\cdot)$ can be realized as a linear operator, and hence $\mathbf{b}(f-g)=\mathbf{b}(f)-\mathbf{b}(g)$. Therefore, (2.63)–(2.64) are immediate from $g\in B^{\alpha_1}$ and $f-g\in B^{\alpha_0}$.

Suppose $\eta<1$. For $\triangle\in\mathcal{T}$, we let $P_\triangle(f):=P_{\triangle,\eta}(f)\in\Pi_k$ be the polynomial from the definition of $p_{m,\eta}(f)$ in (2.20) ($P_\triangle(f)$ is not unique). Thus $P_\triangle(f)\in\Pi_k$ is such that

$$\|f-P_\triangle(f)\|_{L_\eta(\triangle)}\le cE_k(f,\triangle)_\eta. \tag{2.65}$$

We shall next show that for each $\triangle\in\mathcal{T}$ there exists a polynomial $P_\triangle(g)\in\Pi_k$ such that

$$\|g-P_\triangle(g)\|_{L_\eta(\triangle)}\le cE_k(g,\triangle)_\eta \tag{2.66}$$

and

$$\|f-g-(P_\triangle(f)-P_\triangle(g))\|_{L_\eta(\triangle)}\le cE_k(f-g,\triangle)_\eta. \tag{2.67}$$

We consider two cases.

Case 1. $E(f-g)\le E(g)$, where $E(\cdot):=E_k(\cdot,\triangle)_\eta$. Let $R\in\Pi_k$ be such that

$$\|f-g-R\|=E(f-g), \quad\text{where } \|\cdot\|:=\|\cdot\|_{L_\eta(\triangle)}. \tag{2.68}$$

We define $P_\triangle(g) := P_\triangle(f) - R \in \Pi_k$. Then (2.67) holds, by (2.68). We use (2.65) and (2.68) to obtain

$$\begin{aligned}\|g - P_\triangle(g)\| &\le c\|f - P_\triangle(f)\| + c\|f - g - R\| \le cE(f) + cE(f-g)\\ &\le cE(f-g) + cE(g) + cE(f-g) \le cE(g),\end{aligned}$$

which gives (2.66).

Case 2. $E(g) < E(f-g)$. This time we choose $P_\triangle(g) \in \Pi_k$ so that $\|g - P_\triangle(g)\| = E(g)$. Similarly as above, one can show that

$$\|f - g - (P_\triangle(f) - P_\triangle(g))\| \le cE(f-g).$$

Thus the existence of $P_\triangle(g) \in \Pi_k$ satisfying (2.66) and (2.67) is established.

Using the polynomials $P_\triangle(g)$ from above, we define, for $m \in \mathbb{Z}$,

$$p_{m,\eta}(g) := \sum_{\triangle\in\mathcal{T}_m} \mathbb{1}_\triangle \cdot P_\triangle(g) \quad \text{and} \quad p_{m,\eta}(f-g) := \sum_{\triangle\in\mathcal{T}_m} \mathbb{1}_\triangle \cdot (P_\triangle(f) - P_\triangle(g)).$$

Furthermore, as in (2.21) and (2.34), we define

$$Q_{m,\eta}(g) := Q_m(p_{m,\eta}(g)) \quad \text{and} \quad q_{m,\eta}(g) := Q_{m,\eta}(g) - Q_{m-1,\eta}(g).$$

We define $Q_{m,\eta}(f-g)$ and $q_{m,\eta}(f-g)$ in the same way. Finally, we define $\mathbf{b}(g) = (b_\theta(g))_{\theta\in\Theta}$ and $\mathbf{b}(f-g) = (b_\theta(f-g))_{\theta\in\Theta}$ from

$$q_{m,\eta}(g) =: \sum_{\theta\in\Theta_m} b_\theta(g)\varphi_\theta \quad \text{and} \quad q_{m,\eta}(f-g) =: \sum_{\theta\in\Theta_m} b_\theta(f-g)\varphi_\theta, \quad m \in \mathbb{Z}.$$

Evidently, $p_{m,\eta}(f-g) = p_{m,\eta}(f) - p_{m,\eta}(g)$ and since Q_m is a linear operator, it follows that $\mathbf{b}(f-g) = \mathbf{b}(f) - \mathbf{b}(g)$. From this and the fact that $P_\triangle(g)$ satisfies (2.66) and (2.67), using Theorem 2.8, Theorem 2.10, and (2.39), we obtain that $\mathbf{b}(g)$ satisfies (2.63) and (2.64), and hence (2.62) holds. This completes the proof of the lemma. □

By Lemma 2.13, (2.58), and (2.59) (with α_j replaced by α), we obtain

$$\|f\|_{(B^{\alpha_0},B^{\alpha_1})_{\lambda,\tau}} \approx \|\mathbf{b}(f)\|_{(\ell_{\tau_0},\ell_{\tau_1})_{\lambda,\tau}} \approx \|\mathbf{b}(f)\|_{\ell_\tau} \approx \|f\|_{B^\alpha}.$$

Thus the proof of Theorem 2.12 is complete. □

Several remarks are in order.

Remark 2.14. (a) If $p = \infty$, then the B-space $B^\alpha_\tau(\Phi)$ ($\tau := 1/\alpha$) is useful for our goals only if $\alpha \geq 1$. The reason for this is that $B^\alpha_\tau(\Phi)$ is not embedded in C if $\alpha < 1$. Indeed, consider the function $f := \sum_{j=1}^\infty j^{-1}\varphi_{\theta_j}$, where $\theta_j \in \Theta_{m_j}$, $m_1 < m_2 < \cdots$, and $\{\varphi_{\theta_j}\}$ are Courant (or other) elements which overlap so that $\|f\|_\infty \approx \sum_{j=1}^\infty j^{-1} = \infty$. On the other hand (see (2.33)), $|f|_{B^\alpha_\tau(\Phi)} \leq c(\sum_{j=1}^\infty j^{-\tau})^{1/\tau} < \infty$, since $\tau := 1/\alpha > 1$.

(b) We introduced the B-norms $N_{\Phi,\mathbb{S},\eta}(\cdot)$ and $N_{\Phi,Q,\eta}(\cdot)$ with $0 < \eta < p$ (see (2.31) and (2.37)) for the following reason. As we shall see in section 3, normally $\alpha > 1$, and hence $\tau < 1$, which compels us to work in L_τ with $\tau < 1$, which is not a very friendly space. At the same time, if $p > 1$ we can choose $1 \leq \eta < p$ and work in L_η instead.

(c) We also want to explain why we introduce the B-spaces over locally regular (or better) triangulations but not over more general ones. The reason is that if we relax the main conditions (2.1)–(2.2) in the definition of LR-triangulations, then we can hardly work with the B-spaces. In particular, the equivalence of the norms (see Theorem 2.10) fails to exist, which makes it impossible to prove all the results from section 3.

General B-spaces. Given an LR(or better)-triangulation $\mathcal{T}$ and a family of basis functions $\Phi = \Phi_\mathcal{T}$ over $\mathcal{T}$ as in section 2.2, we define the more general B-space $B^\alpha_{pq}(\Phi) = B^\alpha_{pq}(\mathcal{S})$, $\alpha > 0$, $0 < p, q \leq \infty$, as the set of all $f \in L_p(\mathbb{R}^2)$ such that

$$\|f\|_{B^\alpha_{pq}(\Phi)} := \|f\|_p + \left(\sum_{m\in\mathbb{Z}} \left[2^{m\alpha} \left(\sum_{\triangle\in\mathcal{T},\, 2^{-m}\leq|\triangle|<2^{-m+1}} \mathbb{S}_\triangle(f)_p^p \right)^{1/p} \right]^q \right)^{1/q} < \infty,$$

with the ℓ_q-norm replaced by the sup-norm if $q = \infty$, where $\mathbb{S}_\triangle(f)_p$ is as above (see (2.22)). Evidently, $B^\alpha_p(\Phi) = B^\alpha_{pp}(\Phi)$. In going further, the norms in $B^\alpha_\tau(\Phi)$ from (2.31), (2.32), and (2.37) can be generalized accordingly. In the present article, we do not explore the B-spaces in such generality because the space scale $B^\alpha_\tau(\Phi)$ is sufficient for our goal of characterizing the approximation rates of nonlinear n-term approximation from differentiable piecewise polynomials.

Fat B-spaces: The link to Besov spaces. Suppose $\mathcal{T}$ is an arbitrary SLR-triangulation of $\mathbb{R}^2$. The *fat B-space* $\mathbb{B}^{\alpha k}_\tau(\mathcal{T})$ with $k \geq 1$ and α, τ as in the definition of $B^\alpha_\tau(\mathcal{T})$ (section 2.3) is defined (see [38]) as the set of all functions $f \in L_\tau(\mathbb{R}^2)$ such that

$$\|f\|_{\mathbb{B}^{\alpha k}_\tau(\mathcal{T})} := \left(\sum_{\triangle\in\mathcal{T}} [|\triangle|^{-\alpha} E_k(f, \Omega_\triangle)_\tau]^\tau \right)^{1/\tau} \approx \left(\sum_{\triangle\in\mathcal{T}} [|\triangle|^{-\alpha} \omega_k(f, \Omega_\triangle)_\tau]^\tau \right)^{1/\tau} < \infty,$$

where $E_k(f, \Omega_\triangle)_\tau$ is the error of L_τ-approximation to f on $\Omega_\triangle := \Omega^1_\triangle$ from Π_k and $\omega_k(f, \Omega_\triangle)_\tau$ is the local L_τ-modulus of smoothness of f on $\Omega_\triangle$. (Recall that $E_k(f, \Omega_\triangle)_\tau \approx \omega_k(f, \Omega_\triangle)_\tau$ by Whitney's theorem (2.16), since $\mathcal{T}$ is an SLR-triangulation.) Furthermore, other equivalent norms in $\mathbb{B}^{\alpha k}_\tau(\mathcal{T})$ as well as more general fat B-spaces $\mathbb{B}^{\alpha k}_{pq}(\mathcal{T})$ can be defined as in [38].

Suppose that $\Phi = \Phi_\mathcal{T}$ is a hierarchical family of basis functions over $\mathcal{T}$ as described in section 2.2. Assuming $\Pi_k \subset \mathcal{S}_m \subset \mathcal{S}^{k,r}_m(\mathcal{T})$ for all $m \in \mathbb{Z}$ (that is, $\tilde{k} = k$ in the notation of section 2.2), we have for $f \in L_\tau$ and $\triangle \in \mathcal{T}_m$,

$$E_k(f, \Omega^\ell_\triangle)_\tau \leq c \sum_{\triangle'\in\mathcal{T}_m,\, \triangle'\subset\Omega^\ell_\triangle} E_k(f, \Omega_{\triangle'})_\tau,$$

which implies $\|f\|_{B^\alpha_\tau(\Phi_\mathcal{T})} \le c\|f\|_{\mathbb{B}^{\alpha k}_\tau(\mathcal{T})}$. Therefore, the space $\mathbb{B}^{\alpha k}_\tau(\mathcal{T})$ is a good candidate to replace $B^\alpha_\tau(\Phi)$ in nonlinear spline approximation, but this is only possible if $0 < \alpha < \alpha_0$ for some $\alpha_0 < \infty$, which we do not compute here. The problem with the space $\mathbb{B}^{\alpha k}_\tau(\mathcal{T})$ is that $\|\varphi_\theta\|_{\mathbb{B}^{\alpha k}_\tau(\mathcal{T})} < \infty$ only for $0 < \alpha < \alpha_0$. (See Theorem 2.15 in the case of regular triangulations.) Therefore, the basic norm equivalence results (Theorem 2.10) hold only for a restricted range of α. Thus, $\mathbb{B}^{\alpha k}_\tau(\mathcal{T})$ is simply not the "right" space for the specific problem at hand if $\alpha \ge \alpha_0$. It is too "fat." However, the spaces $\mathbb{B}^{\alpha k}_\tau(\mathcal{T})$ are still noteworthy since they are less sensitive to small perturbations of the triangulation $\mathcal{T}$ and are technically easier. We believe that a situation will present itself when they will be the "right" spaces.

Comparison between regular B-spaces and Besov spaces. We begin by recalling the definition of the classical Besov space by moduli of smoothness. So, the space $B^s_q(L_p) := B^s_q(L_p(\mathbb{R}^2))$, $s > 0$, $1 \le p, q \le \infty$, is defined as the set of all functions $f \in L_p(\mathbb{R}^2)$ such that

$$\|f\|_{B^s_q(L_p)} := \left(\int_0^\infty (t^{-s}\omega_k(f,t)_p)^q \frac{dt}{t} \right)^{1/q} < \infty \tag{2.69}$$

($\|f\|_p$ is usually added to the right-hand side above), where $k := [s]+1$ and $\omega_k(f,t)_p$ is the kth modulus of smoothness of f in $L_p(\mathbb{R}^2)$, i.e., $\omega_k(f,t)_p := \sup_{|h|\le t} \|\Delta^k_h(f,\cdot)\|_p$. It is well known that whenever $1 \le p \le \infty$, if in (2.69) k is replaced by any other $k > s$, then the resulting space would be the same with an equivalent norm. However, the situation is totally different when $p < 1$, and this is a reason for introducing k as a parameter of the Besov spaces in the following.

As elsewhere, let us assume that $0 < p < \infty$ and $\alpha > 0$, or $p = \infty$ and $\alpha \ge 1$, and in both cases $1/\tau := \alpha + 1/p$. Let $k \ge 1$. We define the space $B^{2\alpha,k}_\tau(L_\tau)$ as the Besov space $B^{2\alpha}_\tau(L_\tau)$ (see (2.69)), where k and α are independent of each other. These are the spaces that naturally occur in nonlinear spline approximation (see [53]).

Suppose that $\mathcal{T}^*$ is a regular triangulation of $\mathbb{R}^2$ (see section 2.1). Then as shown in [38], $\mathbb{B}^{\alpha k}_\tau(\mathcal{T}^*) = B^{2\alpha,k}_\tau(L_\tau)$ with equivalent norms. (Notice that the smoothness parameters of B-spaces and Besov spaces are normalized differently and α corresponds to 2α.)

Let us now assume that $\Phi_{\mathcal{T}^*} = \{\varphi_\theta\}$ is a family of basis functions over $\mathcal{T}^*$ as in section 2.2 such that $\Pi_k \subset \mathcal{S}_m \subset \mathcal{S}^{k,r}_m$ ($m \in \mathbb{Z}$), where $r \ge 0$ and $k > r$. As we mentioned above, the fat B-space $\mathbb{B}^{\alpha k}_\tau(\mathcal{T}^*)$, and hence the Besov space $B^{2\alpha,k}_\tau(L_\tau)$, is a good candidate to replace the B-space $B^\alpha_\tau(\Phi_{\mathcal{T}^*})$ in nonlinear n-term approximation from $\Phi_{\mathcal{T}^*}$. We next spell out the exact conditions for equivalence of the corresponding norms.

THEOREM 2.15. *Under the above assumptions, if* $0 < \alpha < r + 1 + 1/p$, *then*

$$B^{2\alpha,k}_\tau(L_\tau) = B^\alpha_\tau(\Phi_{\mathcal{T}^*}) \tag{2.70}$$

with equivalent norms. Furthermore, if a single basis function $\varphi_\theta \in \Phi_{\mathcal{T}^*}$ *does not belong to* C^{r+1}, *then the equivalence is no longer true when* $\alpha \ge r + 1 + 1/p$. *More precisely, for such* φ_θ *and* α, $\|\varphi_\theta\|_{B^{2\alpha,k}_\tau(L_\tau)} = \infty$, *while* $\|\varphi_\theta\|_{B^\alpha_\tau(\Phi_{\mathcal{T}^*})} \approx \|\varphi_\theta\|_p$.

Proof. As we mentioned before, $\|f\|_{B^\alpha_\tau(\Phi_{\mathcal{T}^*})} \le c\|f\|_{\mathbb{B}^{\alpha k}_\tau(\mathcal{T}^*)}$ for $f \in \mathbb{B}^{\alpha k}_\tau(\mathcal{T}^*)$, and also we have $\|f\|_{\mathbb{B}^{\alpha k}_\tau(\mathcal{T}^*)} \approx \|f\|_{B^{2\alpha,k}_\tau(L_\tau)}$, exactly as in Theorem 2.25 from [38]. Therefore,

$$\|f\|_{B^\alpha_\tau(\Phi_{\mathcal{T}^*})} \le c\|f\|_{B^{2\alpha,k}_\tau(L_\tau)} \text{ for } f \in B^{2\alpha,k}_\tau(L_\tau).$$

The proof of the reverse estimate follows in the footsteps of the proof of Theorem 2.28 from [38], and we shall indicate only the differences. Using the conditions on $\Phi_{\mathcal{T}^*}$ and the fact that $\mathcal{T}^*$ is regular, one can show by straightforward calculations that, for each $\theta \in \Theta(\mathcal{T}^*)$,

$$\omega_k(\varphi_\theta, t)_\tau^\tau \le \begin{cases} c|E_\theta|^{\frac{1}{2}(1-(r+1)\tau)} \cdot t^{1+(r+1)\tau} & \text{if } 0 < t < |E_\theta|^{1/2}, \\ c|E_\theta| & \text{if } t \ge |E_\theta|^{1/2}. \end{cases} \tag{2.71}$$

Moreover, both sides of (2.71) are equivalent if φ_θ does not belong to C^{r+1}. In going further, one uses (2.71) exactly as in [38] to complete the proof of the theorem. ∎

Remark 2.16. An interesting situation occurs when $p = \infty$ and $r = 0$. Then there is no α for which (2.70) holds. This is the case when $\Phi_{\mathcal{T}^*}$ is the set of all Courant elements generated by $\mathcal{T}^*$ (a regular triangulation).

Comparison between different B-spaces and Besov spaces. Suppose $\Phi_{\mathcal{T}}$ is a family of basis functions associated with an SLR-triangulation $\mathcal{T}$ which allows arbitrarily sharp angles. Then some extremely "skinny" basis functions $\varphi_\theta \in \Phi_{\mathcal{T}}$ (with elongated level curves) will occur. It is easily seen that such functions have huge Besov norms (see [38]) compared to their L_p-norms as well as their $B(\Phi_{\mathcal{T}})$-norms (see Theorem 3.2 below) for any smoothness $\alpha > 0$. Therefore, the B-spaces for such a triangulation are essentially different from Besov spaces. The situation is quite similar when comparing two B-spaces over different triangulations. Therefore, the B-spaces change substantially with the triangulations, thus making the search for the "right" triangulation mentioned in the introduction a meaningful task. In contrast to this, the standard Besov spaces can be used only to characterize the approximation power of piecewise polynomials over *regular* triangulations.

B-spaces over compact domains. B-spaces can be introduced on an arbitrary compact polygonal domain $E \subset \mathbb{R}^2$. A substantial difference would be in assuming that each triangulation $\mathcal{T}$ of E is of the form $T = \bigcup_{m=0}^{\infty} \mathcal{T}_m$, where T_0 is an initial level (triangulation of E) and $\mathcal{T}_1, \mathcal{T}_2, \dots$ are consecutive refinements of $\mathcal{T}_0$. This approach is important for the applications (see [39]).

B-spaces in dimensions $d > 2$. Multilevel triangulations and B-spaces can be introduced in much the same way in dimensions $d > 2$. Of course, then the triangles should be replaced by simplices, thus making some geometric argumentation of this section essentially more involved. In particular, the property (e) of a multilevel triangulation should be extended to all faces of the simplices in $\mathcal{T}_m$, thus saying that there are at most N_0 simplices in $\mathcal{T}_m$ attached to a particular face. The "no hanging vertices" condition (d) should be replaced by the condition that each facet of a simplex in $\mathcal{T}_m$ is a common facet of exactly two simplices in $\mathcal{T}_m$. The minimal angle condition appearing in the definition of regular triangulations and in (2.4) should be replaced by the *shape regularity* condition that postulates the existence of an upper bound on the ratio of the diameter of a simplex and the diameter of the inscribed sphere. In conditions (2.1)–(2.3) the area should be replaced by the d-dimensional volume.

B-spaces in dimension $d = 1$. B-spaces can be introduced in the univariate case, but none will give anything new, and hence they are not needed. The key fact is that, in the univariate case, the Bernstein inequality involving Besov spaces holds with no restrictions on the smoothness parameter $\alpha < \infty$ (see [53]).

In a nutshell, the essence of the spaces we considered in this section is the following. The Besov spaces are based on local polynomial approximation over regular multilevel triangulations, which is explicitly shown in [38]. When the regular triangulations are replaced by SLR-triangulations, then the Besov spaces become fat

B-spaces, which further evolve to B-spaces when the local polynomial approximation is replaced by local spline approximation.

The B-spaces are closely related to certain anisotropic maximal functions, nonclassical differentiability, and other problems, which are beyond the scope of this article.

3. Nonlinear n-term spline approximation. In this section, we assume that $\mathcal{T}$ is a locally regular (or better) triangulation of $\mathbb{R}^2$. Also, we assume that $\Phi = \Phi_{\mathcal{T}}$ is a hierarchical family of basis functions over $\mathcal{T}$ (see section 2.2). Notice that Φ is not a basis; Φ is redundant. We consider nonlinear n-term approximation from Φ in $L_p(\mathbb{R}^2)$ $(0 < p \le \infty)$, where we identify $L_\infty(\mathbb{R}^2)$ as $C_0(\mathbb{R}^2)$. We let $\Sigma_n(\Phi)$ denote the nonlinear set consisting of all splines s of the form

$$s = \sum_{\theta \in \mathcal{M}} a_\theta \varphi_\theta,$$

where $\mathcal{M} \subset \Theta(\mathcal{T})$, $\#\mathcal{M} \le n$, and $\mathcal{M}$ may vary with s. We denote by $\sigma_n(f,\Phi)_p$ the error of L_p-approximation to $f \in L_p(\mathbb{R}^2)$ from $\Sigma_n(\Phi)$:

$$\sigma_n(f,\Phi)_p := \inf_{s \in \Sigma_n(\Phi)} \|f - s\|_p.$$

Our goal is to characterize the approximation spaces generated by nonlinear n-term approximation from Φ. To this end we next prove a pair of companion Jackson and Bernstein estimates. We shall utilize the B-spaces $B^\alpha_\tau(\Phi)$ introduced in section 2. We assume that $0 < p < \infty$ and $\alpha > 0$, or $p = \infty$ and $\alpha \ge 1$. In both cases, $1/\tau := \alpha + 1/p$ $(1/\infty := 0)$.

THEOREM 3.1 (Jackson estimate). *If $f \in B^\alpha_\tau(\Phi)$, then*

$$\sigma_n(f,\Phi)_p \le c n^{-\alpha} \|f\|_{B^\alpha_\tau(\Phi)}, \tag{3.1}$$

with c independent of f and n.

In the case $0 < p < \infty$, this theorem follows by the general Theorem 3.4 from [38], in view of the results of section 2. For completeness, we shall give its short proof in the appendix. The proof when $p = \infty$ can be carried out as the proof of Theorem 4.1 from [39] but is a little longer, and so we shall skip it.

THEOREM 3.2 (Bernstein estimate). *If $s \in \Sigma_n(\Phi)$, then*

$$\|s\|_{B^\alpha_\tau(\Phi)} \le c n^\alpha \|s\|_p, \tag{3.2}$$

with c independent of s and n.

The proof of this (vital for our development) theorem utilizes the ideas of the proofs of Theorem 3.6 from [38] $(0 < p < \infty)$ and Theorem 4.2 from [39] $(p = \infty)$ but is not identical to them. We shall give the proof in the appendix.

For a fixed $\mathcal{T}$ and $\Phi := \Phi_{\mathcal{T}}$, we set $K(f,t) := K(f,t;L_p,B^\alpha_\tau(\Phi))$ ($L_p := C_0$ if $p = \infty$); see (2.55). The Jackson and Bernstein estimates from Theorems 3.1 and 3.2 imply in a standard way (see, e.g., [55]) the following direct and inverse estimates: For any $\alpha > 0$, if $f \in L_p$, then

$$\sigma_n(f,\Phi)_p \le c K(f, n^{-\alpha}) \tag{3.3}$$

and

$$K(f,n^{-\alpha}) \le c n^{-\alpha} \left(\|f\|_p + \left[\sum_{\nu=1}^{n} \frac{1}{\nu} (\nu^\alpha \sigma_\nu(f,\Phi)_p)^\mu \right]^{1/\mu} \right), \tag{3.4}$$

where $\mu := \min\{p,1\}$ and c is independent of f and n.

An immediate consequence of (3.3) and (3.4) is that $\sigma_n(f,\Phi)_p = O(n^{-\gamma})$, $0<\gamma<\alpha$, if and only if $K(f,n^{-\alpha}) = O(n^{-\gamma})$. More generally, these estimates enable us to characterize the approximation spaces generated by nonlinear n-term approximation from Φ. We define the approximation space $A_q^\gamma := A_q^\gamma(\Phi, L_p)$, $\alpha>0$, $0<q\le\infty$, as the set of all functions $f\in L_p$ such that

$$\|f\|_{A_q^\gamma} := \|f\|_p + \left(\sum_{n=1}^{\infty}(n^\gamma \sigma_n(f,\Phi)_p)^q \frac{1}{n}\right)^{1/q} < \infty$$

with the ℓ_q-norm replaced by the sup-norm if $q=\infty$ as usual.

The direct and inverse estimates (3.3)–(3.4) readily imply (see, e.g., [55]) the following characterization of the approximation spaces.

THEOREM 3.3. *If $0<\gamma<\alpha$ and $0<q\le\infty$, then*

$$A_q^\gamma(\Phi, L_p) = (L_p, B_\tau^\alpha(\Phi))_{\frac{\gamma}{\alpha},q}$$

with equivalent norms.

In one specific case the interpolation spaces can be identified as B-spaces.

THEOREM 3.4. *Suppose $0<p<\infty$ and $\alpha>0$, or $p=\infty$ and $\alpha>1$, and let $\tau := (\alpha+1/p)^{-1}$. Then*

$$A_\tau^\alpha(\Phi, L_p) = B_\tau^\alpha(\Phi) \tag{3.5}$$

with equivalent norms.

The following interpolation result is immediate from Theorems 3.3 and 3.4.

COROLLARY 3.5. *Suppose p, α, and $\tau =: \tau(\alpha)$ are as in the hypothesis of Theorem 3.4, and let $\beta>\alpha$ and $\tau(\beta) := (\beta+1/p)^{-1}$. Then*

$$(L_p, B_{\tau(\beta)}^\beta(\Phi))_{\frac{\alpha}{\beta},\tau(\alpha)} = B_{\tau(\alpha)}^\alpha(\Phi) \tag{3.6}$$

with equivalent norms.

Proof of Theorem 3.4. We shall employ the idea of the proof of Theorem 3.3 in [30]. We shall use abbreviated notation: $A_q^\alpha := A_q^\alpha(\Phi, L_p)$, $B_\tau^\alpha := B_\tau^\alpha(\Phi)$, and the like. For any $\beta>0$, we denote $\tau(\beta) := (\beta+1/p)^{-1}$.

We first prove the following continuous embedding:

$$A_\mu^\beta \subset B_{\tau(\beta)}^\beta \quad \text{with} \quad \mu := \min\{\tau(\beta), 1\}. \tag{3.7}$$

Indeed, suppose $f\in A_\mu^\beta$, and let $s_m\in\Sigma_m$ be such that

$$\|f-s_m\|_p \le 2\sigma_m(f)_p. \tag{3.8}$$

Since $\sigma_m(f)_p\to 0$, we have $f = s_1 + \sum_{\nu=1}^\infty (s_{2^\nu} - s_{2^{\nu-1}})$ in L_p (uniformly if $p=\infty$), and hence ($\mu\le 1$)

$$\|f\|_{B_{\tau(\beta)}^\beta}^\mu \le \|s_1\|_{B_{\tau(\beta)}^\beta}^\mu + \sum_{\nu=1}^\infty \|s_{2^\nu} - s_{2^{\nu-1}}\|_{B_{\tau(\beta)}^\beta}^\mu. \tag{3.9}$$

We apply the Bernstein estimate from Theorem 3.2 to $s_{2^\nu} - s_{2^{\nu-1}} \in \Sigma_{2^{\nu+1}}$ and use (3.8) to obtain

$$\|s_{2^\nu} - s_{2^{\nu-1}}\|_{B_{\tau(\beta)}^\beta} \le c2^{\nu\beta}\|s_{2^\nu} - s_{2^{\nu-1}}\|_p \le c2^{\nu\beta}(\sigma_{2^\nu}(f)_p + \sigma_{2^{\nu-1}}(f)_p),$$

and similarly $\|s_1\|_{B^\beta_{\tau(\beta)}} \le c(\|f\|_p + \sigma_1(f)_p)$. Using these in (3.9) implies

$$\|f\|^\mu_{B^\beta_{\tau(\beta)}} \le c\|f\|^\mu_p + c\sum_{\nu=1}^\infty (2^{\nu\beta}\sigma_{2^\nu}(f)_p)^\mu \le c\|f\|^\mu_{A^\beta_\mu},$$

which is (3.7).

Second, the Jackson estimate from Theorem 3.1 gives the continuous embedding

$$B^\beta_{\tau(\beta)} \subset A^\beta_\infty. \tag{3.10}$$

A third important ingredient in this proof is the fact that the approximation spaces A^α_q are invariant under interpolation (see [31, 52]): If $\alpha_0, \alpha_1 > 0$ and $0 < q_1, q_2, q \le \infty$, then

$$(A^{\alpha_0}_{q_0}, A^{\alpha_1}_{q_1})_{\lambda,q} = A^\alpha_q, \quad \text{where } \alpha = (1-\lambda)\alpha_0 + \lambda\alpha_1 \text{ with } 0 < \lambda < 1. \tag{3.11}$$

Now we choose α_0 and α_1 so that $0 < \alpha_0 < \alpha < \alpha_1$ ($\alpha_0 := 1$ if $p = \infty$). Also, we select $0 < \lambda < 1$ so that $\alpha = (1-\lambda)\alpha_0 + \lambda\alpha_1$. Furthermore, we set $\tau_j := (\alpha_j + 1/p)^{-1}$ and $\mu_j := \min\{\tau_j, 1\}$, $j = 0, 1$. By Theorem 2.12, we have

$$(B^{\alpha_0}_{\tau_0}, B^{\alpha_1}_{\tau_1})_{\lambda,\tau} = B^\alpha_\tau.$$

We use this, (3.7), (3.10), and (3.11) to obtain the following continuous embeddings:

$$A^\alpha_\tau = (A^{\alpha_0}_{\mu_0}, A^{\alpha_1}_{\mu_1})_{\lambda,\tau} \subset B^\alpha_\tau = (B^{\alpha_0}_{\tau_0}, B^{\alpha_1}_{\tau_1})_{\lambda,\tau} \subset (A^{\alpha_0}_\infty, A^{\alpha_1}_\infty)_{\lambda,\tau} = A^\alpha_\tau,$$

which give (3.5). □

Algorithms. In [39], there are three algorithms developed for n-term Courant element approximation in L_p ($0 < p \le \infty$). These algorithms can be immediately adapted to nonlinear n-term approximation from any family of differentiable spline basis functions $\Phi_\mathcal{T}$ on a compact polygonal domain $E \subset \mathbb{R}^2$. It is an integral part of our program that using the machinery of the B-spaces, Jackson and Bernstein estimates, interpolation, etc. developed in this article, we can prove that these algorithms achieve the rate of the best n-term approximation. This aspect of our theory will not be elaborated on here (see [39]).

Approximation from the libraries $\{\Phi_\mathcal{T}\}_\mathcal{T}$. An important element of our concept for nonlinear spline approximation is the introduction of another level of nonlinearity by allowing the triangulation $\mathcal{T}$ to vary. For a given SRL(or LR)-triangulation $\mathcal{T}$, let $\Phi_\mathcal{T}$ be a family of spline basis functions like the ones considered in section 2.2. Now, without changing the nature of the basis elements from $\Phi_\mathcal{T}$, we let $\mathcal{T}$ vary and obtain a collection (library) of basis families $\{\Phi_\mathcal{T}\}_\mathcal{T}$. We denote

$$\sigma_n(f)_p := \inf_\mathcal{T} \sigma_n(f, \Phi_\mathcal{T})_p,$$

where the infimum is taken over all SLR-triangulations $\mathcal{T}$ with fixed parameters, and we also assume that the parameters of $\Phi_\mathcal{T}$ are fixed. The following theorem is immediate from the Jackson estimate in Theorem 3.1. We shall assume again that $0 < p < \infty$ and $\alpha > 0$, or $p = \infty$ and $\alpha \ge 1$, and in both cases, $1/\tau := \alpha + 1/p$.

THEOREM 3.6. *If* $\inf_\mathcal{T} \|f\|_{B^\alpha_\tau(\Phi_\mathcal{T})} < \infty$, *then*

$$\sigma_n(f)_p \le cn^{-\alpha} \inf_\mathcal{T} \|f\|_{B^\alpha_\tau(\Phi_\mathcal{T})}$$

with c depending only on p, α, and the parameters of $\mathcal{T}$ and $\Phi_\mathcal{T}$.

The ultimate *open problem* here is to characterize the approximation spaces generated by $\{\sigma_n(f)_p\}$ for a given library of basis functions $\{\Phi_{\mathcal{T}}\}_{\mathcal{T}}$.

Global smoothness of functions: How to measure it? Here we come to one of the fundamental questions in approximation theory (and not only there) of how the global smoothness of the functions should be measured.

In the case of nonlinear n-term L_p-approximation from a *single* basis family $\Phi_{\mathcal{T}}$, a function f should be considered of smoothness $\alpha > 0$ if $\|f\|_{B^\alpha_\tau(\Phi_{\mathcal{T}})} < \infty$. Then the rate of n-term L_p-approximation of f from $\Phi_{\mathcal{T}}$ is $O(n^{-\alpha})$ (roughly). If we consider nonlinear n-term approximation from a given *library* of basis families $\{\Phi_{\mathcal{T}}\}_{\mathcal{T}}$ ($\mathcal{T}$ is allowed to vary), then a function f should naturally be considered of smoothness $\alpha > 0$ if $\inf_{\mathcal{T}} \|f\|_{B^\alpha_\tau(\Phi_{\mathcal{T}})} < \infty$, which means that there exists a triangulation $\mathcal{T} := \mathcal{T}_f$ such that $\|f\|_{B^\alpha_\tau(\Phi_{\mathcal{T}})} < \infty$. Then the rate of n-term L_p-approximation of f from the library $\{\Phi_{\mathcal{T}}\}_{\mathcal{T}}$ is $O(n^{-\alpha})$. It is crystal clear to us that no single (super) space can do the job in this case. It is an *open problem* to develop an algorithm for finding, for a given function f, an optimal (or near optimal) triangulation, i.e., a triangulation $\mathcal{T}_f$ for which f exhibits maximal (near maximal) smoothness, using the space scale $B^\alpha_\tau(\Phi_{\mathcal{T}_f})$. It is also an *open problem* whether, for a given function $f \in L_p$, there exists a single triangulation $\mathcal{T}_f$ such that, for all $n \geq 1$, the n-term L_p-approximation of f from the library $\{\Phi_{\mathcal{T}}\}_{\mathcal{T}}$ can be realized by n-term approximation from $\Phi_{\mathcal{T}_f}$ and, consequently, characterized by the B-spaces $B^\alpha_\tau(\Phi_{\mathcal{T}_f})$ via interpolation.

Another important related issue for discussion is the smoothness of the approximating tool $\Phi_{\mathcal{T}} := \{\varphi_\theta\}$ ($\mathcal{T}$ fixed). Clearly, in nonlinear approximation, there is no saturation, which means that the corresponding approximation spaces A^γ_q are nontrivial for all $0 < \gamma < \infty$. Therefore, the smoothness spaces to be used should naturally be designed so that the functions $\{\varphi_\theta\}$ are infinitely smooth with respect to these spaces. This has been one of the guiding principles to us in constructing the B-spaces. Thus each basis function $\varphi_\theta \in \Phi$ is infinitely smooth with respect to the scale of B-spaces $B^\alpha_\tau(\Phi)$, which is reflected in the fact that $\|\varphi_\theta\|_{B^\alpha_\tau(\Phi)} \leq c\|\varphi_\theta\|_p$ for $0 < \alpha < \infty$ (see Theorem 3.2). This makes it possible that in our direct, inverse, and characterization theorems we impose no restrictions on the rate of approximation $\alpha < \infty$ (see Theorems 3.1–3.4). Also, this explains the complete success of Besov spaces in univariate nonlinear spline approximation (see [53]) and why Besov spaces are not quite suitable in dimensions $d > 1$. The latter remark needs a few words of explanation: First, by allowing triangulations with arbitrarily sharp angles, we allow very "skinny" basis functions with huge Besov norms compared to their L_p-norms (see [38]), which precludes the use of Besov spaces in such situations. Second, even when working on regular triangulations, the use of Besov spaces is restricted by the Besov smoothness (regularity) of the basis functions (see Theorem 2.15), while B-spaces impose no restrictions on the rates of approximation.

Spline wavelets (prewavelets) and frames. In the case of uniform triangulations, spline wavelets play an essential role in practical algorithms. It would be desirable to have compactly supported wavelet (prewavelet) bases or frames generated by (differentiable) spline basis families $\Phi_{\mathcal{T}}$ over LR- or SLR-triangulations $\mathcal{T}$. To our knowledge there are no constructions of this type available, as for now. Moreover, there is some evidence that such constructions would be too complicated and impractical for general triangulations. However, continuous spline prewavelets on regular triangulations with uniform dyadic refinements are available from [21, 34, 58]. (See also [47].) Evidently, nonlinear n-term approximation from compactly supported spline wavelets or frames, generated by Courant elements or a smoother spline basis

family $\Phi_{\mathcal{T}}$, cannot give a better rate of convergence than nonlinear n-term approximation from $\Phi_{\mathcal{T}}$. We hope that efficient algorithms for n-term approximation from such families may provide a substitute for wavelet methods in situations where the latter are difficult to apply and, in particular, for approximation in L_∞.

Adaptive tree approximation. This is a method for nonlinear approximation from piecewise polynomials on (single level) triangular partitions, which has been developed recently in [5, 7]. In [5], algorithms are developed which achieve the rate of the best adaptive tree approximation, while in [7] the rates of approximation are related to the smoothness of the functions in terms of Besov spaces. There are substantial distinctions between this approach and the one in the present article. Namely, the approximation schemes from [5, 7] use "single level" piecewise polynomials on triangulations which satisfy the minimal angle condition, while here we use multilevel (multiscale) piecewise polynomial bases over triangulations which allow arbitrarily sharp angles. Therefore, the notion of "best approximation" in [5, 7] is quite different from the one used here. Substantial progress has been made in [6] in applying the adaptive tree approximation method for numerical (finite element) solution of PDEs.

4. Construction of differentiable basis functions. In this section, we give, for any SLR-triangulation, a construction of differentiable spline basis in $\mathcal{S}_m^{k,r}$, $r \ge 1$, $k > 4r+1$, satisfying the conditions from section 2.2. In general, we follow the scheme of [22]; however, appropriate modifications in the construction and in the proofs have to be made since we do not assume that the triangulation is regular. In particular, we replace the standard normal derivatives to the edges by derivatives in affine invariant directions; see the definition of $D_{\mu(e,\triangle)}$ below. Since our construction is also applicable to nonnested triangulations (see Remark 4.8), we formulate the results here for a *fixed* level $\mathcal{T}_m$ assuming only conditions (a), (d)–(f), and (2.3) of section 2.1 and making sure that the constants in (2.11) and (2.12) depend only on k, r, N_0, and δ_2.

4.1. Nodal functionals. As before, let $\mathcal{V}_m$ and $\mathcal{E}_m$ be the sets of all vertices and all edges of $\mathcal{T}_m$, respectively. We shall describe the basis functions for $\mathcal{S}_m = \mathcal{S}^{k,r}(\mathcal{T}_m)$, $k > 4r+1$, with the aid of the so-called *nodal functionals* defined on $\mathcal{S}^{k,r}(\mathcal{T}_m)$. These are certain linear functionals involving the values of the splines and their derivatives at specific points in $\mathbb{R}^2$. The functional corresponding to the simple evaluation of the splines at $\xi \in \mathbb{R}^2$ will be denoted by δ_ξ.

Of particular interest as evaluation points are the vertices $v \in \mathcal{V}_m$, where we also need the derivative evaluation functionals of type $\delta_v D_e^\alpha$ with e being any edge in $\mathcal{E}_m$ emanating from v, and $\delta_v D_{e_1}^\alpha D_{e_2}^\beta$, where e_1, e_2 are adjacent edges emanating from v. Here $D_{[v,\tilde v]}^\alpha s$ denotes the derivative of s of order α in the direction of the interval $[v, \tilde v]$, weighted with the length of $[v, \tilde v]$, namely,

$$D_{[v,\tilde v]}^\alpha s := \Big((\tilde v_x - v_x)D_x + (\tilde v_y - v_y)D_y\Big)^\alpha s,$$
$$v = (v_x, v_y), \quad \tilde v = (\tilde v_x, \tilde v_y).$$

Note that, due to this weighting, the corresponding *Markov inequality* reads as follows:

$$\|D_{[v,\tilde v]}^\alpha p\|_{L_\infty[v,\tilde v]} \le c\|p\|_{L_\infty[v,\tilde v]}, \qquad p \in \Pi_k, \tag{4.1}$$

where c depends only on k and α.

Let $\triangle_1, \triangle_2 \in \mathcal{T}_m$ share an edge e. Since every $s \in \mathcal{S}^{k,r}(\mathcal{T}_m)$ is continuous, the two polynomial patches $s|_{\triangle_1}$ and $s|_{\triangle_2}$ coincide along e. Therefore, $\delta_v D_e^\alpha s$ may be computed for any $\alpha = 0, 1, \dots$ as either $\delta_v D_e^\alpha (s|_{\triangle_1})$ or $\delta_v D_e^\alpha (s|_{\triangle_2})$ with the same

result. Similarly, let $e_1, e_2 \in \mathcal{E}_m$ be two edges of a triangle $\triangle \in \mathcal{T}_m$ with a common vertex v. Then $\delta_v D_{e_1}^\alpha D_{e_2}^\beta s$ denotes the mixed derivative of s at v in the directions of e_1 and e_2 away from v. If $\alpha + \beta \le r$, this derivative is uniquely defined. If $\alpha + \beta > r$, the result may depend on the choice of the polynomial patch of s attached to v. We follow the convention to always take $\delta_v D_{e_1}^\alpha D_{e_2}^\beta s := \delta_v D_{e_1}^\alpha D_{e_2}^\beta (s|_\triangle)$, where $\triangle$ is the above triangle formed by e_1, e_2.

We shall also need functionals evaluating at some points on any edge e the derivatives of the spline in an affine invariant direction not parallel to e. Let $e = [v_1, v_2] \in \mathcal{E}_m$, and let $\triangle_e = [v_1, v_2, v_3] \in \mathcal{T}_m$ be a triangle attached to e. Denote by $\mu(e, \triangle)$ the median of $\triangle$ connecting the middle point $(v_1 + v_2)/2$ of e with the third vertex v_3 of $\triangle$. For any point $\xi \in e$, $\delta_\xi D_{\mu(e,\triangle)}$ will denote the derivative at ξ in the direction pointing into the half-plane containing $\triangle$ parallel to $\mu(e, \triangle)$, weighted with the length of $\mu(e, \triangle)$. For each edge $e \in \mathcal{E}_m$, we choose one of the two triangles attached to e and denote it by $\triangle_e$. (Note that this selection of $\triangle_e$ is not unique, but as will be seen it will cause no problems for the basis construction.)

Remark 4.1. For later references, we note here that any nodal functional $\eta : \mathcal{S}^{k,r}(\mathcal{T}_m) \to \mathbb{R}$ of the above type can be *extended* to a linear functional $\tilde\eta : \mathcal{S}^{k,-1}(\mathcal{T}_m) \to \mathbb{R}$ such that $\tilde\eta(s) = \eta(s)$ as long as $s \in \mathcal{S}^{k,r}(\mathcal{T}_m)$. Indeed, if the definition of η involves δ_ξ for some point $\xi \in \cup_{e\in\mathcal{E}_m} e$, then we choose one of the triangles $\triangle \in \mathcal{T}_m$ containing ξ and use the corresponding value of $s|_\triangle$ or its derivatives at ξ to define $\tilde\eta(s)$ for any $s \in \mathcal{S}^{k,-1}(\mathcal{T}_m)$. The only restriction on the choice of $\triangle$ is that it must be consistent with the above rules for $\delta_v D_e^\alpha$, $\delta_v D_{e_1}^\alpha D_{e_2}^\beta$, and $\delta_\xi D_{\mu(e,\triangle)}$. Clearly, the extension of this type is *not unique*. Moreover, convex combinations of evaluations of the restrictions of s to different triangles can also be used.

4.2. Characterization of differentiability. Let L be a straight line dividing $\mathbb{R}^2$ into two half-planes $H, \tilde H$. Given $p, \tilde p \in \Pi_k$, we define a piecewise polynomial function s by setting $s|_H = p$, $s|_{\tilde H} = \tilde p$. To check whether s is differentiable across L, we choose two points u, v on L, as well as two points $w, \tilde w$ in the interiors of H and $\tilde H$, respectively. We set $\triangle := [u, v, w]$, $\tilde\triangle := [u, v, \tilde w]$, $e := [u, v]$, $\mu := [u, w]$, $\tilde\mu := [u, \tilde w]$, $\theta := \angle e\mu$, $\tilde\theta := \angle \tilde\mu e$. The proof of the following lemma can be found in [17, 25].

LEMMA 4.2. *Let $0 \le r < k$. Then $s \in C^r(\mathbb{R}^2)$ if and only if*

$$\delta_u D_{\tilde\mu}^\alpha D_e^{q-\alpha} \tilde p = \sum_{\beta=0}^{\alpha} (-1)^\beta \binom{\alpha}{\beta} \left(\frac{\sin(\theta+\tilde\theta)}{|e|}\right)^{\alpha-\beta} \left(\frac{\sin\tilde\theta}{|\mu|}\right)^{\beta} \left(\frac{\sin\theta}{|\tilde\mu|}\right)^{-\alpha} \delta_u D_\mu^\beta D_e^{q-\beta} p \tag{4.2}$$

for all $\alpha = 0, \dots, r$ and $q = \alpha, \dots, k-1$.

It is readily seen that (4.2) can be reformulated as follows:

$$\delta_u D_{\tilde\mu}^\alpha D_e^{q-\alpha} \tilde p = \sum_{\beta=0}^{\alpha} (-1)^\beta \binom{\alpha}{\beta} \left(\sigma|\triangle^*|\right)^{\alpha-\beta} \frac{|\tilde\triangle|^\beta}{|\triangle|^\alpha} \delta_u D_\mu^\beta D_e^{q-\beta} p, \tag{4.3}$$

where $\sigma := \operatorname{sgn}\sin(\theta + \tilde\theta)$ and $\triangle^* := [u, w, \tilde w]$. (This identity simplifies in an obvious way when $|\triangle^*| = 0$.)

See [22] for a discussion of the relationship between these *nodal conditions of differentiability* and the well-known Bernstein–Bézier conditions.

4.3. Construction of basis splines. Consider the following set $\mathcal{N}_m$ of nodal functionals on $\mathcal{S}^{k,r}(\mathcal{T}_m)$,

$$\mathcal{N}_m := \left(\bigcup_{v\in\mathcal{V}_m} \mathcal{N}_m^v\right) \cup \left(\bigcup_{e\in\mathcal{E}_m} \mathcal{N}_m^e\right) \cup \left(\bigcup_{\triangle\in\mathcal{T}_m} \mathcal{N}_m^\triangle\right), \tag{4.4}$$

where for each $\triangle = [v_1, v_2, v_3] \in \mathcal{T}_m$,

$$\mathcal{N}_m^{\triangle} := \{\eta_\xi^{\triangle} := \delta_\xi \,:\, \xi \in \Xi_{\triangle}\},$$

$$\Xi_{\triangle} := \left\{ \frac{i_1 v_1 + i_2 v_2 + i_3 v_3}{k-1} \,:\quad i_1 + i_2 + i_3 = k-1, \quad i_1, i_2, i_3 > r \right\} \subset \triangle,$$

for each edge $e = [v_1, v_2] \in \mathcal{E}_m$,

$$\mathcal{N}_m^{e} := \{\eta_{q,\xi}^{e} := \delta_\xi D^q_{\mu(e,\triangle_e)} \,:\, q = 0, \dots, r, \quad \xi \in \Xi_{e,q}\},$$

$$\Xi_{e,q} := \left\{ \frac{i_1 v_1 + i_2 v_2}{k-q-1} \,:\, i_1 + i_2 = k-q-1, \quad i_1, i_2 > 2r - q \right\} \subset e,$$

and for each vertex $v \in \mathcal{V}_m$,

$$\mathcal{N}_m^{v} := \bigcup_{q=0}^{2r} \mathcal{N}_m^{v,q},$$

with $\mathcal{N}_m^{v,q}$, $q = 0, \dots, 2r$, being defined as follows. Let $\triangle^{[i]} = [v, v_i, v_{i+1}]$, $i = 1, \dots, N_v$, be the triangles in $\mathcal{T}_m$ attached to v in counterclockwise order, $v_{N_v+\ell} = v_\ell$, and let $e_i = [v, v_i]$. We set

$$\mathcal{N}_m^{v,0} := \{\eta^{v,0} := \delta_v\},$$

$$\mathcal{N}_m^{v,q} := \{\eta_{i,\alpha}^{v,q} := \delta_v D_{e_i}^{q-\alpha} D_{e_{i+1}}^{\alpha} \,:\, i = 1, \dots, N_v,\ \alpha = 0, \dots, q-1\}, \quad q \ge 1.$$

Note that $\mathcal{N}_m^{\triangle}$ or $\mathcal{N}_m^{e}$ might be empty for some combinations of r, k, e.g., $\mathcal{N}_m^{\triangle} = \mathcal{N}_m^{e} = \emptyset$ if $r = 0, k = 2$, or $\mathcal{N}_m^{\triangle} = \emptyset$ if $r = 1, k = 6$. This, however, does not cause any problem for the construction below.

In view of (4.2), the functionals in $\mathcal{N}_m^{v,q}$ are not linearly independent on $\mathcal{S}^{k,r}(\mathcal{T}_m)$ if $q \ge 1$. Namely, the following conditions hold for all $s \in \mathcal{S}^{k,r}(\mathcal{T}_m)$, $v \in \mathcal{V}_m$, $q = 1, \dots, 2r$:

$$\eta_{i,\alpha}^{v,q}(s) = \sum_{\beta=0}^{\alpha} (-1)^\beta \binom{\alpha}{\beta} \left(\frac{\sin(\theta_{i-1}+\theta_i)}{|e_i|}\right)^{\alpha-\beta} \left(\frac{\sin\theta_i}{|e_{i-1}|}\right)^{\beta} \left(\frac{\sin\theta_{i-1}}{|e_{i+1}|}\right)^{-\alpha} \eta_{i-1,q-\beta}^{v,q}(s), \tag{4.5}$$

$$\alpha = 1, \dots, \min\{r, q\}, \quad i = 1, \dots, N_v,$$

where $\theta_i := \angle e_i e_{i+1}$, $\eta_{i,q}^{v,q} := \eta_{i+1,0}^{v,q}$.

The following key lemma is instrumental in constructing the basis functions.

LEMMA 4.3. *There is a unique spline $s \in \mathcal{S}^{k,r}(\mathcal{T}_m)$ such that*

$$\begin{cases} \eta_\xi^{\triangle}(s) = a_\xi^{\triangle}, & \xi \in \Xi_{\triangle},\ \triangle \in \mathcal{T}_m, \\ \eta_{q,\xi}^{e}(s) = a_{q,\xi}^{e}, & \xi \in \Xi_{e,q},\ q = 0, \dots, r,\ e \in \mathcal{E}_m, \\ \eta^{v,0}(s) = a^{v,0}, & v \in \mathcal{V}_m, \\ \eta_{i,\alpha}^{v,q}(s) = a_{i,\alpha}^{v,q}, & i = 1, \dots, N_v,\ \alpha = 0, \dots, q-1,\ q = 1, \dots, 2r,\ v \in \mathcal{V}_m, \end{cases} \tag{4.6}$$

for any given $a_\xi^{\triangle}, a_{q,\xi}^{e}, a^{v,0} \in \mathbb{R}$ and any $a_{i,\alpha}^{v,q} \in \mathbb{R}$ satisfying

$$a_{i,\alpha}^{v,q} = \sum_{\beta=0}^{\alpha} (-1)^\beta \binom{\alpha}{\beta} \left(\frac{\sin(\theta_{i-1}+\theta_i)}{|e_i|}\right)^{\alpha-\beta} \left(\frac{\sin\theta_i}{|e_{i-1}|}\right)^{\beta} \left(\frac{\sin\theta_{i-1}}{|e_{i+1}|}\right)^{-\alpha} a_{i-1,q-\beta}^{v,q}, \tag{4.7}$$

$$\alpha = 1, \dots, \min\{r, q\}, \quad i = 1, \dots, N_v.$$

Moreover, for each $\triangle \in \mathcal{T}_m$,

$$\|s|_{\triangle}\|_{L_\infty(\triangle)} \le c\,\delta_2^{-2r} \max_{\eta\in\mathcal{N}_m(\triangle)} |\eta(s)|, \tag{4.8}$$

where c *is a constant depending only on* k, *and*

$$\mathcal{N}_m(\triangle) := \left(\bigcup_{v\in\mathcal{V}_m\cap\triangle} \mathcal{N}_m^v\right) \cup \left(\bigcup_{\substack{e\in\mathcal{E}_m\\ e\subset\triangle}} \mathcal{N}_m^e\right) \cup \mathcal{N}_m^{\triangle}.$$

Proof. We first determine $s|_e$ for each $e = [v_1, v_2] \in \mathcal{E}_m$ using the fact that $s|_e$, as a univariate function on the interval e, is a polynomial $s_{e,0}$ of degree at most $k-1$. Therefore, $s_{e,0}$ is uniquely determined by the following k Hermite interpolation conditions:

$$\begin{cases} \delta_{v_1} s_{e,0} = a^{v_1,0}, \quad \delta_{v_2} s_{e,0} = a^{v_2,0}, \\ \delta_{v_1} D_e^\gamma s_{e,0} = a_{i,0}^{v_1,\gamma}, \quad \delta_{v_2} D_e^\gamma s_{e,0} = a_{j,0}^{v_2,\gamma}, \quad \gamma = 1,\dots,2r, \\ \delta_\xi s_{e,0} = a_{0,\xi}^e, \qquad \xi \in \Xi_{e,0}, \end{cases} \tag{4.9}$$

where we assume that e is the ith edge emanating from v_1 and the jth edge emanating from v_2.

We next determine $s_{e,q} := (D^q_{\mu(e,\triangle_e)} s)|_e$, $q = 1,\dots,r$. Let $\triangle_e = [v_1, v_2, v_3]$. Then $D^q_{\mu(e,\triangle_e)} = (D_{[v_1,v_3]} - \frac{1}{2} D_{[v_1,v_2]})^q$. Therefore, for $\gamma = 0,\dots,2r-q$,

$$\delta_{v_1} D_e^\gamma s_{e,q} = \sum_{\ell=0}^{q} (-1)^\ell 2^{-\ell} \delta_{v_1} D_{e_i}^{\gamma+\ell} D_{e_{i+1}}^{q-\ell} s = \sum_{\ell=0}^{q} (-1)^\ell 2^{-\ell} \eta_{i,q-\ell}^{v_1,q+\gamma}(s).$$

Similarly, since $D^q_{\mu(e,\triangle_e)} = (D_{[v_2,v_3]} - \frac{1}{2} D_{[v_2,v_1]})^q$, we have for $\gamma = 0,\dots,2r-q$,

$$\begin{aligned} \delta_{v_2} D_e^\gamma s_{e,q} &= \sum_{\ell=0}^{q} (-1)^\ell 2^{-\ell} \delta_{v_2} D_{e_j}^{\gamma+\ell} D_{e_{j-1}}^{q-\ell} s \\ &= \sum_{\ell=0}^{q-1} (-1)^\ell 2^{-\ell} \eta_{j-1,\gamma+\ell}^{v_2,q+\gamma}(s) + (-1)^q 2^{-q} \eta_{j,0}^{v_2,q+\gamma}(s). \end{aligned}$$

In addition, we have for each $\xi \in \Xi_{e,q}$,

$$\delta_\xi s_{e,q} = \delta_\xi D^q_{\mu(e,\triangle_e)} s = \eta_{q,\xi}^e(s).$$

Thus, for each $q = 1,\dots,r$, the univariate polynomial $s_{e,q}$ of degree $k-1-q$ is uniquely determined by the $k-q$ Hermite interpolation conditions

$$\begin{cases} \delta_{v_1} D_e^\gamma s_{e,q} = \displaystyle\sum_{\ell=0}^{q} (-1)^\ell 2^{-\ell} a_{i,q-\ell}^{v_1,q+\gamma}, \quad \gamma = 0,\dots,2r-q, \\ \delta_{v_2} D_e^\gamma s_{e,q} = \displaystyle\sum_{\ell=0}^{q-1} (-1)^\ell 2^{-\ell} a_{j-1,\gamma+\ell}^{v_2,q+\gamma} + (-1)^q 2^{-q} a_{j,0}^{v_2,q+\gamma}, \quad \gamma = 0,\dots,2r-q, \\ \delta_\xi s_{e,q} = a_{q,\xi}^e, \quad \xi \in \Xi_{e,q}. \end{cases} \tag{4.10}$$

Let $\tilde{\triangle} = [v_1, v_2, \tilde{v}_3] \in \mathcal{T}_m$ be the second triangle attached to e. We set

$$\tilde{a}^e_{0,\xi} := a^e_{0,\xi}, \qquad \xi \in \Xi_{e,0},$$
$$\tilde{a}^e_{q,\xi} := \sum_{\ell=0}^{q} (-1)^\ell \binom{q}{\ell} (2\sigma|\triangle^*|)^{q-\ell} |\tilde{\triangle}|^\ell |\triangle_e|^{-q}\, \delta_\xi D_e^{q-\ell} s_{e,\ell},$$
$$\xi \in \Xi_{e,q}, \quad q = 1, \dots, r,$$

where $\triangle^* := [\bar{v}, v_3, \tilde{v}_3]$, $\bar{v} = (v_1 + v_2)/2$, and $\sigma := \operatorname{sgn} \sin(\widehat{v_3 \bar{v} v_2} + \widehat{\tilde{v}_3 \bar{v} v_2})$, where $\widehat{uvw}$ denotes the angle determined by u, v, w with vertex at v. (It may happen that $|\triangle^*| = 0$.) Since $\triangle^* \subset \operatorname{conv}(\triangle_e \cup \tilde{\triangle})$, we have

$$|\triangle^*|^{q-\ell} |\tilde{\triangle}|^\ell |\triangle_e|^{-q} \le \delta_2^{-q+\ell} |\tilde{\triangle}|^q |\triangle_e|^{-q} \le \delta_2^{-2q+\ell}. \tag{4.11}$$

We now construct each polynomial patch $s|_\triangle$, $\triangle \in \mathcal{T}_m$, of the spline s as the unique solution of the following interpolation problem:

$$\begin{cases} \delta_\xi(s|_\triangle) = a^\triangle_\xi, & \xi \in \Xi_\triangle, \\ \delta_\xi D^q_{\mu(e,\triangle)}(s|_\triangle) = a^e_{q,\xi}, & \xi \in \Xi_{e,q},\ q = 0, \dots, r,\ e \subset \triangle \ \text{ if } \triangle_e = \triangle, \\ \delta_\xi D^q_{\mu(e,\triangle)}(s|_\triangle) = \tilde{a}^e_{q,\xi}, & \xi \in \Xi_{e,q},\ q = 0, \dots, r,\ e \subset \triangle \ \text{ if } \triangle_e \ne \triangle, \\ \delta_v(s|_\triangle) = a^{v,0}, & v \in \triangle, \\ \delta_v D^{q-\alpha}_{e_i} D^\alpha_{e_{i+1}}(s|_\triangle) = a^{v,q}_{i,\alpha}, & \alpha = 0, \dots, q,\ q = 1, \dots, 2r,\ v \in \triangle, \\ & (i \text{ is such that } e_i, e_{i+1} \subset \triangle). \end{cases} \tag{4.12}$$

Since (4.12) is a standard finite element interpolation scheme for bivariate polynomials of degree $k-1$ (see, e.g., [57] or Lemma 3.7 in [25]), the polynomial $s|_\triangle$ is uniquely determined.

We now show that the piecewise polynomial s constructed in this way lies in the space $\mathcal{S}^{k,r}(\mathcal{T}_m)$; i.e., it is r times differentiable. To this end we consider any edge $e = [v_1, v_2] \in \mathcal{E}_m$. As before, let $\triangle_e = [v_1, v_2, v_3]$, and let $\tilde{\triangle} = [v_2, v_1, \tilde{v}_3]$ be the second triangle attached to e, and we again assume that e is the ith edge $e_{1,i}$ emanating from v_1 and at the same time the jth edge $e_{2,j}$ emanating from v_2. Then we have

$$e_{1,i-1} = [v_1, \tilde{v}_3], \quad e_{1,i} = [v_1, v_2], \quad e_{1,i+1} = [v_1, v_3],$$
$$e_{2,j-1} = [v_2, v_3], \quad e_{2,j} = [v_2, v_1], \quad e_{2,j+1} = [v_2, \tilde{v}_3].$$

Obviously, for each $q = 0, \dots, r$, $D^q_{\mu(e,\triangle_e)}(s|_{\triangle_e})|_e = s_{e,q}$ satisfies the interpolation conditions (4.9) if $q = 0$ or (4.10) if $q > 0$. We set

$$\hat{s}_{e,q} := D^q_{\mu(e,\tilde{\triangle})}(s|_{\triangle_e})|_e.$$

The desired differentiability of s will follow if we show that

$$\hat{s}_{e,q} = \tilde{s}_{e,q} := D^q_{\mu(e,\tilde{\triangle})}(s|_{\tilde{\triangle}})|_e, \qquad q = 0, \dots, r. \tag{4.13}$$

By (4.12) we have

$$\delta_{v_1}(s|_{\triangle_e}) = \delta_{v_1}(s|_{\tilde{\triangle}}) = a^{v_1,0},$$
$$\delta_{v_1} D^{q-\alpha}_{e_{1,i}} D^\alpha_{e_{1,i+1}}(s|_{\triangle_e}) = a^{v_1,q}_{i,\alpha}, \quad \alpha = 0, \dots, q-1,\ q = 1, \dots, 2r,$$
$$\delta_{v_1} D^{q-\alpha}_{e_{1,i-1}} D^\alpha_{e_{1,i}}(s|_{\tilde{\triangle}}) = a^{v_1,q}_{i-1,\alpha}, \quad \alpha = 0, \dots, q-1,\ q = 1, \dots, 2r,$$

which in view of (4.7) imply

$$\delta_{v_1} D^{q-\alpha}_{e_{1,i}} D^{\alpha}_{e_{1,i+1}}(s|_{\triangle_e}) = \delta_{v_1} D^{q-\alpha}_{e_{1,i}} D^{\alpha}_{e_{1,i+1}}(s|_{\bar{\triangle}}),$$
$$\alpha = 0, \dots, \min\{r, q\}, \quad q = 0, \dots, 2r,$$

and hence

$$\delta_{v_1} D^{\gamma}_{e}(\hat{s}_{e,q} - \tilde{s}_{e,q}) = 0, \qquad \gamma = 0, \dots, 2r - q, \quad q = 0, \dots, r.$$

Similarly, we get

$$\delta_{v_2} D^{\gamma}_{e}(\hat{s}_{e,q} - \tilde{s}_{e,q}) = 0, \qquad \gamma = 0, \dots, 2r - q, \quad q = 0, \dots, r.$$

In addition, a simple calculation relying on (4.3) shows that

$$\delta_{\xi} \hat{s}_{e,q} = \tilde{a}^{e}_{q,\xi}, \qquad \xi \in \Xi_{e,q}, \quad q = 0, \dots, r,$$

so that by (4.12),

$$\delta_{\xi}(\hat{s}_{e,q} - \tilde{s}_{e,q}) = 0, \qquad \xi \in \Xi_{e,q}, \quad q = 0, \dots, r.$$

Since $\hat{s}_{e,q} - \tilde{s}_{e,q}$ satisfies homogeneous interpolation conditions of a well-posed Hermite scheme, (4.13) follows.

The uniqueness of s is clear from the above proof, since $s = 0$ if the numbers in the right-hand side of (4.6) are all zeros.

It remains to prove (4.8). Since $s_{e,q}$ satisfies the interpolation conditions (4.9) if $q = 0$ or (4.10) if $q > 0$,

$$\|s_{e,q}\|_{L_\infty(e)} \le c \max\{\eta(s) \,:\, \eta \in \mathcal{N}^{v_1}_m \cup \mathcal{N}^{v_2}_m \cup \mathcal{N}^{e}_m\}, \qquad q = 0, \dots, r,$$

where c depends only on k. In view of (4.11) and Markov inequality (4.1), we have

$$|\tilde{a}^{e}_{q,\xi}| \le c\, \delta_2^{-2q} \|s_{e,q}\|_{L_\infty(e)}, \qquad q = 0, \dots, r,$$

and (4.8) follows by the properties of the interpolation problem (4.12); see Lemma 3.9 in [25]. □

For each $v \in \mathcal{V}_m$ and $q = 1, \dots, 2r$, we denote by $R^{v,q}_m$ the $(\min\{r, q\}N_v \times qN_v)$-matrix of differentiability conditions (4.5). Let the vectors

$$a^{v,q,j}, \qquad j = 1, \dots, \rho_{v,q} := qN_v - \operatorname{rank}(R^{v,q}_m),$$

form an orthonormal basis for the null space of $R^{v,q}_m$,

$$\operatorname{null}(R^{v,q}_m) := \{a \in \mathbb{R}^{qN_v} \,:\, R^{v,q}_m a = 0\}.$$

For convenience, we shall use the double indices introduced in the definition of $\mathcal{N}^{v,q}_m$ also for the components of $a^{v,q,j}$:

$$a^{v,q,j}_{i,\alpha}, \qquad i = 1, \dots, N_v, \quad \alpha = 0, \dots, q - 1. \tag{4.14}$$

We set

$$\eta^{v,q,j} := \sum_{i=1}^{N_v} \sum_{\alpha=0}^{q-1} a_{i,\alpha}^{v,q,j} \eta_{i,\alpha}^{v,q}, \qquad j = 1, \dots, \rho_{v,q}, \tag{4.15}$$

$$\tilde{\mathcal{N}}_m^{v,q} := \{\eta^{v,q,j} \,:\, j = 1, \dots, \rho_{v,q}\}, \quad q = 1, \dots, 2r,$$
$$\tilde{\mathcal{N}}_m^{v} := \mathcal{N}_m^{v,0} \cup \bigcup_{q=1}^{2r} \tilde{\mathcal{N}}_m^{v,q}, \qquad v \in \mathcal{V}_m,$$
$$\tilde{\mathcal{N}}_m := \left(\bigcup_{v \in \mathcal{V}_m} \tilde{\mathcal{N}}_m^{v} \right) \cup \left(\bigcup_{e \in \mathcal{E}_m} \mathcal{N}_m^{e} \right) \cup \left(\bigcup_{\triangle \in \mathcal{T}_m} \mathcal{N}_m^{\triangle} \right),$$

and define the set

$$\Phi_m = \{\varphi_\eta \,:\, \eta \in \tilde{\mathcal{N}}_m\}$$

of the *basis functions* for $\mathcal{S}^{k,r}(\mathcal{T}_m)$ by the duality condition,

$$\mu(\varphi_\eta) = \begin{cases} 1 & \text{if } \mu = \eta, \\ 0 & \text{if } \mu \in \tilde{\mathcal{N}}_m \setminus \{\eta\}. \end{cases} \tag{4.16}$$

To see that the above definition is correct we have to check that for each $\eta \in \tilde{\mathcal{N}}_m$ there exists a unique φ_η satisfying (4.16). This follows by Lemma 4.3. Indeed, since the vectors $a^{v,q,j}$ are orthonormal, we have

$$\eta_{i,\alpha}^{v,q} = \sum_{j=1}^{\rho_{v,q}} a_{i,\alpha}^{v,q,j} \eta^{v,q,j}, \qquad i = 1, \dots, N_v, \quad \alpha = 0, \dots, q-1.$$

Therefore, for a fixed η, the numbers

$$a_{i,\alpha}^{v,q} := \eta_{i,\alpha}^{v,q}(\varphi_\eta), \qquad i = 1, \dots, N_v, \quad \alpha = 0, \dots, q-1,$$

satisfy (4.7), which ensures the applicability of Lemma 4.3.

4.4. Properties of basis splines. It follows by Lemma 4.3 that every spline $s \in \mathcal{S}^{k,r}(\mathcal{T}_m)$ is uniquely determined by the sequence $(\eta(s))_{\eta \in \tilde{\mathcal{N}}_m}$; i.e., s has a unique representation

$$s = \sum_{\eta \in \tilde{\mathcal{N}}_m} a_\eta \varphi_\eta, \qquad a_\eta = \eta(s) \in \mathbb{R}.$$

Furthermore, (4.8) immediately implies

$$\operatorname{supp} \varphi_\eta \subseteq \begin{cases} \operatorname{star}(v) & \text{if } \eta \in \tilde{\mathcal{N}}_m^{v} \text{ for a vertex } v \in \mathcal{V}_m, \\ \operatorname{star}(e) & \text{if } \eta \in \mathcal{N}_m^{e} \text{ for an edge } e \in \mathcal{E}_m, \\ \triangle & \text{if } \eta \in \mathcal{N}_m^{\triangle} \text{ for a triangle } \triangle \in \mathcal{T}_m, \end{cases} \tag{4.17}$$

$$\|\varphi_\eta\|_{L_\infty(\mathbb{R}^2)} \le c\, \delta_2^{-2r}. \tag{4.18}$$

By using Markov inequality it is easy to show that

$$|\eta(s)| \le \tilde{c} \begin{cases} \|s\|_{L_\infty(\mathrm{star}\,(v))} & \text{if } \eta \in \tilde{\mathcal{N}}_m^v \text{ for a vertex } v \in \mathcal{V}_m, \\ \|s\|_{L_\infty(\mathrm{star}\,(e))} & \text{if } \eta \in \mathcal{N}_m^e \text{ for an edge } e \in \mathcal{E}_m, \\ \|s\|_{L_\infty(\triangle)} & \text{if } \eta \in \mathcal{N}_m^\triangle \text{ for a triangle } \triangle \in \mathcal{T}_m, \end{cases} \tag{4.19}$$

with $\tilde{c}$ a constant depending only on k, r and N_0.

Thus, we showed that the basis $\Phi_m = \{\varphi_\eta : \eta \in \tilde{\mathcal{N}}_m\}$ satisfies all requirements of section 2.2 with $\mathcal{S}_m = \mathcal{S}^{k,r}(\mathcal{T}_m)$ and $\tilde{k} = k$. (Obviously, $\Pi_k \subset \mathcal{S}_m$ and $\mathcal{S}^{k,r}(\mathcal{T}_m) \subset \mathcal{S}^{k,r}(\mathcal{T}_{m+1})$ if $\mathcal{T}_{m+1}$ is a refinement of $\mathcal{T}_m$.) More precisely, we have the following result.

THEOREM 4.4. *Let $r \ge 0$, $k > 4r+1$. Suppose that $\mathcal{T}_m$ satisfies* (a), (d)–(f), *and* (2.3) *of section* 2.1. *Then the basis functions $\varphi_\eta \in \mathcal{S}^{k,r}(\mathcal{T}_m)$ ($\eta \in \tilde{\mathcal{N}}_m$) constructed above have the following properties:*

(a) *For any $s \in \mathcal{S}^{k,r}(\mathcal{T}_m)$ there exists a unique sequence of real coefficients $(a_\eta)_{\eta\in\tilde{\mathcal{N}}_m}$ such that*

$$s = \sum_{\eta\in\tilde{\mathcal{N}}_m} a_\eta \varphi_\eta,$$

with $a_\eta = \eta(s)$, $\eta \in \tilde{\mathcal{N}}_m$.

(b) *For each $\eta \in \tilde{\mathcal{N}}_m$ there is a vertex $v = v_\eta \in \mathcal{V}_m$ such that*

$$\begin{aligned} \operatorname{supp}\varphi_\eta &\subset \mathrm{star}\,(v) =: E_\eta, \\ \|\varphi_\eta\|_{L_\infty(\mathbb{R}^2)} &= \|\varphi_\eta\|_{L_\infty(E_\eta)} \le M_1, \\ |\eta(s)| &\le M_2 \|s\|_{L_\infty(E_\eta)}, \qquad s \in \mathcal{S}^{k,r}(\mathcal{T}_m), \end{aligned}$$

where M_1, M_2 are positive constants depending only on k, r, δ_2, and N_0.

In particular, by the proof of Lemma 2.3, we have the following stability property of Φ_m.

THEOREM 4.5. *The basis Φ_m is L_p-stable for all $0 < p \le \infty$; i.e., for any sequence $(a_\eta)_{\eta\in\tilde{\mathcal{N}}_m}$,*

$$\Big\| \sum_{\eta\in\tilde{\mathcal{N}}_m} a_\eta \varphi_\eta \Big\|_{L_p(\mathbb{R}^2)} \approx \left(\sum_{\eta\in\tilde{\mathcal{N}}_m} \|a_\eta\varphi_\eta\|^p_{L_p(\mathbb{R}^2)} \right)^{1/p},$$

where the constants of equivalence depend only on p, k, r, δ_2, and N_0. In the case $p = \infty$ the ℓ_p-norm in the right-hand side is replaced by the sup-norm.

The linear functionals $\lambda_\eta : \mathcal{S}^{k,-1}(\mathcal{T}_m) \cap L_\infty(E_\eta) \to \mathbb{R}$, $\eta \in \tilde{\mathcal{N}}_m$, with properties

$$\lambda_\eta(s|_{E_\eta}) = \eta(s), \quad s \in \mathcal{S}^{k,r}(\mathcal{T}_m),$$

$$|\lambda_\eta(f)| \le M_2 \|f\|_{L_\infty(E_\eta)}, \quad f \in \mathcal{S}^{k,-1}(\mathcal{T}_m)|_{E_\eta} \cap L_\infty(E_\eta),$$

needed in the definition of the projector Q_m (see (2.18)) can now be defined in a *constructive* manner. Indeed, we first extend each functional $\eta \in \mathcal{N}_m$ to a functional $\tilde{\eta}$ defined on $\mathcal{S}^{k,-1}(\mathcal{T}_m)$, according to Remark 4.1, and then set

$$\lambda_\eta := \tilde{\eta} \qquad \text{if} \qquad \eta \in \left(\bigcup_{e\in\mathcal{E}_m} \mathcal{N}_m^e \right) \cup \left(\bigcup_{\triangle\in\mathcal{T}_m} \mathcal{N}_m^\triangle \right)$$

and

$$\lambda_\eta := \sum_{i=1}^{N_v}\sum_{\alpha=0}^{q-1} a_{i,\alpha}^{v,q,j}\tilde\eta_{i,\alpha}^{v,q} \qquad \text{if} \qquad \eta^{v,q,j} = \sum_{i=1}^{N_v}\sum_{\alpha=0}^{q-1} a_{i,\alpha}^{v,q,j}\eta_{i,\alpha}^{v,q} \in \bigcup_{v\in\mathcal{V}_m}\tilde{\mathcal{N}}_m^v.$$

By (2.22), Q_m can be extended to the operator $Q_{m,p} : L_p^{\rm loc} \to \mathcal{S}^{k,r}(\mathcal{T}_m)$ whose local approximation power is described in the following theorem (see Lemma 2.5).

THEOREM 4.6. *Suppose $f \in L_p^{\rm loc}$, $0 < p \le \infty$ ($f \in C$ if $p = \infty$). Then*

$$\|f - Q_{m,p}(f)\|_{L_p(\triangle)} \le c\mathbb{S}_\triangle(f)_p \le cE_k(f,\Omega_\triangle)_p, \quad \triangle \in \mathcal{T}_m,$$

where $\Omega_\triangle := \Omega_\triangle^1$ is the union of all triangles in $\mathcal{T}_m$ that have a common vertex with $\triangle$, and the constant c depends only on p, k, r, δ_2, and N_0.

To show that the assumption that the triangulations $\mathcal{T}_m$ satisfy (2.3) cannot be omitted, we consider the following example.

Example 4.7. Suppose $\mathcal{T}_m$ has an edge $e = [v,u]$ with two triangles $\triangle = \triangle_e = [v,u,w]$ and $\tilde\triangle = [v,u,\tilde w]$ attached to e such that $u = v + (2^{-M}\alpha, 0)$, $w = v + (-\alpha,\alpha)$, $\tilde w = v + (-\alpha,-\alpha)$, where the positive numbers M, α depend on m. Evidently, $|\mathrm{conv}\,(\triangle_e \cup \tilde\triangle)|/|\triangle_e| = 2(2^M+1)$, and (2.3) will be violated if M grows unboundedly with m, while the maximal angle of the two triangles is $3\pi/4$, thus allowing the maximal angle condition (2.5) to hold. Note that such configurations of triangles are possible for a sequence of levels of an LR-triangulation $\mathcal{T}$ with the corresponding M's tending to infinity; see section 2.1 of [38]. Choosing $k = 6$ and $r = 1$, we consider the basis functions $\varphi_\eta \in \mathcal{S}^{6,1}(\mathcal{T}_m)$, $\eta \in \tilde{\mathcal{N}}_m$, constructed according to the above algorithm. We next show that the basis $\Phi_m = \{\varphi_\eta : \eta \in \tilde{\mathcal{N}}_m\}$ is *instable*; i.e., Theorem 4.5 does not hold for it. (Therefore, neither Φ_m nor a renorming of it satisfies the requirements of section 2.2.) More precisely, we show that the constant function $\mathbb{1}_{\mathbb{R}^2}(x) \equiv 1$, $x \in \mathbb{R}^2$, does not have an L_∞-stable expansion with respect to Φ_m. We have

$$\|\mathbb{1}_{\mathbb{R}^2}\|_{L_\infty(\mathbb{R}^2)} = 1, \qquad \mathbb{1}_{\mathbb{R}^2} = \sum_{\eta\in\tilde{\mathcal{N}}_m} \eta(\mathbb{1}_{\mathbb{R}^2})\varphi_\eta.$$

Now choose $\eta = \eta^{v,0} = \delta_v \in \mathcal{N}_m^{v,0}$. Since $\eta(\mathbb{1}_{\mathbb{R}^2}) = 1$, the instability of Φ_m will follow if we show that $\|\varphi_\eta\|_{L_\infty(\mathbb{R}^2)}$ is unbounded as $M \to \infty$. By (4.12),

$$\begin{aligned}\delta_\xi D^1_{\mu(e,\tilde\triangle)}(\varphi_\eta|_{\tilde\triangle}) &= \tilde a^e_{1,\xi}\\ &= -2\tfrac{|\triangle^*|}{|\triangle_e|}\delta_\xi D_e^1 s_{e,0} - \tfrac{|\tilde\triangle|}{|\triangle_e|}\delta_\xi s_{e,1},\end{aligned}$$

where $\xi = (v+u)/2$, $\triangle^* = [\xi, w, \tilde w]$, $s_{e,0} = \varphi_\eta|_e$, $s_{e,1} = (D^1_{\mu(e,\triangle_e)}\varphi_\eta)|_e$. Obviously, $|\tilde\triangle|/|\triangle_e| = 1$, and

$$|\triangle^*|/|\triangle_e| = \left(|\mathrm{conv}\,(\triangle_e \cup \tilde\triangle)| - \tfrac{|\triangle_e|+|\tilde\triangle|}{2}\right)/|\triangle_e| = 2^{M+1} + 1.$$

The univariate polynomial $s_{e,0}$ of degree 5 is determined by the Hermite interpolation conditions (4.9) that take in our case the form

$$\delta_v s_{e,0} = 1, \quad \delta_u s_{e,0} = \delta_u D_e^1 s_{e,0} = \delta_u D_e^2 s_{e,0} = \delta_v D_e^1 s_{e,0} = \delta_v D_e^2 s_{e,0} = 0.$$

An elementary computation shows that $\delta_\xi D_e^1 s_{e,0} = -15/8$. By (4.10), we immediately get $\delta_\xi s_{e,1} = a^e_{1,\xi} = \eta^e_{1,\xi}(\varphi_\eta) = 0$. Thus,

$$\delta_\xi D^1_{\mu(e,\tilde\triangle)}(\varphi_\eta|_{\tilde\triangle}) = \frac{15}{4}(2^{M+1}+1) \to \infty \quad \text{as} \quad M \to \infty.$$

In view of Markov inequality, $\|\varphi_\eta\|_{L_\infty(\mathbb{R}^2)} \geq c|\delta_\xi D^1_{\mu(e,\tilde{\triangle})}(\varphi_\eta|_{\tilde{\triangle}})|$, and we get the desired unboundedness of $\|\varphi_\eta\|_{L_\infty(\mathbb{R}^2)}$ for sufficiently large M.

Remark 4.8. It is clear that Theorems 4.4–4.6 are valid for any sequence of levels $\mathcal{T}_m$ satisfying the hypotheses of Theorem 4.4; i.e., nestedness and other additional assumptions on $\{\mathcal{T}_m\}$ stated in section 2.1 are not needed for these results.

Remark 4.9. It is an important property of the basis functions φ_η constructed above that they are invariant under affine transforms. More precisely, let $\mathcal{T}_m$ satisfy the hypotheses of Theorem 4.4, and let $A : \mathbb{R}^2 \to \mathbb{R}^2$ be an affine transform. We set $A(\mathcal{T}_m) = \{A(\triangle) : \triangle \in \mathcal{T}_m\}$, and $A\eta(s) := \delta_{A(v)} D^\alpha_{A(e_1)} D^\beta_{A(e_2)} s$, for each nodal functional η of the form $\eta(s) = \delta_v D^\alpha_{e_1} D^\beta_{e_2} s$ and extend the operator A linearly to the linear combinations of the nodal functionals such as those occurring in (4.15). Then, clearly, the sets of nodal functionals $\mathcal{N}_m$ and $\mathcal{N}^A_m$ defined by (4.4) for $\mathcal{T}_m$ and $A(\mathcal{T}_m)$, respectively, satisfy $\mathcal{N}^A_m = \{A\eta : \eta \in \mathcal{N}_m\}$. (We used here, in particular, the fact that $\mu(A(e), A(\triangle_e)) = A(\mu(e, \triangle_e))$.) Moreover, since the matrices $R^{v,q}_m$ of the differentiability conditions (4.5) are affine invariant (see (4.3)), we also have $\tilde{\mathcal{N}}^A_m = \{A\eta : \eta \in \tilde{\mathcal{N}}_m\}$ for the appropriate sets $\tilde{\mathcal{N}}_m, \tilde{\mathcal{N}}^A_m$ defined as in the construction above, provided we choose the same orthonormal vectors (4.14) in both cases. Let now $\Phi_m = \{\varphi_\eta : \eta \in \tilde{\mathcal{N}}_m\} \subset \mathcal{S}^{k,r}(\mathcal{T}_m)$ and $\Phi^A_m = \{\varphi_{A\eta} : \eta \in \tilde{\mathcal{N}}_m\} \subset \mathcal{S}^{k,r}(A(\mathcal{T}_m))$ be the spline bases dual to $\tilde{\mathcal{N}}_m$ and $\tilde{\mathcal{N}}^A_m$, respectively. Since $\varphi_\eta(A\cdot)$, $\eta \in \tilde{\mathcal{N}}_m$, obviously satisfy the same duality relations, we conclude that $\varphi_{A\eta} = \varphi_\eta(A\cdot)$, $\eta \in \tilde{\mathcal{N}}_m$, which is the desired affine invariance.

Remark 4.10. Our construction is extendable to the spaces $\mathcal{S}^{k,r}(\mathcal{T}_m)$, $k > r2^d + 1$, in dimensions $d > 2$. To this end the algorithm given in [22] should be extended to SLR-triangulations in $\mathbb{R}^d$. In particular, the orthogonal directions of derivatives used in [22] should be replaced by appropriate affine invariant directions.

Remark 4.11. If the triangulation covers only a compact domain E, then usual modifications of basis functions corresponding to boundary edges or vertices (see [22, 23]) lead to the desired stable local bases.

Remark 4.12. In this section, we extended to the setting of SLR-triangulations the bivariate version of nodal stable local basis construction of [22, 23], which was originally designed for regular triangulations. The scheme from [27] can be used as an alternative means of constructing stable local bases for $\mathcal{S}^{k,r}(\mathcal{T}_m)$, $k > 3r + 2$, in dimension $d = 2$. Such a development would take advantage of the affine invariance of the Bernstein–Bézier representation of piecewise polynomials. We elected to utilize the scheme from [22] instead, since it is available for any number of variables and allows an effective numerical implementation as shown (for $r = 1, 2$, $d = 2$) in [23]. Also, we want to pay heed to two more spline basis constructions (for regular triangulations in dimension $d = 2$) that allow the same kind of extension to SLR-triangulations: (a) stable local bases for $\mathcal{S}^{k,1}(\mathcal{T}_m)$, $k > 5$, constructed in [26]; (b) locally stable bases on nested triangulations ($k > 4r + 1$) [24]. Note that the stable local bases for superspline subspaces of $\mathcal{S}^{k,r}(\mathcal{T}_m)$ [16, 17, 44, 57] cannot be used since these spaces are not nested for nested triangulations, while the earlier local spline bases for $\mathcal{S}^{k,r}(\mathcal{T}_m)$ [1, 8, 18, 35, 36, 48] are known to be unstable for certain triangulations.

Remark 4.13. It is easy to see that, in the case $r = 0$, the above basis reduces to the classical Lagrange finite element basis for $\mathcal{S}^{k,0}(\mathcal{T}_m)$, $k > 1$. Since δ_2 disappears from (4.8) when $r = 0$, Theorems 4.4–4.6 hold for locally regular triangulations; i.e., the SLR assumption (2.3) is not needed in this case. (Note that δ_2 and N_0 completely disappear from Theorem 4.4, and δ_2 is replaced by δ_1 in Theorems 4.5–4.6.) For $r = 0$, $k = 2$, we get the Courant elements, and the only essential difference

to the construction from [38] is that we rely here on the extensions of linear functionals described in Remark 4.1 rather than on the explicit quasi interpolant for continuous piecewise linear functions adopted in [38]. Both approaches obviously lead to the same B-spaces.

5. Spline bases on special triangulations. There are several constructions of differentiable spline bases fitting into our scheme that are only available for specific multilevel triangulations. Since these triangulations have a special structure or even are uniform, the corresponding libraries $\{\Phi_{\mathcal{T}}\}$ of bases are not as rich as the one of the previous section associated with arbitrary SLR-triangulations. Moreover, the necessity to maintain the structure of the triangulation highly reduces the variety of refinement methods that can be used (whereas, e.g., *local* refinement by bisection can be used with bases on arbitrary triangulations). On the other hand, bases on special triangulations usually allow a smaller degree of piecewise polynomials for a given order of differentiability as well as a simpler and more efficient practical implementation.

In this section, we review some known constructions of this type. (Note that only box splines are available for more than two variables.)

5.1. Box splines. As usual, we consider only splines of two variables. Let $\Xi = [\xi_1 \cdots \xi_n]$ be a full rank $2 \times n$ matrix with columns ξ_i in $\mathbb{Z}^2 \setminus 0$. The *box spline* $M_\Xi : \mathbb{R}^2 \to \mathbb{R}^2$ associated with Ξ is defined by its Fourier transform

$$\hat{M}_\Xi(u) = \prod_{\nu=1}^{n} \frac{1 - e^{-i\xi_\nu u}}{i\xi_\nu u}, \qquad u \in \mathbb{R}^2,$$

where $\xi_\nu u$ denotes the inner product of the two vectors.

We now review the basic properties of box splines (see [9]), in order to verify the requirements of section 2.2. It is well known that M_Ξ has a compact support,

$$\operatorname{supp} M_\Xi = \left\{ \sum_{\nu=1}^{n} t_\nu \xi_\nu \ : \ 0 \le t_\nu \le 1 \right\}. \tag{5.1}$$

The box spline basis functions at the mth level are defined by

$$\varphi_{m,j} = M_\Xi(2^m \cdot - j), \qquad j \in \mathbb{Z}^2.$$

We set

$$\Phi_m = \{\varphi_{m,j} \ : \ j \in \mathbb{Z}^2\}, \qquad m \in \mathbb{Z},$$

and

$$\mathcal{S}_m = \left\{ \sum_{j \in \mathbb{Z}^2} a_{m,j} \varphi_{m,j} \ : \ a_{m,j} \in \mathbb{R} \right\}, \qquad m \in \mathbb{Z},$$

where the series converges everywhere since for every $x \in \mathbb{R}^2$ and $m \in \mathbb{Z}$ only a finite number of $\varphi_{m,j}(x)$ ($j \in \mathbb{Z}^2$) are nonzero. Clearly, any affine change of variables $Q : \mathbb{R}^2 \to \mathbb{R}^2$ gives rise to basis functions $\varphi_{m,j}(Qx)$ that satisfy the conditions of section 2.2 if and only if the $\varphi_{m,j}$ do. Therefore, we do not distinguish between constructions that can be transformed into each other by such a method.

Since

$$\frac{\hat{M}_\Xi(2u)}{\hat{M}_\Xi(u)} = \prod_{\nu=1}^{n} \frac{1+e^{-i\xi_\nu u}}{2},$$

M_Ξ is a finite linear combination of $M_\Xi(2\cdot - j)$, $j \in \mathbb{Z}^2$, which implies that

$$\mathcal{S}_m \subset \mathcal{S}_{m+1}, \qquad m \in \mathbb{Z}.$$

Let

$$r(\Xi) := \max\{r \,:\, \text{any } 2\times(n-r) \text{ submatrix of } \Xi \text{ has rank } 2\} - 1$$

and

$$k(\Xi) := n - 1.$$

The elements of $\mathcal{S}_m$ are $r(\Xi)$ times differentiable piecewise polynomials of degree $k(\Xi)-1$ with respect to the rectilinear partition $\mathcal{T}_m^\Xi$ of $\mathbb{R}^2$ determined by the straight lines

$$H_\nu + 2^{-m} j, \qquad j \in \mathbb{Z}^2, \quad \nu = 1, \dots, n,$$

where

$$H_\nu := \{t\xi_\nu \,:\, t \in \mathbb{R}\}.$$

Thus,

$$\mathcal{S}_m \subset \mathcal{S}^{k(\Xi), r(\Xi)}(\mathcal{T}_m^\Xi).$$

Moreover,

$$\Pi_{\tilde{k}(\Xi)} \subset \mathcal{S}_m, \qquad m \in \mathbb{Z},$$

and $\Pi_{\tilde{k}(\Xi)+1} \not\subset \mathcal{S}_m$, where

$$\tilde{k}(\Xi) = r(\Xi) + 2.$$

It is well known that the translates of a box spline are not always linearly independent. In fact, Φ_m is a basis for $\mathcal{S}_m$ ($m \in \mathbb{Z}$) if and only if the matrix Ξ is *unimodular*; i.e., each nonsingular 2×2 submatrix of Ξ has determinant ± 1. This condition implies substantial restrictions on Ξ. Namely, up to an affine change of variables, Ξ must have the form

$$\Xi = [\underbrace{e_1 \cdots e_1}_{n_1} \underbrace{e_2 \cdots e_2}_{n_2} \underbrace{e_3 \cdots e_3}_{n_3}],$$

where $e_1 = \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$, $e_2 = \left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$, $e_3 = \left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]$, $n_1, n_2 \geq 1$, $n_3 \geq 0$, and $n_1 + n_2 + n_3 = n$. It is easy to see that

$$r(\Xi) = n - \max\{n_1, n_2, n_3\} - 2$$

and that $\mathcal{T}_m^{\Xi}$ is either a tensor product mesh if $n_3 = 0$ or a *three-directional mesh* $\mathcal{T}_m^{(1)}$ defined by the straight lines $x_1 = 2^{-m}j$, $x_2 = 2^{-m}j$, $x_1 - x_2 = 2^{-m}j$ ($j \in \mathbb{Z}^2$) in $\mathbb{R}^2$ if $n_3 \geq 1$. Since only the latter case leads to a multilevel triangulation, we assume that $n_3 \geq 1$.

It remains to verify (2.10)–(2.12). By (5.1), the support of M_Ξ is the hexagon with vertices $(0,0)$, $(n_1,0)$, $(0,n_2)$, (n_1+n_3,n_3), (n_3,n_2+n_3), (n_1+n_3,n_2+n_3), which implies (2.10) with $\ell \leq \lfloor n/2 \rfloor$. Obviously, (2.11) is valid with $M_1 = \|M_\Xi\|_{L_\infty}$. Finally, it is easy to show (2.12) by using the constructions of dual functionals $\lambda_j : \mathcal{S}_0 \to \mathbb{R}$ ($j \in \mathbb{Z}^2$), with $\lambda_j(\varphi_{0,k}) = \delta_{j,k}$, given, e.g., in [19, 37, 41].

Let us mention the following two cases that are perhaps most relevant in applications:

$$\begin{aligned} &\text{(a)} \quad n_1 = n_2 = 2, \quad n_3 = 1, \quad \mathcal{S}_m \subset \mathcal{S}^{4,1}(\mathcal{T}_m^{(1)}), \quad \tilde{k} = 3, \\ &\text{(b)} \quad n_1 = n_2 = n_3 = 2, \quad \mathcal{S}_m \subset \mathcal{S}^{5,2}(\mathcal{T}_m^{(1)}), \quad \tilde{k} = 4. \end{aligned}$$

5.2. Other spline bases on uniform triangulations. There are some other spline basis constructions for the three-directional mesh $\mathcal{T}_m^{(1)}$; see, e.g., [15, 56]. However, to our knowledge, none of them simultaneously satisfies the requirements of nestedness of the spaces, stability, and locality of the basis functions. The situation is better for the *four-directional mesh* $\mathcal{T}_m^{(2)}$ obtained from $\mathcal{T}_m^{(1)}$ by adding the straight lines $x_1 + x_2 = 2^{-m}j$ ($j \in \mathbb{Z}^2$). Since $\mathcal{T}_m^{(2)}$ is a special case of a so-called FVS-triangulation (see section 5.3), finite element bases for $\mathcal{S}^{4,1}(\mathcal{T}_m^{(2)})$ are available and satisfy the conditions of section 2.2. Some recent alternative constructions of stable local bases for $\mathcal{S}^{4,1}(\mathcal{T}_m^{(2)})$ can be found in [13, 28, 42, 49]. Moreover, a stable local basis for $\mathcal{S}^{7,2}(\mathcal{T}_m^{(2)})$ is also constructed in [28]. Finally, we want to mention the stable local basis from [33] for C^1 quadratic splines with respect to a sequence of triangulation levels that can be called the *six-directional meshes*.

5.3. Refinable composite finite elements. Multilevel and hierarchical bases play an important role in the modern theory and practice of numerical methods for PDEs; see, e.g., [51]. Classical smooth finite elements [14] give rise to stable local spline bases on triangulations satisfying the minimal angle condition. (Note that it should be possible to replace this condition of regularity with SLR.) However, there are difficulties in using them to build nested spline spaces on multilevel triangulations; see [11, 20]. Although the "polynomial" finite elements (e.g., the Argyris element) are available for arbitrary triangulations, they lead to *superspline* spaces [57] that lack nestedness for nested triangulations (*levels* in the terminology of our section 2). In contrast to them, "composite" finite elements require a special structure of the levels $\mathcal{T}_m$, e.g., a Clough–Tocher or Powell–Sabin split, which is not always compatible with nested refinements with other desirable properties like boundedness of the valence of the vertices. In fact, we are aware of only two cases when composite finite elements are *refinable*, i.e., provide stable local bases for certain multilevel triangulations. First, this is true for the triangulations obtained by the *Powell–Sabin* 12*-split*; see [50] for the relevant construction of stable local bases for C^1 quadratics and cubics. The other case is that of *FVS-triangulations* obtained from arbitrary strictly convex *quadrangulations* by adding two diagonals of each quadrilateral; see, e.g., [20, 43]. Here, a well-known composite finite element due to Fraeijs de Veubeke and Sander gives rise to a stable local basis for C^1 cubics, while for higher orders of differentiability only nonnested superspline-type constructions are known [40, 45, 46].

Appendix A.

Proof of Theorem 2.9. Denote briefly $\mathrm{N} := (\sum_{\theta\in\Theta} \|c_\theta\varphi_\theta\|_p^\tau)^{1/\tau}$.

Case 1. $0 < p \le 1$. Since $\tau < p \le 1$, we have

$$\Big\| \sum_{\theta\in\Theta} |c_\theta\varphi_\theta(\cdot)| \Big\|_p \le \left(\sum_{\theta\in\Theta} \|c_\theta\varphi_\theta\|_p^p \right)^{1/p} \le \left(\sum_{\theta\in\Theta} \|c_\theta\varphi_\theta\|_p^\tau \right)^{1/\tau}.$$

Case 2. $p = \infty$. Since $\tau = 1/\alpha \le 1$, then (2.42) is obvious.

Case 3. $1 < p < \infty$. We need the following lemma.

LEMMA A.1. *Let* $g := \sum_{\theta\in\mathcal{M}} |c_\theta\varphi_\theta|$, *where* $\#\mathcal{M} < \infty$ *and* $\|c_\theta\varphi_\theta\|_p \le L$ *for* $\theta \in \mathcal{M}$. *Then*

$$\|g\|_p \le cL(\#\mathcal{M})^{1/p},$$

with c *independent of* $\mathcal{M}$ *and* $(c_\theta)_{\theta\in\mathcal{M}}$.

Proof. Using the properties of Φ, we have (recall that $\operatorname{supp}\varphi_\theta \subset E_\theta := \operatorname{star}^\ell(v_\theta)$ and $\|\varphi_\theta\|_\infty \approx |E_\theta|^{-1/p}\|\varphi_\theta\|_p$ by (2.14))

$$\|g\|_p \le \Big\| \sum_{\theta\in\mathcal{M}} \|c_\theta\varphi_\theta\|_\infty \cdot \mathbb{1}_{E_\theta}(\cdot) \Big\|_p \le cL \Big\| \sum_{\theta\in\mathcal{M}} |E_\theta|^{-1/p} \cdot \mathbb{1}_{E_\theta}(\cdot) \Big\|_p.$$

We define $E := \bigcup_{\theta\in\mathcal{M}} E_\theta$ and $E(x) := \min\{|E_\theta| : \theta \in \mathcal{M} \text{ and } E_\theta \ni x\}$ for $x \in E$. By the properties of the LR-triangulations, it follows that

$$\sum_{\theta\in\mathcal{M}} |E_\theta|^{-1/p} \cdot \mathbb{1}_{E_\theta}(x) \le cE(x)^{-1/p}\mathbb{1}_E(x), \quad x \in \mathbb{R}^2.$$

On the other hand,

$$E(x)^{-1} = \max_{\theta\in\mathcal{M},\ E_\theta\ni x} |E_\theta|^{-1} \le \sum_{\theta\in\mathcal{M}} |E_\theta|^{-1}\mathbb{1}_{E_\theta}(x).$$

Therefore,

$$\begin{aligned}\|g\|_p &\le cL\|E(\cdot)^{-1/p}\|_{L_p} = cL\left(\int_E E(x)^{-1}\,dx \right)^{1/p} \\ &\le cL\left(\sum_{\theta\in\mathcal{M}} |E_\theta|^{-1} \int_{\mathbb{R}^2} \mathbb{1}_{E_\theta}(x)\,dx \right)^{1/p} = cL(\#\mathcal{M})^{1/p}. \qquad \square\end{aligned}$$

We define

$$\mathcal{F}_\mu := \{\theta : 2^{-\mu}\mathrm{N} \le \|c_\theta\varphi_\theta\|_p < 2^{-\mu+1}\mathrm{N}\},$$

where $\mathrm{N} := (\sum_{\theta\in\Theta} \|c_\theta\varphi_\theta\|_p^\tau)^{1/\tau}$. Then

$$\bigcup_{\nu\le\mu} \mathcal{F}_\nu = \{\theta \ : \ \|c_\theta\varphi_\theta\|_p \ge 2^{-\mu}\mathrm{N}\},$$

and hence

$$\#\mathcal{F}_\mu \le \sum_{\nu\le\mu} \#\mathcal{F}_\nu = \#\left(\bigcup_{\nu\le\mu}\mathcal{F}_\nu\right) \le 2^{\mu\tau}. \tag{A.1}$$

We set $F_\mu := \sum_{\theta\in\mathcal{F}_\mu} |c_\theta\varphi_\theta|$. Using Lemma A.1 and (A.1), we obtain

$$\begin{aligned}\Big\|\sum_{\theta\in\Theta}|c_\theta\varphi_\theta(\cdot)|\Big\|_p &\le \Big\|\sum_{\mu=0}^\infty F_\mu(\cdot)\Big\|_p \le \sum_{\mu=0}^\infty \|F_\mu\|_p \le c\sum_{\mu=0}^\infty 2^{-\mu}\mathrm{N}(\#\mathcal{F}_\mu)^{1/p}\\ &\le c\mathrm{N}\sum_{\mu=0}^\infty 2^{-\mu(1-\tau/p)} \le c\mathrm{N}\sum_{\mu=0}^\infty 2^{-\mu\tau\alpha} \le c\mathrm{N}.\end{aligned}$$

This completes the proof of Theorem 2.9. □

Proof of Theorem 3.1 (the case $0<p<\infty$). Suppose $f\in B^\alpha_\tau(\Phi)$, where $\alpha>0$, $1/\tau = \alpha+1/p$, $0<p<\infty$. By (2.40), f can be represented in the form $f=\sum_{\theta\in\Theta} b_\theta\varphi_\theta$ with the series converging absolutely a.e. in $\mathbb{R}^2$ and in L_p. We denote briefly $N(f) := N_{\Phi,Q,\tau}(f) := (\sum_{\theta\in\Theta}\|b_\theta\varphi_\theta\|_p^\tau)^{1/\tau} \approx \|f\|_{B^\alpha_\tau(\Phi)}$.

Suppose that $(b_{\theta_j}\varphi_{\theta_j})_{j=1}^\infty$ is a rearrangement of the sequence $(b_\theta\varphi_\theta)_{\theta\in\Theta}$ such that $\|b_{\theta_1}\varphi_{\theta_1}\|_p \ge \|b_{\theta_2}\varphi_{\theta_2}\|_p \ge\cdots$. Set $s_n := \sum_{j=1}^n b_{\theta_j}\varphi_{\theta_j}$, $s_n\in\Sigma_n(\Phi)$.

Case 1. $0<p\le 1$. To estimate $\|f-s_n\|_p$ we shall use the following simple inequality [38]: If $x_1\ge x_2\ge\cdots\ge 0$ and $0<\tau<p$, then

$$\left(\sum_{j=n+1}^\infty x_j^p\right)^{1/p} \le n^{1/p-1/\tau}\left(\sum_{j=1}^\infty x_j^\tau\right)^{1/\tau}. \tag{A.2}$$

We use Theorem 2.9 and apply (A.2) with $x_j := \|b_{\theta_j}\varphi_{\theta_j}\|_p$ to obtain

$$\begin{aligned}\|f-s_n\|_p \le \Big\|\sum_{j=n+1}^\infty |b_{\theta_j}\varphi_{\theta_j}|\Big\|_p &\le \left(\sum_{j=n+1}^\infty \|b_{\theta_j}\varphi_{\theta_j}\|_p^p\right)^{1/p}\\ &\le n^{1/p-1/\tau}\left(\sum_{j=1}^\infty \|b_{\theta_j}\varphi_{\theta_j}\|_p^\tau\right)^{1/\tau} = n^{-\alpha}N(f),\end{aligned}$$

which proves Theorem 3.1 in Case 1.

Case 2. $1<p<\infty$. We proceed quite similarly as in the proof of Theorem 2.9. We set $\mathcal{F}_\mu := \{\theta : 2^{-\mu}N(f)\le\|b_\theta\varphi_\theta\|_p < 2^{-\mu+1}N(f)\}$ and $F_\mu := \sum_{\theta\in\mathcal{F}_\mu}|b_\theta\varphi_\theta|$.

Fix $m\ge 1$ and set $M := \lceil 2^{m\tau}\rceil$. As in the proof of Theorem 2.9 (see (A.1)), $\#\mathcal{F}_m \le \sum_{\nu\le m}\#\mathcal{F}_\nu \le 2^{m\tau}\le M$. Using Lemma A.1, we obtain

$$\begin{aligned}\|f-s_M\|_p \le \Big\|\sum_{\mu=m+1}^\infty F_\mu\Big\|_p &\le \sum_{\mu=m+1}^\infty \|F_\mu\|_p\\ &\le c\sum_{\mu=m+1}^\infty 2^{-\mu}N(f)(\#\mathcal{F}_\mu)^{1/p} \le cN(f)\sum_{\mu=m+1}^\infty 2^{-\mu(1-\tau/p)}\\ &\le cN(f)2^{-m(1-\tau/p)} \le cM^{-1/\tau+1/p}N(f) = cM^{-\alpha}N(f).\end{aligned}$$

This estimate readily implies (3.1). □

Proof of Theorem 3.2. *Step* 1. With this step we lay some groundwork that is needed for the proof of the Bernstein inequality. Let $\mathcal{T}$ be an arbitrary LR-triangulation and suppose Λ is a finite subset of $\mathcal{T}$. The set Λ generates a certain tree structure that we want to bring up in what follows.

We say that $\triangle \in \mathcal{T}$ is a *branching triangle* if at least two children of $\triangle$ have descendants in Λ. Let $\tilde{\Lambda}$ denote the extension of Λ obtained by adding all branching triangles and all children of branching triangles if they are not already in Λ. By considering the *tree* of the ancestors of all triangles in Λ, it is not difficult to see that the total number of branching triangles does not exceed $\#\Lambda - 1$. Since the number of children of a triangle is bounded by M_0, we conclude that $\#\tilde{\Lambda} \le c\#\Lambda$.

Furthermore, for a later use in Step 3, we call $\triangle \in \mathcal{T} \setminus \tilde{\Lambda}$ a *chain triangle* if at least one of its descendants belongs to Λ. The set of all chain triangles will be denoted by Γ. By construction, for each $\triangle \in \Gamma$ there is a unique largest triangle $\tilde{\triangle} \in \tilde{\Lambda}$ contained in $\triangle$. We set $K_\triangle := \triangle \setminus \tilde{\triangle}$ and $\mu_\triangle := m - \tilde{m}$, where $\triangle \in \mathcal{T}_m$ and $\tilde{\triangle} \in \mathcal{T}_{\tilde{m}}$. We denote by $\tilde{\Gamma}$ the set of all $\triangle \in \Gamma$ for which there is a $\triangle' \in \tilde{\Lambda}$ containing $\triangle$. It is easy to see that $\tilde{\Gamma}$ is the disjoint union of *finite chains*, i.e., sets λ of the form $\lambda = \{\triangle_1, \dots, \triangle_\nu\} \subset \tilde{\Gamma}$ $(\nu \ge 1)$, where $\triangle''_\lambda \supset \triangle_1 \supset \cdots \supset \triangle_\nu \supset \triangle'_\lambda$ for some $\triangle'_\lambda, \triangle''_\lambda \in \tilde{\Lambda}$, and $\triangle_1$ is a child of $\triangle''_\lambda$, $\triangle_j$ is a child of $\triangle_{j-1}$, $\nu = 2, \dots, \nu$, and $\triangle'_\lambda$ is a child of $\triangle_\nu$. Similarly, $\Gamma \setminus \tilde{\Gamma}$ is the disjoint union of *infinite chains* $\lambda = \{\dots, \triangle_{-2}, \triangle_{-1}\} \subset \Gamma$, where $\cdots \supset \triangle_{-2} \supset \triangle_{-1} \supset \triangle'_\lambda$ for some $\triangle'_\lambda \in \tilde{\Lambda}$, and $\triangle_j$ is a child of $\triangle_{j-1}$, $\nu = -1, -2, \dots$, and $\triangle'_\lambda$ is a child of $\triangle_{-1}$. We let $\mathcal{L}$ and $\mathcal{L}^\infty$ denote the sets of all finite, respectively, infinite chains in Γ. Clearly, $\#\mathcal{L} \le \#\tilde{\Lambda}$ and $\#\mathcal{L}^\infty \le \#\tilde{\Lambda}$.

Step 2. For the proof of the theorem in the case $0 < p < \infty$, we need the following lemma.

LEMMA A.2. *Suppose* $s = \sum_{\triangle\in\Lambda} \mathbb{1}_\triangle \cdot P_\triangle$, *where* $P_\triangle \in \Pi_k$ $(k \ge 1)$, $\Lambda \subset \mathcal{T}$ *with* $\mathcal{T}$ *an LR-triangulation, and* $\#\Lambda < \infty$. *Then*

$$\left(\sum_{\triangle\in\Lambda} |\triangle|^{-\alpha\tau} \|s\|^\tau_{L_\tau(\triangle)} \right)^{1/\tau} \le c(\#\Lambda)^\alpha \|s\|_p,$$

with c *independent of* s *and* Λ.

Proof. We adopt all necessary notation from Step 1 above with Λ from the hypotheses of the lemma. Since $\#\tilde{\Lambda} \le c\#\Lambda$ and $s = \sum_{\triangle\in\tilde{\Lambda}} \mathbb{1}_\triangle \cdot P_\triangle$, where $P_\triangle = 0$ for $\triangle \in \tilde{\Lambda}\setminus\Lambda$, we may assume without loss of generality that $\tilde{\Lambda} = \Lambda$; i.e., the branching triangles and their children are contained in Λ.

Let $\triangle_1, \dots, \triangle_m$ be all nonbranching triangles in Λ. It is not difficult to see that for each of them there are only two possibilities: either $\triangle_i$ does not contain any other $\triangle \in \Lambda$ (in which case we call $\triangle_i$ a *final triangle*) or there is a unique largest triangle $\tilde{\triangle}_i \in \Lambda$ strictly contained in $\triangle_i$. We define the *rings* $K_i := \triangle_i \setminus \tilde{\triangle}_i$, $i = 1, \dots, m$, where $\tilde{\triangle}_i := \emptyset$ for a final triangle $\triangle_i$. Obviously, K_i have pairwise disjoint interiors, and $s|_{K_i} = P_i|_{K_i}$, for some $P_i \in \Pi_k$, $i = 1, \dots, m$. Since all children of branching triangles are in Λ, we have for each $\triangle \in \Lambda$,

$$\triangle = \bigcup_{\substack{i=1\\ \triangle_i\subset\triangle}}^{m} K_i \quad \text{and} \quad s|_\triangle = \sum_{\substack{i=1\\ \triangle_i\subset\triangle}}^{m} \mathbb{1}_{K_i} \cdot P_i.$$

Therefore,

$$\begin{aligned}\sum_{\triangle\in\Lambda}|\triangle|^{-\alpha\tau}\|s\|^\tau_{L_\tau(\triangle)} &= \sum_{\triangle\in\Lambda}|\triangle|^{-\alpha\tau}\sum_{\substack{i=1\\ \triangle_i\subset\triangle}}^m \|s\|^\tau_{L_\tau(K_i)}\\ &= \sum_{i=1}^m \|s\|^\tau_{L_\tau(K_i)}\sum_{\triangle\in\Lambda,\,\triangle\supset\triangle_i}|\triangle|^{-\alpha\tau}\\ &= \sum_{i=1}^m \|s\|^\tau_{L_\tau(K_i)}|\triangle_i|^{-\alpha\tau}\sum_{\triangle\in\Lambda,\,\triangle\supset\triangle_i}(|\triangle_i|/|\triangle|)^{\alpha\tau}\\ &\le c\sum_{i=1}^m \|s\|^\tau_{L_\tau(K_i)}|\triangle_i|^{-\alpha\tau},\end{aligned}$$

where we once switched the order of summation and used (2.9). Since $|\tilde\triangle_i|\le\rho|\triangle_i|$, we have by Lemma 2.2,

$$\|P_i\|_{L_\tau(K_i)}\approx|K_i|^{1/\tau-1/p}\|P_i\|_{L_p(K_i)}\approx|\triangle_i|^\alpha\|P_i\|_{L_p(K_i)},$$

which implies $\|s\|^\tau_{L_\tau(K_i)}|\triangle_i|^{-\alpha\tau}\approx\|s\|^\tau_{L_p(K_i)}$, $i=1,\dots,m$. Now by Hölder's inequality,

$$\sum_{i=1}^m\|s\|^\tau_{L_p(K_i)}\le\left(\sum_{i=1}^m\|s\|^p_{L_p(K_i)}\right)^{\tau/p}m^{1-\tau/p}\le(\#\Lambda)^{\alpha\tau}\|s\|^\tau_p,$$

and the proof is complete. □

Step 3. Let $s\in\Sigma_n(\Phi)$ and suppose that $s=:\sum_{\theta\in\mathcal{M}}c_\theta\varphi_\theta$, where $\mathcal{M}\subset\Theta(\mathcal{T})$ and $\#\mathcal{M}\le n$. Let Λ be the set of all triangles $\triangle\in\mathcal{T}$ which are involved in all $E_\theta:=\operatorname{supp}\varphi_\theta$, $\theta\in\mathcal{M}$. Then $s=\sum_{\triangle\in\Lambda}s_\triangle$, where $s_\triangle=:\mathbb{1}_\triangle\cdot P_\triangle$, $P_\triangle\in\Pi_k$. Evidently, by (2.7), $\#\Lambda\le c^\star(N_0,\ell)\,\#\mathcal{M}\le cn$.

We first extend Λ to $\tilde\Lambda$ as in Step 1 above and introduce some auxiliary sets of triangles needed for the forthcoming arguments. We set

$$\begin{aligned}\tilde\Lambda^*_m&:=\{\triangle\in\mathcal{T}_m:\Omega^\ell_\triangle\supset\triangle'\ \text{for some}\ \triangle'\in\tilde\Lambda\cap\mathcal{T}_m\},\\ \tilde\Lambda^{**}_m&:=\{\triangle\in\mathcal{T}_m:\Omega^{2\ell}_\triangle\supset\triangle'\ \text{for some}\ \triangle'\in\tilde\Lambda\cap\mathcal{T}_m\},\qquad m\in\mathbb{Z},\end{aligned}$$

and also

$$\tilde\Lambda^*:=\bigcup_{m\in\mathbb{Z}}\tilde\Lambda^*_m,\qquad\tilde\Lambda^{**}:=\bigcup_{m\in\mathbb{Z}}\tilde\Lambda^{**}_m.$$

Note that $\triangle,\triangle'\in\mathcal{T}_m$ and $\triangle'\subset\Omega^\ell_\triangle$ imply $\triangle\subset\Omega^\ell_{\triangle'}$, and hence

$$\tilde\Lambda^*_m=\{\triangle\in\mathcal{T}_m:\triangle\subset\Omega^\ell_{\triangle'}\ \text{for some}\ \triangle'\in\tilde\Lambda\cap\mathcal{T}_m\}.$$

Therefore, by (2.8), $\#\tilde\Lambda^*_m\le c^{\star\star}(N_0,\ell)\,\#(\tilde\Lambda\cap\mathcal{T}_m)$, and it follows that $\#\tilde\Lambda^*\le cn$. Similarly, $\#\tilde\Lambda^{**}\le c^{\star\star}(N_0,2\ell)(\#\tilde\Lambda)\le cn$. It is clear that $\tilde\Lambda\subset\tilde\Lambda^*\subset\tilde\Lambda^{**}$.

We now proceed to estimate $|s|^\tau_{B^\alpha_\tau(\mathcal{T})}:=\sum_{\triangle\in\mathcal{T}}|\triangle|^{-\alpha\tau}\mathbb{S}_\triangle(s)^\tau_\tau$. Let

$$s_m:=\sum_{\mu\le m}\sum_{\theta\in\mathcal{M}\cap\Theta_\mu}c_\theta\varphi_\theta,\qquad m\in\mathbb{Z}.$$

Then $s_m \in \mathcal{S}_m$, and hence $\mathbb{S}_\triangle(s)_\tau = \mathbb{S}_\triangle(s - s_m)_\tau$ if $\triangle \in \mathcal{T}_m$. For each $\triangle \in \mathcal{T}$, we shall use one of the following two obvious bounds for $\mathbb{S}_\triangle(s)_\tau$:

$$\mathbb{S}_\triangle(s)_\tau \le \|s\|_{L_\tau(\Omega^\ell_\triangle)}, \tag{A.3}$$

$$\mathbb{S}_\triangle(s)_\tau \le \|s - s_m\|_{L_\tau(\Omega^\ell_\triangle)}, \qquad \triangle \in \mathcal{T}_m. \tag{A.4}$$

Namely, (A.3) will be applied to the triangles $\triangle$ in the set $\tilde\Lambda^* \subset \mathcal{T}$ defined above, while (A.4) will be used for all remaining triangles in $\mathcal{T}$.

For the next estimates, we shall consider separately the cases $0 < p < \infty$ and $p = \infty$.

Case 1. $0 < p < \infty$. We consider two possibilities for each $\triangle \in \mathcal{T}$: $\triangle \in \tilde\Lambda^*$ or $\triangle \in \mathcal{T} \setminus \tilde\Lambda^*$.

(a) If $\triangle \in \tilde\Lambda^*_m$, then for each $\triangle' \in \mathcal{T}_m$ such that $\triangle' \subset \Omega^\ell_\triangle$, we have $\triangle' \in \tilde\Lambda^{**}_m$ and, in view of (2.2), $|\triangle'| \le c|\triangle|$. Hence, by (A.3),

$$\begin{aligned}
\sum_{\triangle\in\tilde\Lambda^*_m} |\triangle|^{-\alpha\tau}\mathbb{S}_\triangle(s)^\tau_\tau &\le \sum_{\triangle\in\tilde\Lambda^*_m} |\triangle|^{-\alpha\tau} \sum_{\triangle'\in\tilde\Lambda^{**}_m,\, \triangle'\subset\Omega^\ell_\triangle} \|s\|^\tau_{L_\tau(\triangle')} \\
&\le c \sum_{\triangle\in\tilde\Lambda^*_m} \sum_{\triangle'\in\tilde\Lambda^{**}_m,\, \triangle'\subset\Omega^\ell_\triangle} |\triangle'|^{-\alpha\tau}\|s\|^\tau_{L_\tau(\triangle')} \\
&= c \sum_{\triangle'\in\tilde\Lambda^{**}_m} \sum_{\triangle\in\tilde\Lambda^*_m,\, \Omega^\ell_\triangle\supset\triangle'} |\triangle'|^{-\alpha\tau}\|s\|^\tau_{L_\tau(\triangle')} \\
&\le c \sum_{\triangle'\in\tilde\Lambda^{**}_m} |\triangle'|^{-\alpha\tau}\|s\|^\tau_{L_\tau(\triangle')},
\end{aligned}$$

where we have switched the order of summation and taken into account the fact that $\#\{\triangle \in \tilde\Lambda^*_m : \Omega^\ell_\triangle \supset \triangle'\} = \#\{\triangle \in \tilde\Lambda^*_m : \triangle \subset \Omega^\ell_{\triangle'}\} \le c^{\star\star}(N_0, \ell)$, by (2.8). It follows that

$$\begin{aligned}
\sum_{\triangle\in\tilde\Lambda^*} |\triangle|^{-\alpha\tau}\mathbb{S}_\triangle(s)^\tau_\tau &\le c \sum_{\triangle\in\tilde\Lambda^{**}} |\triangle|^{-\alpha\tau}\|s\|^\tau_{L_\tau(\triangle)} \\
&\le c(\#\tilde\Lambda^{**})^{\alpha\tau}\|s\|^\tau_p \le cn^{\alpha\tau}\|s\|^\tau_p,
\end{aligned} \tag{A.5}$$

where we applied Lemma A.2 to s with Λ replaced by $\tilde\Lambda^{**}$, which is obviously legitimate since $\tilde\Lambda^{**} \supset \Lambda$.

(b) Now suppose $\triangle \in \mathcal{T}_m \setminus \tilde\Lambda^*_m$. Then $\Omega^\ell_\triangle = \bigcup_{j=1}^{n_\triangle} \triangle_j$ for some $\triangle_j \in \mathcal{T}_m \setminus \tilde\Lambda$, $j = 1, \dots, n_\triangle$, with $n_\triangle \le c^{\star\star} \le 3N_0^{2\ell-1}$ (see (2.8)). We have, using (A.4),

$$\mathbb{S}_\triangle(s)^\tau_\tau = \mathbb{S}_\triangle(s - s_m)^\tau_\tau \le \sum_{j=1}^{n_\triangle} \|s - s_m\|^\tau_{L_\tau(\triangle_j)}. \tag{A.6}$$

If $\triangle_j \notin \Gamma$, then it has no descendants in Λ, and hence $s|_{\triangle_j} = s_m|_{\triangle_j}$, and

$$\|s - s_m\|_{L_\tau(\triangle_j)} = 0, \qquad \triangle_j \notin \Gamma. \tag{A.7}$$

Suppose $\triangle_j \in \Gamma$; i.e., it is a chain triangle. Let $\tilde\triangle_j$ be the unique largest triangle of $\tilde\Lambda$ contained in $\triangle_j$, and let $K_{\triangle_j} = \triangle_j \setminus \tilde\triangle_j$ and $\mu_{\triangle_j} = m - \tilde m$ be defined as in Step 1.

It is clear that in this case $s|_{K_{\triangle_j}} = s_m|_{K_{\triangle_j}} = \mathbb{1}_{K_{\triangle_j}} \cdot P_{\triangle_j}$ and $s_m|_{\triangle_j} = \mathbb{1}_{\triangle_j} \cdot P_{\triangle_j}$ for some $P_{\triangle_j} \in \Pi_k$. Therefore,

$$\|s - s_m\|^\tau_{L_\tau(\triangle_j)} = \|s - s_m\|^\tau_{L_\tau(\tilde\triangle_j)} \le c\|s\|^\tau_{L_\tau(\tilde\triangle_j)} + c\|P_{\triangle_j}\|^\tau_{L_\tau(\tilde\triangle_j)}.$$

If $\triangle_j \in \Gamma \setminus \tilde\Gamma$, then clearly $s_m|_{\triangle_j} = 0$, and we have

$$\|s - s_m\|_{L_\tau(\triangle_j)} = \|s\|_{L_\tau(\tilde\triangle_j)}, \qquad \triangle_j \in \Gamma \setminus \tilde\Gamma. \tag{A.8}$$

Assume that $\triangle_j \in \tilde\Gamma$. By Lemma 2.2,

$$\begin{aligned}\|P_{\triangle_j}\|^\tau_{L_\tau(\tilde\triangle_j)} &\le |\tilde\triangle_j|\|P_{\triangle_j}\|^\tau_{L_\infty(\triangle_j)} \le c|\tilde\triangle_j|\|P_{\triangle_j}\|^\tau_{L_\infty(K_{\triangle_j})}\\ &\le c|\tilde\triangle_j||K_{\triangle_j}|^{-\tau/p}\|P_{\triangle_j}\|^\tau_{L_p(K_{\triangle_j})} \le c|\tilde\triangle_j||\triangle_j|^{\alpha\tau-1}\|s\|^\tau_{L_p(K_{\triangle_j})}.\end{aligned}$$

By (2.1), $|\tilde\triangle_j| \le \rho^{\mu_{\triangle_j}}|\triangle_j|$, and we arrive at the inequality

$$\|s - s_m\|^\tau_{L_\tau(\triangle_j)} \le c\|s\|^\tau_{L_\tau(\tilde\triangle_j)} + c\rho^{\mu_{\triangle_j}}|\triangle_j|^{\alpha\tau}\|s\|^\tau_{L_p(K_{\triangle_j})}, \qquad \triangle_j \in \tilde\Gamma. \tag{A.9}$$

From (A.6)–(A.9) and (2.2), we obtain

$$\begin{aligned}\sum_{\triangle\in\mathcal{T}\setminus\tilde\Lambda^*} |\triangle|^{-\alpha\tau}\mathbb{S}_\triangle(s)^\tau_\tau &= \sum_{m\in\mathbb{Z}}\sum_{\triangle\in\mathcal{T}_m\setminus\tilde\Lambda^*_m} |\triangle|^{-\alpha\tau}\mathbb{S}_\triangle(s)^\tau_\tau\\ &\le c\sum_{\triangle\in\Gamma}|\triangle|^{-\alpha\tau}\|s\|^\tau_{L_\tau(\tilde\triangle)} + c\sum_{\triangle\in\tilde\Gamma}\rho^{\mu_\triangle}\|s\|^\tau_{L_p(K_\triangle)}\\ &=: \Sigma_1 + \Sigma_2.\end{aligned}$$

Trivially,

$$\|s\|_{L_\tau(\tilde\triangle)} \le \sum_{\triangle'\in\tilde\Lambda,\,\triangle'\subset\triangle}\|s\|_{L_\tau(\triangle')}, \qquad \triangle\in\Gamma.$$

Switching the order of summation, we find

$$\begin{aligned}\Sigma_1 &\le c\sum_{\triangle'\in\tilde\Lambda}\|s\|^\tau_{L_\tau(\triangle')}\sum_{\triangle\in\Gamma,\,\triangle\supset\triangle'}|\triangle|^{-\alpha\tau}\\ &\le c\sum_{\triangle'\in\tilde\Lambda}\|s\|^\tau_{L_\tau(\triangle')}|\triangle'|^{-\alpha\tau}\sum_{\triangle\in\Gamma,\,\triangle\supset\triangle'}(|\triangle'|/|\triangle|)^{\alpha\tau}\\ &\le c\sum_{\triangle'\in\tilde\Lambda}|\triangle'|^{-\alpha\tau}\|s\|^\tau_{L_\tau(\triangle')} \le c(\#\tilde\Lambda)^{\alpha\tau}\|s\|^\tau_p,\end{aligned} \tag{A.10}$$

where we also used (2.9) and applied Lemma A.2 to s with Λ replaced by $\tilde\Lambda$.

To estimate Σ_2 we shall use the representation of $\tilde\Gamma$ as a disjoint union of chains: $\tilde\Gamma = \bigcup_{\lambda\in\mathcal{L}}\lambda$. Let $\lambda\in\mathcal{L}$ and suppose $\lambda = \{\triangle_1,\dots,\triangle_\nu\}$, where $\triangle''_\lambda \supset \triangle_1 \supset \cdots \supset \triangle_\nu \supset \triangle'_\lambda$ with $\triangle'_\lambda, \triangle''_\lambda \in \tilde\Lambda$. Then $\mu_{\triangle_i} \ge \nu - i + 1$. Therefore,

$$\sum_{\triangle\in\lambda}\rho^{\mu_\triangle}\|s\|^\tau_{L_p(K_\triangle)} \le \|s\|^\tau_{L_p(\triangle''_\lambda\setminus\triangle'_\lambda)}\sum_{j=1}^{\nu}\rho^{\nu-j+1} \le c\|s\|^\tau_{L_p(K_\lambda)},$$

where $K_\lambda := \triangle''_\lambda \setminus \triangle'_\lambda$. It is easy to see that the sets K_λ, $\lambda \in \mathcal{L}$, have pairwise disjoint interiors. Summing over all $\lambda \in \mathcal{L}$, we obtain by Hölder's inequality

$$\Sigma_2 \le c \sum_{\lambda\in\mathcal{L}} \|s\|^\tau_{L_p(K_\lambda)} \le c\left(\sum_{\lambda\in\mathcal{L}} \|s\|^p_{L_p(K_\lambda)} \right)^{\tau/p} (\#\mathcal{L})^{1-\tau/p} \le c(\#\tilde\Lambda)^{\alpha\tau}\|s\|^\tau_p.$$

From this estimate and (A.10), we find

$$\sum_{\triangle\in\mathcal{T}\setminus\tilde\Lambda^*} |\triangle|^{-\alpha\tau}\mathbb{S}_\triangle(s)^\tau_\tau \le c(\#\tilde\Lambda)^{\alpha\tau}\|s\|^\tau_p \le cn^{\alpha\tau}\|s\|^\tau_p.$$

Combining this with (A.5) gives $\|s\|^\tau_{B^\alpha_\tau(\Phi)} \le cn^{\alpha\tau}\|s\|^\tau_p$; i.e., (3.2) holds.

Case 2. $p = \infty$. The proof in this case is easier. We consider as before two possibilities for each $\triangle \in \mathcal{T}$: $\triangle \in \tilde\Lambda^*$ or $\triangle \in \mathcal{T} \setminus \tilde\Lambda^*$.

(a) For $\triangle \in \tilde\Lambda^*$, we obtain by (2.2)

$$|\triangle|^{-1}\mathbb{S}_\triangle(s)^\tau_\tau \le |\triangle|^{-1}\|s\|^\tau_{L_\tau(\Omega^\ell_\triangle)} \le |\triangle|^{-1}|\Omega^\ell_\triangle|\|s\|^\tau_\infty \le c\|s\|^\tau_\infty.$$

Therefore,

$$\sum_{\triangle\in\tilde\Lambda^*} |\triangle|^{-1}\mathbb{S}_\triangle(s)^\tau_\tau \le c\|s\|^\tau_\infty(\#\tilde\Lambda^*) \le cn\|s\|^\tau_\infty. \tag{A.11}$$

(b) Let $\triangle \in \mathcal{T}_m \setminus \tilde\Lambda^*_m$. Then $\Omega^\ell_\triangle =: \bigcup_{j=1}^{n_\triangle} \triangle_j$ for some $\triangle_j \in \mathcal{T}_m \setminus \tilde\Lambda$, $j = 1, \dots, n_\triangle$, with $n_\triangle \le c^{\star\star} < 3N_0^{2\ell-1}$ (see (2.8)). We have (see (A.4))

$$\mathbb{S}_\triangle(s)^\tau_\tau = \mathbb{S}_\triangle(s - s_m)^\tau_\tau \le \sum_{j=1}^{n_\triangle} \|s - s_m\|^\tau_{L_\tau(\triangle_j)}.$$

As in Case 1, if $\triangle_j \notin \Gamma$, then $\|s - s_m\|_{L_\tau(\triangle_j)} = 0$, and if $\triangle_j \in \Gamma$, then $s|_{K_{\triangle_j}} = s_m|_{K_{\triangle_j}} = \mathbb{1}_{K_{\triangle_j}} \cdot P_{\triangle_j}$ and $s_m|_{\triangle_j} = \mathbb{1}_{\triangle_j} \cdot P_{\triangle_j}$ for some $P_{\triangle_j} \in \Pi_k$. Therefore,

$$\begin{aligned}\|s - s_m\|^\tau_{L_\tau(\triangle_j)} &= \|s - s_m\|^\tau_{L_\tau(\tilde\triangle_j)} \\ &\le c|\tilde\triangle_j|(\|s\|^\tau_\infty + \|P_{\triangle_j}\|^\tau_{L_\infty(\tilde\triangle_j)}) \le c|\tilde\triangle_j|\|s\|^\tau_\infty,\end{aligned}$$

where we used the inequalities $\|P_{\triangle_j}\|_{L_\infty(\tilde\triangle_j)} \le \|P_{\triangle_j}\|_{L_\infty(\triangle_j)} \le c\|P_{\triangle_j}\|_{L_\infty(K_{\triangle_j})} \le c\|s\|_\infty$ (see Lemma 2.2). From the above, we infer by (2.2)

$$|\triangle|^{-1}\mathbb{S}_\triangle(s)^\tau_\tau \le c\|s\|^\tau_\infty \sum_{1\le j\le n_\triangle,\, \triangle_j\in\Gamma\cap\mathcal{T}_m} |\tilde\triangle_j|/|\triangle_j|,$$

and hence, using (2.2) and the fact that each $\triangle' \in \Gamma \cap \mathcal{T}_m$ can belong to $\le c^{\star\star}$ sets $\Omega^\ell_\triangle$, we obtain

$$\begin{aligned}\sum_{\triangle\in\mathcal{T}_m\setminus\tilde\Lambda^*_m} |\triangle|^{-1}\mathbb{S}_\triangle(s)^\tau_\tau &\le c\|s\|^\tau_\infty \sum_{\triangle\in\Gamma\cap\mathcal{T}_m} |\tilde\triangle|/|\triangle| \\ &\le c\|s\|^\tau_\infty \sum_{\triangle\in\Gamma\cap\mathcal{T}_m} \rho^{\mu_\triangle}.\end{aligned}$$

Summing over $m \in \mathbb{Z}$, we find

$$\sum_{\triangle\in\mathcal{T}\setminus\tilde{\Lambda}^*} |\triangle|^{-1}\mathbb{S}_\triangle(s)_\tau^\tau \le c\|s\|_\infty^\tau \sum_{\triangle\in\Gamma} \rho^{\mu_\triangle} \le c\|s\|_\infty^\tau\,(\#\mathcal{L}+\#\mathcal{L}^\infty) \le cn\|s\|_\infty^\tau.$$

We couple this with (A.11) to obtain $\|s\|_{B_\tau^\alpha(\mathcal{T})}^\tau \le cn\|s\|_\infty^\tau$, which is (3.2). ∎

REFERENCES

[1] P. Alfeld, B. Piper, and L. L. Schumaker, *Minimally supported bases for spaces of bivariate piecewise polynomials of smoothness r and degree d ≥ 4r+1*, Comput. Aided Geom. Design, 4 (1987), pp. 105–123.

[2] I. Babuška and A. K. Aziz, *On the angle condition in the finite element method*, SIAM J. Numer. Anal., 13 (1976), pp. 214–226.

[3] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction*, Grundlehren Math. Wiss. 223, Springer-Verlag, Berlin, New York, 1976.

[4] C. Bennett and R. Sharpley, *Interpolation of Operators*, Pure Appl. Math. 129, Academic Press, Boston, 1988.

[5] P. Binev and R. DeVore, *Fast Computation in Adaptive Tree Approximation*, http://www.math.sc.edu/∼imip/02.html.

[6] P. Binev, W. Dahmen, and R. DeVore, *Adaptive Finite Element Methods with Convergence Rates*, IGPM Report 219, RWTH Aachen, Aachen, Germany, 2002, http://elc2.igpm.rwth-aachen.de/∼dahmen/.

[7] P. Binev, W. Dahmen, R. DeVore, and P. Petrushev, *Approximation classes for adaptive methods*, Serdica Math. J., 28 (2002), pp. 391–416.

[8] C. de Boor and K. Höllig, *Approximation power of smooth bivariate pp functions*, Math. Z., 197 (1988), pp. 343–363.

[9] C. de Boor, K. Höllig, and S. D. Riemenschneider, *Box Splines*, Springer-Verlag, New York, 1993.

[10] C. de Boor and R. Q. Jia, *A sharp upper bound on the approximation order of smooth bivariate pp functions*, J. Approx. Theory, 72 (1993), pp. 24–33.

[11] J. Bramble and X. Zhang, *Multigrid methods for the biharmonic problem discretized by conforming C^1 finite elements on nonnested meshes*, Numer. Funct. Anal. Optim., 1 (1995), pp. 835–846.

[12] S. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Texts Appl. Math. 15, Springer-Verlag, New York, 1994.

[13] M. D. Buhmann, O. Davydov, and T. N. T. Goodman, *Cubic spline prewavelets on the four-directional mesh*, Found. Comp. Math., 3 (2003), pp. 113–133.

[14] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[15] C. K. Chui, *Multivariate Splines*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 54, SIAM, Philadelphia, 1988.

[16] C. K. Chui, D. Hong, and R.-Q. Jia, *Stability of optimal order approximation by bivariate splines over arbitrary triangulations*, Trans. Amer. Math. Soc., 347 (1995), pp. 3301–3318.

[17] C. K. Chui and M.-J. Lai, *Multivariate vertex splines and finite elements*, J. Approx. Theory, 60 (1990), pp. 245–343.

[18] C. K. Chui and M.-J. Lai, *On bivariate super vertex splines*, Constr. Approx., 6 (1990), pp. 399–419.

[19] W. Dahmen and C. A. Micchelli, *On the local linear independence of translates of a box spline*, Studia Math., 82 (1985), pp. 243–262.

[20] W. Dahmen, P. Oswald, and X.-Q. Shi, *C^1-hierarchical bases*, J. Comput. Appl. Math., 51 (1994), pp. 37–56.

[21] W. Dahmen and R. Stevenson, *Element-by-element construction of wavelets satisfying stability and moment conditions*, SIAM J. Numer. Anal., 37 (1999), pp. 319–352.

[22] O. Davydov, *Stable local bases for multivariate spline spaces*, J. Approx. Theory, 111 (2001), pp. 267–297.

[23] O. Davydov, *On the computation of stable local bases for bivariate polynomial splines*, in Trends in Approximation Theory, K. Kopotun, T. Lyche, and M. Neamtu, eds., Vanderbilt University Press, Nashville, TN, 2001, pp. 85–94.

[24] O. Davydov, *Locally stable spline bases on nested triangulations*, in Approximation Theory X: Wavelets, Splines, and Applications, C. K. Chui, L. L. Schumaker, and J. Stöckler, eds., Vanderbilt University Press, Nashville, TN, 2002, pp. 231–240.
[25] O. Davydov, G. Nürnberger, and F. Zeilfelder, *Bivariate spline interpolation with optimal approximation order*, Constr. Approx., 17 (2001), pp. 181–208.
[26] O. Davydov and L. L. Schumaker, *Stable local nodal bases for C^1 bivariate polynomial splines*, in Curve and Surface Fitting: Saint-Malo 1999, A. Cohen, C. Rabut, and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 2000, pp. 171–180.
[27] O. Davydov and L. L. Schumaker, *On stable local bases for bivariate polynomial spline spaces*, Constr. Approx., 18 (2002), pp. 87–116.
[28] O. Davydov and F. Zeilfelder, *Scattered data fitting by direct extension of local polynomials with bivariate splines*, Adv. Comput. Math., to appear.
[29] R. DeVore, B. Jawerth, and V. Popov, *Compression of wavelet decompositions*, Amer. J. Math., 114 (1992), pp. 737–785.
[30] R. DeVore, P. Petrushev, and X. Yu, *Nonlinear wavelet approximation in the space $C(\mathbb{R}^d)$*, in Progress in Approximation Theory, A. A. Gonchar and E. B. Saff, eds., Springer-Verlag, New York, 1992, pp. 261–283.
[31] R. DeVore and V. Popov, *Interpolation spaces and nonlinear approximation*, in Function Spaces and Applications (Lund, 1986), Lecture Notes in Math. 1302, Springer-Verlag, Berlin, 1988, pp. 191–205.
[32] R. DeVore and V. Popov, *Interpolation of Besov spaces*, Trans. Amer. Math. Soc., 305 (1988), pp. 297–314.
[33] M. Dæhlen, T. Lyche, K. Mørken, R. Schneider, and H.-P. Seidel, *Multiresolution analysis over triangles, based on quadratic Hermite interpolation*, J. Comput. Appl. Math., 119 (2000), pp. 97–114.
[34] M. S. Floater and E. G. Quak, *Linear independence and stability of piecewise linear prewavelets on arbitrary triangulations*, SIAM J. Numer. Anal., 38 (2000), pp. 58–79.
[35] D. Hong, *Spaces of bivariate spline functions over triangulation*, Approx. Theory Appl., 7 (1991), pp. 56–75.
[36] A. Kh. Ibrahim and L. L. Schumaker, *Super spline spaces of smoothness r and degree $d \geq 3r+2$*, Constr. Approx., 7 (1991), pp. 401–423.
[37] W. Jianzhong, *On dual basis of bivariate box spline*, Approx. Theory Appl., 3 (1987), pp. 153–163.
[38] B. Karaivanov and P. Petrushev, *Nonlinear piecewise polynomial approximation beyond Besov spaces*, Appl. Comput. Harmon. Anal., to appear.
[39] B. Karaivanov, P. Petrushev, and R. Sharpley, *Algorithms for nonlinear piecewise polynomial approximation: Theoretical aspects*, Trans. Amer. Math. Soc., 355 (2003), pp. 2585–2631.
[40] M. Laghchim-Lahlou and P. Sablonnière, *Quadrilateral finite elements of FVS type and class C^ρ*, Numer. Math., 70 (1995), pp. 229–243.
[41] M.-J. Lai, *A remark on translates of a box spline*, Approx. Theory Appl., 5 (1989), pp. 97–104.
[42] M.-J. Lai, *Approximation order from bivariate C^1-cubics on a four-directional mesh is full*, Comput. Aided Geom. Design, 11 (1994), pp. 215–223.
[43] M.-J. Lai, *Bivariate spline spaces on FVS-triangulations*, in Approximation Theory VIII, Vol. 1: Approximation and Interpolation, C. K. Chui and L. L. Schumaker, eds., World Scientific, Singapore, 1995, pp. 309–316.
[44] M.-J. Lai and L. L. Schumaker, *On the approximation power of bivariate splines*, Adv. Comput. Math., 9 (1998), pp. 251–279.
[45] M.-J. Lai and L. L. Schumaker, *On the approximation power of splines on triangulated quadrangulations*, SIAM J. Numer. Anal., 36 (1999), pp. 143–159.
[46] M.-J. Lai and L. L. Schumaker, *Quadrilateral macroelements*, SIAM J. Math. Anal., 33 (2002), pp. 1107–1116.
[47] T. Lyche, K. Mørken, and E. Quak, *Theory and algorithms for non-uniform spline wavelets*, in Multivariate Approximation and Applications, N. Dyn, D. Leviatan, D. Levin, and A. Pinkus, eds., Cambridge University Press, Cambridge, UK, 2001, pp. 152–187.
[48] J. Morgan and R. Scott, *A nodal basis for C^1 piecewise polynomials of degree $n \geq 5$*, Math. Comp., 29 (1975), pp. 736–740.
[49] G. Nürnberger, L. L. Schumaker, and F. Zeilfelder, *Local Lagrange interpolation by bivariate C^1 cubic splines*, in Mathematical Methods in CAGD: Oslo 2000, T. Lyche and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 2001, pp. 393–404.
[50] P. Oswald, *Hierarchical conforming finite element methods for the biharmonic equation*, SIAM J. Numer. Anal., 29 (1992), pp. 1610–1625.

[51] P. Oswald, *Multilevel Finite Element Approximation*, Teubner, Stuttgart, 1994.
[52] J. Peetre and G. Sparr, *Interpolation of normed Abelian groups*, Ann. Mat. Pura Appl. (4), 92 (1972), pp. 217–262.
[53] P. Petrushev, *Direct and converse theorems for rational and spline approximation and Besov spaces*, in Function Spaces and Applications (Lund, 1986), Lecture Notes in Math. 1302, Springer-Verlag, Berlin, 1988, pp. 363–377.
[54] P. Petrushev, *Multivariate n-term rational and piecewise polynomial approximation*, J. Approx. Theory, 121 (2003), pp. 158–197.
[55] P. Petrushev and V. Popov, *Rational Approximation of Real Functions*, Cambridge University Press, Cambridge, UK, 1987.
[56] P. Sablonnière, *Quasi-interpolants associated with H-splines on a three direction mesh*, J. Comput. Appl. Math., 66 (1996), pp. 433–442.
[57] L. L. Schumaker, *On super splines and finite elements*, SIAM J. Numer. Anal., 26 (1989), pp. 997–1005.
[58] R. Stevenson, *Piecewise linear (pre-)wavelets on non-uniform meshes*, in Multigrid Methods, V. W. Hackbusch and G. Wittum, eds., Springer-Verlag, Berlin, 1998, pp. 306–319.

SIAM J. MATH. ANAL.
Vol. 35, No. 3, pp. 759–805

SEQUENCES OF SINGULARLY PERTURBED FUNCTIONALS GENERATING FREE-DISCONTINUITY PROBLEMS*

MASSIMILIANO MORINI†

Abstract. We prove that a wide class of singularly perturbed functionals generates as Γ-limit a functional related to a free-discontinuity problem. Several applications of the result are shown.

Key words. free-discontinuity problems, singular perturbations, Γ-convergence

AMS subject classification. 49K10

DOI. 10.1137/S0036141001395388

1. Introduction. Many models in the fields of fracture mechanics and computer vision lead to free-discontinuity problems, that is, to the minimization of functionals defined in spaces of discontinuous functions (namely, BV and SBV) involving energies with a bulk part and a surface part concentrated along the (free-)discontinuity zone. This paper is concerned with the variational approximation in the sense of De Giorgi's Γ-convergence of such energies by smooth functionals defined in Sobolev spaces. The issue of finding a smooth approximation of free-discontinuity problems is important for two main reasons:

(a) the numerical treatment for regular functionals defined in Sobolev spaces is much easier;
(b) the existence of such an approximation allows for the definition of a parabolic evolution model as a limit of the gradient flows of the approximating functionals.

For a general survey on free-discontinuity problems and their approximation we refer to [10] and [6].

When the volume part of the energy is given by $\int_\Omega |\nabla u|^2\,dx$, heuristic considerations suggest using, as approximating functionals, energies of the form

$$\frac{1}{\varepsilon}\int_\Omega f(\sqrt{\varepsilon}|\nabla u|)\,dx,$$

where $f : [0,+\infty) \to [0,+\infty)$ is quadratic near the origin and with finite limit at infinity. However, an easy convexity argument shows that energies of this kind Γ-converge to the zero functional. Various methods have been developed to bypass this convexity constraint, and most of them exploit De Giorgi's suggestion of using suitable nonlocal versions of the functionals above (see [13], [21], [17]). The approach we consider here is based on singular perturbations and consists in adding a "small" term depending on second derivatives: the idea is to control the oscillations of minimizing sequences by penalizing abrupt changes of the gradient. So we are led to consider energies of the form

$$\frac{1}{\varepsilon}\int_\Omega f(\sqrt{\varepsilon}|\nabla u|)\,dx + r(\varepsilon)\int_\Omega \|\nabla^2 u\|^2\,dx, \tag{1.1}$$

where $r(\varepsilon)$ is a function which vanishes as $\varepsilon \to 0^+$.

*Received by the editors September 18, 2001; accepted for publication (in revised form) February 21, 2003; published electronically October 2, 2003.
http://www.siam.org/journals/sima/35-3/39538.html
†Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (morini@asdf4.math.cmu.edu).

The first step in this direction was taken by Alicandro, Braides, and Gelli in [2]: they showed that the one-dimensional functionals

$$\frac{1}{\varepsilon}\int_0^1 f(\sqrt{\varepsilon}|u'|)\,dx + \varepsilon^3 \int_0^1 |u''|^2\,dx,$$

with $f(t) := \alpha t^2 \wedge \beta$, Γ-converge to a functional of the form

$$\alpha \int_0^1 |u'|^2\,dx + c(\beta)\sum_{S_u} \sqrt{u^+ - u^-};$$

later, Alicandro and Gelli treated the N-dimensional case (see [3]). We intend to extend the results above to general functionals of the form (1.1), where f is still quadratic near the origin but possibly unbounded. In fact we set the problem in a much more general framework and investigate the asymptotic behavior of the sequence

$$F_\varepsilon(u) := \int_\Omega f_\varepsilon(|\nabla u|)\,dx + (r(\varepsilon))^3 \int_\Omega \|\nabla^2 u\|^2\,dx, \tag{1.2}$$

where (f_ε) is any family of positive nondecreasing functions with a convex or convex–concave shape (i.e., there exists $x_\varepsilon > 0$ such that f_ε is convex in $[0, x_\varepsilon]$ and concave in $[x_\varepsilon, +\infty)$); let us remark that such a structural assumption is quite natural for this kind of problem (see, for example, [14], [15], [22]). In the main theorem of the paper (Theorem 3.2) we prove that the Γ-limits of (1.2) are related to the pointwise limits of $f_\varepsilon(t)$ and of $r(\varepsilon)f_\varepsilon(t/r(\varepsilon))$: if for an infinitesimal subsequence (ε_n) we have

(a) $f_{\varepsilon_n} \to g$ pointwise,

(b) $r(\varepsilon_n)f_{\varepsilon_n}(\cdot/r(\varepsilon_n)) \to b$ pointwise,

then (F_{ε_n}) Γ-converges to a functional F defined on $BV(\Omega)$ and of the form

$$F(u) = \int_\Omega f(|\nabla u|)\,dx + \int_{S_u} \varphi(u^+ - u^-) + C|D^c u|, \tag{1.3}$$

where C (possibly equal to $+\infty$), f, and φ are defined in terms of g and b.

The presence of the second derivatives in the approximating functionals determines a restriction on the regularity and on the growth of the jump-functions φ that can possibly appear in the limit. For instance, it turns out that φ always satisfies the growth condition

$$C_1(\sqrt{z} - 1) \le \varphi(z) \le C_2(z + 1) \qquad \forall z \ge 0$$

for suitable C_1, $C_2 > 0$; in particular, the Mumford–Shah functional is not reachable by our procedure. However, since for any positive, convex, and superlinear function g and for any positive and concave function b with $\lim_{t\to 0^+} b(t)/t = +\infty$, it is possible to construct a family (f_ε) and a rescaling function $r(\varepsilon)$ such that conditions (a) and (b) above are fulfilled, we see that a wide class of free-discontinuity functionals with φ satisfying (3.57) can be approximated.

In section 4 we apply our theorem to prove that if f is quadratic near the origin, sublinear, and concave at infinity, then there exists a rescaling function $r(\varepsilon)$ (explicitly given in terms of f) such that the family (1.1) Γ-converges, up to passing to a subsequence, to a free-discontinuity functional of the form (1.3). All the possible Γ-limits generated by a family of functionals as in (1.1) are classified. The rescaling $r(\varepsilon)$ is unique up to asymptotic equivalence.

In a recent paper [9] Bouchitté, Dubs, and Seppecher considered the one-dimensional functionals

$$F_\varepsilon(u) := \int_I \frac{|u'|^2}{1+(\varepsilon|u'|)^p}\,dx + \varepsilon^{\frac{3p}{p-1}\vee 4}\int_I |u''|^2\,dx$$

defined in $W^{2,2}(I)$ and proved that the Γ-limit is given by

$$F(u) := \int_I |u'|^2\,dx + k_p \sum_{x\in S_u} (u^+ - u^-)^{\frac{4-p}{2+p}\vee 0}.$$

When $p \le 2$, their result is a particular case of ours (but it is proved with different methods); on the contrary, the case $p > 2$ is not included in our treatment since the potential $f(t)$ becomes decreasing and degenerates at infinity; note that the use of a degenerate potential allows for the approximation of the Mumford–Shah functional (the case $p > 4$).

Let us also point out that our theorem can be used in particular to handle the so-called Perona–Malik functional

$$\frac{1}{\varepsilon}\int_\Omega \log(1+\varepsilon|\nabla u|^2)\,dx;$$

we will show that in this case the right rescaling function is given by $r(\varepsilon) = \frac{\varepsilon}{\log\frac{1}{\varepsilon}}$ and that the family

$$\frac{1}{\varepsilon}\int_\Omega \log(1+\varepsilon|\nabla u|^2)\,dx + \left(\frac{\varepsilon}{\log\frac{1}{\varepsilon}}\right)^3 \int_\Omega \|\nabla^2 u\|^2\,dx$$

Γ-converges to

$$\int_\Omega |\nabla u|^2\,dx + c\int_{S_u} \sqrt{u^+ - u^-}\,d\mathcal{H}^{N-1},$$

with $c > 0$ explicitly computable (see Example 4.8). The Perona–Malik functional was introduced in the context of image processing. Let us briefly recall the problem: if g is the input gray level function representing the original image, the simplest way to smooth and denoise it is by applying a Gaussian convolution kernel; this procedure turns out to be equivalent to considering as a processed image the solution $u(x,t)$ of the heat diffusion equation

$$\frac{\partial}{\partial t}u = \Delta u, \qquad u(x,0) = g(x), \tag{1.4}$$

computed at time t ("t" can be seen as a scale parameter: the greater t, the smaller the scale at which the smoothing occurs).

The main drawback of this approach is that it produces an unconditional smoothing which cannot distinguish between objects and contours, since edges also begin soon to diffuse! To overcome these difficulties Perona and Malik proposed in [26] a model of selective smoothing where the contours are preserved as much as possible: it consists in replacing (1.4) by the nonlinear equation

$$\frac{\partial}{\partial t}u = \operatorname{div}\left(\frac{\nabla u}{1+|\nabla u|^2}\right), \qquad u(x,0) = g(x), \tag{1.5}$$

which is the gradient flow of the (Perona–Malik) functional $\int_\Omega \log(1+|\nabla u|^2)\,dx$. The underlying idea is the following: where $|\nabla u|$ is large, in particular near the edges, the diffusion is low and the contour is "kept," while far from the edges, where the gradient is smaller, u diffuses as in the heat equation. Note that the simultaneous smoothing and edge-detection effects of the equation strongly depend on the particular structure of the function $\log(1+t^2)$: the quadratic behavior near the origin is responsible for the denoising process, while the concave behavior at infinity is responsible for the edge-detection. Our Γ-convergence result says that there is an alternative procedure, based on minimizing the (rescaled) energy instead of considering its gradient flow, which exploits the structure of $\log(1+t^2)$ to produce again a smoothing and edge-detection effect.

Actually the same considerations apply to all functions f satisfying our structure assumptions, and we can think of the functionals $\int f(|\nabla u|)\,dx$ as "generalized Perona–Malik energies" giving rise to "generalized Perona–Malik equations" of the form

$$\frac{\partial}{\partial t}u = \operatorname{div}\,(g(|\nabla u|)\nabla u)\,, \qquad u(x,0)=g(x),$$

with g bounded and decreasing to 0 when $|\nabla u|$ is large.

We want to mention, as a further application of our main result, the study of the asymptotic behavior of the family

$$\frac{1}{\varepsilon}\int_\Omega f(\varepsilon|\nabla u|)\,dx + \varepsilon^3\int_\Omega \|\nabla^2 u\|^2\,dx,$$

where f is nondecreasing, differentiable at the origin, with nonzero derivative, and concave at infinity: the Γ-limit turns out to be a functional defined in $BV(\Omega)$ and of the form

$$f'(0)\int_\Omega |\nabla u|\,dx + \int_{S_u}\varphi(u^+-u^-)\,d\mathcal{H}^{N-1} + f'(0)|D^c u|, \tag{1.6}$$

with φ explicitly characterized in terms of f. Again, as f varies among all the admissible potentials, a wide class of jump-functions φ can be generated (see Theorem 4.11 and Example 4.12).

Some final remarks are in order. All the convergence results we mentioned above are completely proved in the one-dimensional case; in N dimensions one can prove the following. Let (F_n) be a sequence of one-dimensional functionals converging to F and denote by (F_n^N) and F^N their respective N-dimensional versions; then we show that $\Gamma\text{-}\lim_n F_n^N(u) = F^N(u)$ if u satisfies

$$\exists\, u_k \to u \quad \text{such that (s.t.)} \quad \mathcal{H}^{N-1}(S_{u_k}) < +\infty \quad \text{and} \quad F^N(u_k)\to F^N(u).$$

The class of such functions coincides with the whole space when the bulk part of F^N is linear (as in (1.6)); we believe that the same occurs when F^N is defined in SBV, but at the moment such a technical result is not available, and, in fact, the complete representation of the Γ-limit holds for functions, with discontinuity set of finite $\mathcal{H}^{N-1}$-measure. Let us finally remark that these difficulties arise in the proof of the Γ-lim sup inequality, while the Γ-lim inf inequality is completely proved as well as the equicoerciveness of the approximating functionals, which guarantees the convergence of minimizers.

2. Preliminary results.

2.1. Definitions and general properties of BV functions. In this subsection we fix notation and briefly recall basic definitions and properties from the theory of BV functions: for a general treatment we refer to [6]. The Lebesgue measure and the $(N-1)$-dimensional Hausdorff measure of a set $B \subset \mathbb{R}^N$ are denoted by $\mathcal{L}^N(B)$ and $\mathcal{H}^{N-1}(B)$, respectively. We will often write $|B|$ instead of $\mathcal{L}^N(B)$. Given a measure μ we denote its total variation by $|\mu|$; moreover $\mu\lfloor B$ denotes the restriction of the measure μ to the set B given by $(\mu\lfloor B)(A) = \mu(B \cap A)$.

Let $\Omega \subset \mathbb{R}^N$ be an open set, let $u : \Omega \to \mathbb{R}$ be a measurable function, and let $x \in \Omega$. We denote by $u^+(x)$ and $u^-(x)$, respectively, the upper and lower limits of u at x, defined by

$$u^+(x) := \inf\left\{ t \in \mathbb{R} : \lim_{\rho \to 0^+} \frac{|\{y \in \Omega : |x-y| < \rho,\ u(y) > t\}|}{\rho^N} = 0 \right\},$$

$$u^-(x) := \sup\left\{ t \in \mathbb{R} : \lim_{\rho \to 0^+} \frac{|\{y \in \Omega : |x-y| < \rho,\ u(y) < t\}|}{\rho^N} = 0 \right\}.$$

If $u^+(x) = u^-(x) \in \mathbb{R}$, then the common value of $u^+(x)$ and $u^-(x)$ is called the *approximate limit* of u at the point x and is denoted by $\operatorname{ap-lim}_{y \to x} u(y)$.

We say that u is a *function of bounded variation* in Ω and we write $u \in BV(\Omega)$ if $u \in L^1(\Omega)$ and its distributional derivative is a vector-valued measure Du with finite total variation $|Du|(\Omega)$. Given $u \in BV(\Omega)$, we denote by J_u the set where $u^+ > u^-$ and by S_u the *essential discontinuity set* of u made up of those points x which are not Lebesgue points. It turns out that $J_u \subseteq S_u$ and $\mathcal{H}^{N-1}(S_u \setminus J_u) = 0$. For every $x \notin S_u$ we denote by $\tilde{u}(x)$ the approximate limit of u at x.

The *complete graph* of a function $u \in BV(\Omega)$ is the set

$$\Gamma_u := \{(x,z) \in \Omega \times \mathbb{R} : u^-(x) \le z \le u^+(x)\}.$$

If $u \in BV(\Omega)$, then it can be proved that S_u is countably $(\mathcal{H}^{N-1}, N-1)$ rectifiable, i.e.,

$$S_u = N \cup \bigcup_{i \in \mathbb{N}} K_i,$$

where $\mathcal{H}^{N-1}(N) = 0$, and each K_i is a compact set contained in a C^1 hypersurface; as a consequence we have that for $\mathcal{H}^{N-1}$-a.e. $x \in S_u$ it is possible to define an approximate tangent plane $T_x(S_u)$ and therefore an approximate normal unit vector $\nu_u(x)$ which can be chosen in such a way that

$$\lim_{\rho \to 0^+} \int_{B_\rho^{\nu_u(x)}(x)} |u(y) - u^+(x)|\, dy = 0,$$

where $B_\rho^{\nu_u(x)}(x) := \{y \in B_\rho(x) : (y-x)\cdot\nu_u(x) > 0\}$ (here and in what follows, given x and y in $\mathbb{R}^N$, we denote the scalar product of x and y by $x \cdot y$). For every $u \in BV(\Omega)$, by the Radon–Nikodým theorem we can write $Du = D^a u + D^s u$, where $D^a u$ is absolutely continuous and $D^s u$ is singular with respect to the Lebesgue measure. We denote the density of $D^a u$ with respect to the Lebesgue measure by ∇u. Moreover, we denote the restriction of $D^s u$ to S_u by $D^j u$, and the restriction of $D^s u$ to $\Omega \setminus S_u$ by $D^c u$. It turns out that $D^j u = (u^+ - u^-)\nu_u \mathcal{H}^{N-1}\lfloor S_u$, so that in particular

$$|D^s u| = |D^c u| + (u^+ - u^-)\mathcal{H}^{N-1}\lfloor S_u.$$

We will say that a set E is of *finite perimeter* in Ω if χ_E (i.e., the characteristic function of E) is of bounded variation in Ω. We define $\partial^* E \cap \Omega := S_{\chi_E} \cap \Omega$ to be the *reduced boundary* of E in Ω. Let us recall now the Fleming–Rishel *coarea formula*. Let u be a Lipschitz function and let v belong to $BV(\Omega)$. Then for almost every $t \in \mathbb{R}$ we have that $\{x \in \Omega : u > t\}$ is a set of finite perimeter in Ω and

$$\int_\Omega v|\nabla u|\,dx = \int_{-\infty}^{+\infty} \left(\int_{\partial^*\{u>t\}\cap\Omega} \tilde{v}\, d\mathcal{H}^{N-1} \right) dt. \tag{2.1}$$

We say that u is a *special function of bounded variation*, and we write $u \in SBV(\Omega)$ if $u \in BV(\Omega)$ and $D^c u = 0$. For each $p \geq 1$ the space of all functions $u \in SBV(\Omega)$ such that

$$\nabla u \in L^p(\Omega) \qquad \text{and} \qquad \mathcal{H}^{N-1}(S_u) < +\infty$$

is denoted by $SBV^p(\Omega)$. We consider also the larger space $GBV(\Omega)$, which is composed of all measurable functions $u : \Omega \to \mathbb{R}$ whose truncations $u_k = (u \wedge k) \vee (-k)$ belong to $BV(\Omega)$ for every $k > 0$; finally we set

$$GSBV := \{u \in GBV(\Omega) : |D^c u_k| = 0 \ \forall k > 0\} = \{u \in L^1(\Omega) : u_k \in SBV(\Omega) \ \forall k > 0\}$$

and

$$GSBV^p(\Omega) := \{u \in L^1(\Omega) : u_k \in SBV^p(\Omega) \ \forall k > 0\}.$$

Every $u \in GBV(\Omega) \cap L^1_{loc}(\Omega)$ has a countably $(\mathcal{H}^{N-1}, N-1)$ rectifiable discontinuity set S_u.

We conclude this subsection by recalling a "slicing" result due to Ambrosio (see [5]) and an L^1-precompactness criterion by slicing proved in [1]. First we introduce some notation. Let $\xi \in S^{N-1}$ and let $\Pi_\xi := \{y \in \mathbb{R}^N : y \cdot \xi = 0\}$ be the linear hyperplane orthogonal to ξ. Given $E \subset \mathbb{R}^N$ we denote by $E_\xi \subseteq \Pi_\xi$ the orthogonal projection of E on Π_ξ, and for $y \in \Pi_\xi$ we set $E^y_\xi := \{t \in \mathbb{R} : y + t\xi \in E\}$. Finally for $u : E \to \mathbb{R}$ we define $u^y_\xi : E^y_\xi \to \mathbb{R}$ by $u^y_\xi(t) := u(y + t\xi)$.

THEOREM 2.1.

(a) *Let $u \in BV(\Omega)$. Then for all $\xi \in S^{N-1}$ the function u^y_ξ belongs to $BV(\Omega^y_\xi)$ for $\mathcal{H}^{N-1}$-a.e. $y \in \Pi_\xi$. For such y we have*

$$\begin{aligned} (u^y_\xi)'(t) &= \nabla(y + t\xi) \cdot \xi \qquad \text{for a.e. } t \in \Omega^y_\xi, \\ S_{u^y_\xi} &= (S_u)^y_\xi, \\ u^y_\xi(t\pm) &= u^\pm(y + t\xi) \qquad \text{or} \qquad u^y_\xi(t\pm) = u^\mp(y + t\xi), \end{aligned}$$

according to the case $\nu_u \cdot \xi > 0$ or $\nu_u \cdot \xi < 0$ (the case $\nu_u \cdot \xi = 0$ being negligible). Moreover we have

$$\int_{\Pi_\xi} |D^c u^y_\xi|(A^y_\xi)\, d\mathcal{H}^{N-1}(y) = |D^c u \cdot \xi|(A)$$

for all open subsets $A \subseteq \Omega$, and for all Borel functions g

$$\int_{\Pi_\xi} \sum_{t \in S_{u^y_\xi}} g(t)\, d\mathcal{H}^{N-1}(y) = \int_{S_u} g(x)|\nu_u \cdot \xi|\, d\mathcal{H}^{N-1}.$$

(b) *Conversely, if* $u \in L^1(\Omega)$ *and for all* $\xi \in \{e_1, \dots, e_N\}$ *and for a.e.* $y \in \Pi_\xi$ $u_\xi^y \in BV(\Omega_\xi^y)$ *(*$SBV(\Omega_\xi^y)$*) and*

$$\int_{\Pi_\xi} |Du_\xi^y|\, d\mathcal{H}^{N-1}(y) < +\infty,$$

then $u \in BV(\Omega)$ *(*$SBV(\Omega)$*).*

Given a family $\mathcal{F}$ of functions, for every $\xi \in S^{N-1}$ and $y \in \Pi_\xi$ we set $\mathcal{F}_\xi^y := \{u_\xi^y : u \in \mathcal{F}\}$; moreover we say that a family $\mathcal{F}'$ is δ-close to $\mathcal{F}$ if $\mathcal{F}'$ is contained in a δ-neighborhood of $\mathcal{F}$.

LEMMA 2.2. *Let* $\mathcal{F}$ *be a family of equi-integrable functions belonging to* $L^1(A)$ *and assume that there exists a basis of unit vectors* $\{\xi_1, \dots, \xi_N\}$ *with the property that for every* $i = 1, \dots, N$*, for every* $\delta > 0$*, there exists a family* $\mathcal{F}_\delta$ δ*-close to* $\mathcal{F}$ *such that* $(\mathcal{F}_\delta)_{\xi_i}^y$ *is precompact in* $L^1(A_{\xi_i}^y)$ *for* $\mathcal{H}^{N-1}$*-a.e* $y \in A_{\xi_i}$*. Then* $\mathcal{F}$ *is precompact in* $L^1(A)$*.*

2.2. Semicontinuity and relaxation in BV and SBV. Let $f : \mathbb{R} \to [0, +\infty]$ be convex. Then we define the *recession function* f^∞ of f by

$$f^\infty(z) = \lim_{t \to +\infty} \frac{f(tz)}{t}.$$

Let $\theta : \mathbb{R} \to [0, +\infty]$ be lower semicontinuous and such that there exists $\lim_{t \to 0^+} \theta(t)/t$. Then we can define the *recession function* θ^0 of θ by

$$\theta^0(z) = \lim_{t \to 0^+} \frac{\theta(tz)}{t}.$$

The functions f^∞ and θ^0 turn out to be 1-homogeneous. For every $g, h : \mathbb{R} \to [0, +\infty]$, we define the *inf-convolution of* g *and* h as the function $g \square h$ given by

$$(g \square h)(z) = \inf\{g(x) + h(z - x) : x \in \mathbb{R}\}.$$

Finally we recall that given a function $F : X \to \mathbb{R} \cup +\infty$, where X is a topological space, we denote by $\overline{F}$ the *relaxed functional* of F, i.e., the greatest lower semicontinuous (with respect to the X-topology) functional which is less than F.

The following relaxation result is proved in [8].

THEOREM 2.3 (relaxation in BV). *Let* $f : [0, +\infty) \to [0, +\infty)$ *be a nondecreasing convex function and let* $\varphi : [0, +\infty) \to [0, +\infty)$ *be a concave function. Let* $F : BV(\Omega) \to [0, +\infty]$ *be defined by*

(2.2)
$$F(u) := \begin{cases} \displaystyle\int_\Omega f(|\nabla u|)\, dx + \int_{S_u} \varphi(u^+ - u^-)\, d\mathcal{H}^{N-1} & \text{if } u \in SBV^2(\Omega) \cap L^\infty(\Omega), \\ +\infty & \text{otherwise.} \end{cases}$$

Then the relaxed functional of F *with respect to the* L^1*-metric is given on* BV *by*

$$\overline{F}(u) := \int_\Omega f_1(|\nabla u|)\, dx + \int_{S_u} \varphi_1(u^+ - u^-)\, d\mathcal{H}^{N-1} + \big(f^\infty(1) \wedge \varphi^0(1)\big)\, |D^c u|,$$

where $f_1 := f \square \varphi^0$ *and* $\varphi_1 := \varphi \square f^\infty$*.*

It is possible to prove that $f\triangle\varphi^0 = [f \wedge (\varphi^0 + f(0))]^{**}$, where h^{**} denotes the convexification of h, i.e., the greatest convex and lower semicontinuous function which is smaller than h and, analogously, $\varphi\triangle f^\infty = \operatorname{sub}[\varphi \wedge (f^\infty + \varphi(0))]$, where $\operatorname{sub} h$ denotes the subadditive envelope, i.e., the greatest lower semicontinuous and subadditive function which is smaller than h. Given two Borel functions $\varphi :]0, +\infty[\to [0, +\infty)$ and $f : [0, +\infty] \to [0, +\infty)$, we consider the functional F defined by

$$F(u) = \begin{cases} \displaystyle\int_\Omega f(|\nabla u|)\,dx + \int_{S_u} \varphi(u^+ - u^-)\,d\mathcal{H}^{N-1} & \text{if } u \in GSBV(\Omega), \\ +\infty & \text{otherwise.} \end{cases} \tag{2.3}$$

In [5] the following semicontinuity result is proved.

THEOREM 2.4 (Ambrosio's semicontinuity theorem). *Let $\Omega \subset \mathbb{R}^N$ be an open bounded set. Let $f : [0, +\infty) \to [0, +\infty)$ be a nondecreasing convex function such that $f^\infty(1) = +\infty$ and let $\varphi :]0, +\infty] \to [0, +\infty)$ be a nondecreasing subadditive function such that $b^0(1) = \infty$. Then the functional F defined in* (2.3) *is lower semicontinuous with respect to the L^1-convergence.* □

2.3. A density result in SBV. In analogy with the strong density results of smooth functions in $W^{1,p}(\Omega)$, functions in $SBV^p(\Omega)$ can be approximated in a "strong sense" by functions which have a "regular" jump set and are smooth outside. This can be formally expressed as follows.

Let Ω be an open bounded subset in $\mathbb{R}^N$ with Lipschitz boundary, and denote by $\mathcal{W}(\Omega)$ the space of all functions $w \in SBV(\Omega)$ enjoying the following properties:

(i) $\mathcal{H}^{N-1}(\overline{S}_w \setminus S_w) = 0$;
(ii) $\overline{S}_w$ is the intersection of Ω with the union of a finite number of pairwise disjoint $(N-1)$-simplexes;
(iii) $w \in W^{k,\infty}(\Omega \setminus \overline{S}_w)$ for every $k \in \mathbb{N}$.

Cortesani and Toader proved in [18] the following density result.

THEOREM 2.5. *Let $u \in SBV^p(\Omega) \cap L^\infty(\Omega)$. Then there exists a sequence $(w_j)_j$ in $\mathcal{W}(\Omega)$ such that $w_j \to u$ strongly in $L^1(\Omega)$, $\nabla w_j \to \nabla u$ strongly in $L^p(\Omega, \mathbb{R}^N)$, $\lim_j \|w_j\|_\infty = \|u\|_\infty$, and*

$$\limsup_{j\to\infty} \int_{S_{w_j}} \phi(w_j^+, w_j^-, \nu_{w_j})\,d\mathcal{H}^{N-1} \le \int_{S_u} \phi(u^+, u^-, \nu_u)\,d\mathcal{H}^{N-1}$$

for every upper semicontinuous function $\phi : \mathbb{R} \times \mathbb{R} \times S^{N-1} \to [0, +\infty)$ such that $\phi(a, b, \nu) = \phi(b, a, -\nu)$ for every a, $b \in \mathbb{R}$ and for every $\nu \in S^{N-1}$.

2.4. Γ-convergence. We recall here the definition and the main properties of Γ-convergence: for the general theory we refer to [19] (see also the forthcoming [11]).

DEFINITION 2.6. *Let (X, d) be a metric space and let $F_h : X \to \mathbb{R} \cup \{+\infty\}$ be a sequence of functions. We set*

$$\Gamma\text{-}\liminf_{h\to\infty} F_h(x) := \inf\left\{\liminf_{h\to\infty} F_h(x_h) :\ x_h \to x\right\}$$

and

$$\Gamma\text{-}\limsup_{h\to\infty} F_h(x) := \inf\left\{\limsup_{h\to\infty} F_h(x_h) :\ x_h \to x\right\}.$$

We say that the sequence $(F_h)_{h\in\mathbb{N}}$ Γ*-converges if*

$$\Gamma\text{-}\liminf_{h\to\infty} F_h(x) = \Gamma\text{-}\limsup_{h\to\infty} F_h(x) \qquad \forall x \in X.$$

The common value is called the Γ-limit *and is denoted by* $\Gamma\text{-}\lim_{h\to\infty} F_h$.

DEFINITION 2.7. *We say that the maps* $F_h : X \to \mathbb{R} \cup \{+\infty\}$ *are equicoercive if for every* $t \in \mathbb{R}$ *there exists a compact subset* $K_t \subseteq X$ *such that*

$$\{x \in X : F_h(x) \le t\} \subseteq K_t \qquad \forall h \in \mathbb{N}.$$

The following theorem explains the variational meaning of this kind of convergence.

THEOREM 2.8. *Let* $(F_h)_h$ *be a sequence of equicoercive maps which* Γ*-converges to* F. *Then if* $(x_h)_h$ *is a sequence such that*

$$\lim_{h\to\infty} F_h(x_h) = \lim_{h\to\infty} \inf_X F_h,$$

x_h *is precompact and any cluster point is a minimizer of* F.

We finally recall that given $F : X \to \mathbb{R} \cup \{+\infty\}$, the relaxed functional $\overline{F}$ can be characterized as the Γ-limit of the constant sequence $F_n = F$ for every $n \in \mathbb{N}$.

3. The main convergence result in the one-dimensional case. Let $f_n : [0,+\infty) \to [0,+\infty)$ be a family of continuous nondecreasing functions and let r_n be an infinitesimal sequence of positive real numbers. For any open bounded subset $I \subset \mathbb{R}$ we define

$$F_n(u) := \begin{cases} \displaystyle\int_I f_n(|u'|)\,dx + (r_n)^3 \int_I |u''|^2\,dx & \text{if } u \in W^{2,2}(I), \\ +\infty & \text{otherwise in } L^1(I). \end{cases} \tag{3.1}$$

Moreover, given two functions b, $g : [0,+\infty) \to [0,+\infty)$, we set

$$\mathcal{F}_{b,g}(u) := \begin{cases} \displaystyle\int_I g(|u'|)\,dx + \sum_{S_u} \varphi(u^+ - u^-) & \text{if } u \in SBV(I), \\ +\infty & \text{otherwise in } L^1(I), \end{cases} \tag{3.2}$$

where

$$\varphi(z) := \inf_{\eta>0} \inf \left\{ \int_0^\eta b(|u'|)\,dx + \int_0^\eta |u''|^2\,dx : u \in W^{2,2}(0,\eta), \right. \\ \left. u(0) = 0,\, u(\eta) = z, u'(0) = u'(\eta) = 0 \right\}. \tag{3.3}$$

We denote by $\overline{\mathcal{F}_{b,g}}$ the L^1-relaxation of $\mathcal{F}_{b,g}$. If g is convex and b is convex, concave, or convex–concave, then we can finally define

$$F_{b,g}(u) := \begin{cases} \displaystyle\int_I g_1(|u'|)\,dx + \sum_{S_u} \varphi_1(u^+ - u^-) + (g^\infty(1) \wedge b^0(1))|D^c u| & \text{if } u \in BV(I), \\ +\infty & \text{otherwise in } L^1(I), \end{cases} \tag{3.4}$$

where $g_1 := g \triangle b^0 = [g \wedge (b^0 + g(0))]^{**}$ and $\varphi_1 := \varphi \triangle g^\infty = \mathrm{sub}\,(\varphi \wedge g^\infty)$ (g^∞ and b^0 are the recession functions of g and b, respectively, defined in subsection 2.2).

Remark 3.1. Note that if $g^\infty(1) = b^0(1) = +\infty$, then $F_{b,g} = \mathcal{F}_{b,g}(u)$.

Our main result is stated in the following theorem.

THEOREM 3.2. *Let f_n and r_n be as above, and suppose in addition that the following hold:*

(i) *there exists a nondecreasing function $g : [0, +\infty) \to [0, +\infty)$ such that*

$$f_n(t) \to g(t) \qquad \forall t \in [0, +\infty); \tag{3.5}$$

(ii) *there exists a nondecreasing and continuous function $b : (0, +\infty) \to (0, +\infty)$ such that*

$$r_n f_n\left(\frac{t}{r_n}\right) \to b(t) \qquad \forall t > 0. \tag{3.6}$$

Then

$$\Gamma\text{-}\limsup_{n\to\infty} F_n \le \overline{\mathcal{F}_{b,g}}(u),$$

with respect to the $L^1(I)$-convergence. If in addition we assume

(iii) *one of the two following structural conditions holds true:*

(st1) *f_n is convex for every $n \in \mathbb{N}$;*

(st2) *there exists a sequence $(x_n) \subset (0, +\infty)$ such that $x_n \to +\infty$ and f_n is convex in $[0, x_n]$ and concave in $[x_n, +\infty)$;*

then

$$\Gamma\text{-}\lim_{n\to\infty} F_n = \overline{\mathcal{F}_{b,g}} = F_{b,g}.$$

Finally, every sequence u_n such that $\sup_n (F_n(u_n) + \|u_n\|_1) < +\infty$ is strongly precompact in L^p for every $p \ge 1$.

Remark 3.3. If (iii) holds, then g is convex; assumption (st1) implies that b is in turn convex, while (st2) implies that b is either concave or convex–concave. In all these cases the recession function b^0 is well defined. We finally point out that the equality $\overline{\mathcal{F}_{g,b}} = F_{b,g}$ stated in the last part of the theorem is a consequence of Theorem 2.3 and of the equality $\varphi^0 = b^0$, which will be proved in what follows (see Lemma 3.6).

Remark 3.4. If (st2) holds with $\limsup_{n\to\infty} x_n r_n = c > 0$, then

$$b(t) \ge g^\infty(1)t = g^\infty(t) \qquad \forall t \in [0, c], \tag{3.7}$$

so that, in particular, $b^0(1) \ge g^\infty(1)$. To see this we can suppose that $\lim_n x_n r_n = c$; since the functions f_n converge to g and become convex in larger and larger intervals, we have

$$g'(t-) \le \liminf_{n\to\infty} f_n'(t-) \le \limsup_{n\to\infty} f_n'(t+) \le g'(t+) \tag{3.8}$$

for every $t > 0$. Suppose that $g^\infty(1) \ne 0$ (otherwise the statement is trivial) and let y_k be a divergent sequence such that $g'(y_k) \to g^\infty(1)$. For a fixed k and $\delta \in (0, 1)$ there exists $n_{k,\delta}$ such that $x_n \ge y_k$, $f_n(t) \ge g(y_k)/2$, and $f_n'(t) \ge (1 - \delta)g'(y_k)$ for every $n \ge n_{k,\delta}$, so that, by convexity,

$$f_n(t) \ge \frac{g(y_k)}{2} + (1 - \delta)g'(y_k)(t - y_k) \qquad \forall t \in [y_k, x_n]\ \forall n \ge n_{k,\delta}. \tag{3.9}$$

Fix $t < c$; then, by (3.9),

$$r_n f_n\left(\frac{t}{r_n}\right) \geq r_n \frac{g(y_k)}{2} + r_n(1-\delta)g'(y_k)\left(\frac{t}{r_n} - y_k\right)$$

for every $n \geq \overline{n}$, where $\overline{n} \geq n_{k,\delta}$ is such that $y_k \leq t/r_n \leq x_n$ for every $n \geq \overline{n}$. Passing to the limit in n in the above inequality and taking into account (3.6), we obtain $b(t) \geq (1-\delta)g'(y_k)t$, from which (3.7) follows letting $k \to \infty$ and $\delta \to 0$. With the same proof we see that (st1) implies that $b(t) \geq g^\infty(t)$ for every $t \geq 0$.

Before giving the proof of the theorem we need to state and prove some preparatory lemmas.

LEMMA 3.5. *Suppose that $b(t) = Mt$ for some $M > 0$ and let φ be the function defined in* (3.3). *Then $\varphi(z) = Mz$ for every $z > 0$.*

Proof. Fix $z > 0$ and let (v, η) be an admissible pair for problem (3.3). Then

$$\int_0^\eta M|v'|\,dt + \int_0^\eta |v''|^2\,dt \geq \int_0^\eta M|v'|\,dt \geq Mz,$$

and therefore $\varphi(z) \geq Mz$. Let us prove now the reverse inequality. To this end we construct a sequence of admissible pairs (v_n, η_n) by setting $\eta_n := nz$ and

$$v_n(t) := \begin{cases} \frac{\phi(t)}{n} & \text{if } t \in [0,1), \\ \frac{1}{n} + \frac{1}{n}(t-1) & \text{if } t \in [1, nz-1), \\ z - \frac{\phi(nz-t)}{n} & \text{if } t \in [nz-1, nz], \end{cases}$$

where ϕ is a function belonging to $C^2([0,1])$ and satisfying $\phi(0) = \phi'(0) = 0$, $\phi(1) = \phi'(1) = 1$. We can now estimate

$$\begin{aligned}\varphi(z) &\leq \int_0^{\eta_n} M|v_n'|\,dt + \int_0^{\eta_n} |v_n''|^2\,dt = 2\frac{M}{n}\int_0^1 |\phi'|\,dt + M\frac{nz-2}{n} + 2\frac{1}{n}\int_0^1 |\phi''|^2\,dt \\ &= Mz + O\left(\frac{1}{n}\right),\end{aligned}$$

and therefore, letting $n \to \infty$, we obtain $\varphi(z) \leq Mz$. ∎

LEMMA 3.6. *Let b be as in Remark* 3.3. *Then the function $\varphi : [0, +\infty) \to [0, +\infty)$ defined in* (3.3) *is continuous, nondecreasing, and subadditive, and $\varphi^0(1) = b^0(1)$.*

Proof. The first three properties are easy; let us prove only the last one. We begin with the case

$$b^0(1) = +\infty. \tag{3.10}$$

Claim. For every $\varepsilon > 0$ there exists $\delta > 0$ with the following property: if $z < \delta$ and if (η, u) is an *admissible pair* for problem (3.3) satisfying

$$\int_0^\eta b(|u'|)\,dx + \int_0^\eta |u''|^2\,dx < (1+\varepsilon)\varphi(z), \tag{3.11}$$

then $|u'| \leq \varepsilon$ in $(0, \eta)$.

Suppose by contradiction the existence of $\varepsilon > 0$ and of a sequence $\delta_n \downarrow 0$ such that, for every $n \in \mathbb{N}$, there exist $z_n < \delta_n$ and (η_n, u_n), which satisfies

$$\int_0^{\eta_n} b(|u_n'|)\,dx + \int_0^{\eta_n} |u_n''|^2\,dx < (1+\varepsilon)\varphi(z_n) \tag{3.12}$$

and

$$\|u_n'\|_{L^\infty(0,\eta_n)} > \varepsilon. \tag{3.13}$$

Note that we can suppose $\eta_n > 1$ for every n (u_n can be extended outside the original interval as the constant function z_n); using Hölder's inequality we can estimate for every $x, y \in (0, \eta_n)$

$$|u_n'(x) - u_n'(y)| \le \int_x^y |u_n''|\,dt \le \sqrt{|x-y|}\left(\int_0^{\eta_n} |u_n''|^2\,dt\right)^{\frac{1}{2}} \le C\sqrt{|x-y|},$$

where $C > 0$ is independent of n; by the above estimate and by (3.13) we can deduce the existence of an interval $I_n \subseteq (0, \eta_n)$ such that $|I_n| \ge C'$, with C' independent of n, and $|u_n'| \ge \varepsilon/2$ in I_n. As a consequence we deduce

$$\int_0^\eta b(|u_n'|)\,dx + \int_0^\eta |u_n''|^2\,dx \ge \int_{I_n} b(|u_n'|)\,dx \ge b\left(\frac{\varepsilon}{2}\right) C',$$

which is in contradiction to (3.12) since $\varphi(z_n)$, by continuity, tends to 0. The claim is proved. Given $M > 0$, thanks to (3.10) we can choose ε such that $b(t)/t \ge M$ for every $t \in (0, \varepsilon]$; if $\delta > 0$ is as in the claim above, then for $0 \le z < \delta$ we can estimate

$$\begin{aligned}(1+\varepsilon)\varphi(z) &\ge \int_0^\eta b(|u'|)\,dx + \int_0^\eta |u''|^2\,dx \\ &\ge \int_0^\eta \frac{b(|u'|)}{|u'|}|u'|\,dx \ge M\int_0^\eta |u'|\,dx \ge Mz,\end{aligned}$$

where (η, u) is an admissible pair satisfying (3.11); this concludes the proof when (3.10) holds. Let us suppose now that

$$b^0(1) = C < +\infty. \tag{3.14}$$

Fix $\sigma > 0$ and choose $\varepsilon_\sigma > 0$ such that $b(t) < (C+\sigma)t$ for any $t \in (0, \varepsilon_\sigma)$. Consider the sequence of admissible pairs (η_n, v_n) constructed in the previous lemma; for n large enough we have $\|v_n'\|_\infty \le \varepsilon_\sigma$ and therefore

$$\begin{aligned}\varphi(z) &\le \int_0^{\eta_n} b(|v_n'|)\,dt + \int_0^{\eta_n} |v_n''|^2\,dt \le (C+\sigma)\int_0^{\eta_n} |v_n'|\,dt + \int_0^{\eta_n} |v_n''|^2\,dt \\ &= (C+\sigma)z + O\left(\frac{1}{n}\right).\end{aligned}$$

Letting $n \to \infty$ and $\sigma \to 0$ we obtain

$$\varphi(z) \le Cz \qquad \forall z > 0. \tag{3.15}$$

Finally, arguing exactly as for the other case, we easily obtain $\liminf_{z\to 0^+} \varphi(z)/z \ge C$, which concludes the proof of the lemma. □

LEMMA 3.7. *Let $(u_n)_{n\in\mathbb{N}}$ be a sequence of functions such that $\sup_n F_n(u_n) < +\infty$ and consider the sets $D_n := \{x \in I : |u_n'(x)| > c/r_n\}$, where $c > 0$ is a fixed constant. Then there exists $\overline{n} \in \mathbb{N}$, depending on c, such that*

$$|D_n| \le \left(\frac{2\sup_n F_n(u_n)}{b(c)}\right) r_n$$

for every integer $n \ge \overline{n}$.

Proof. If n is large enough, thanks to (3.6) we have

$$F_n(u_n) \geq \int_{D_n} f_n(|u_n'|)\,dx \geq \frac{r_n}{r_n} f_n\left(\frac{c}{r_n}\right)|D_n| \geq \frac{1}{2r_n} b(c)|D_n|. \qquad \square$$

LEMMA 3.8. *Suppose also that* (iii) *of Theorem* 3.2 *holds true and let* (u_n) *be such that*

$$\sup_n F_n(u_n) < +\infty.$$

Then

$$r_n \|u_n'\|_\infty \leq 2 \sup_n F_n(u_n) + 1 \tag{3.16}$$

for n *large enough. Moreover, if* $g \not\equiv 0$*, there exists a positive constant* C *depending only on* $|I|$*,* g*, and* b *such that*

$$\operatorname{Var} u_n \leq C\left(\sup_n F_n(u_n) + 1\right)^2 \tag{3.17}$$

for n *large enough.*

Proof. Take $c = 1$ and consider the sets D_n defined in the previous lemma; since they are open, we can write $D_n = \bigcup_{k=1}^{\infty}(a_n^k, b_n^k)$. Let y be a point of D_n; therefore there exists $k \in \mathbb{N}$ such that $y \in (a_n^k, b_n^k)$. By Lemma (3.7) and using Hölder's inequality, we have

$$\begin{aligned}
|u_n'(y)| &\leq |u_n'(a_n^k)| + \int_{a_n^k}^{y} |u_n''(t)|\,dt \\
&\leq \frac{1}{r_n} + \frac{|D_n|^{\frac{1}{2}}}{(r_n)^{\frac{3}{2}}}\left((r_n)^3 \int_{a_n^k}^{b_n^k} |u_n''|^2\,dt\right)^{\frac{1}{2}} \\
&\leq \frac{1}{r_n} + 2\frac{(r_n)^{\frac{1}{2}}}{(r_n)^{\frac{3}{2}}} \sup_n F_n(u_n) = \left(2\sup_n F_n(u_n) + 1\right)\frac{1}{r_n}
\end{aligned}$$

so that (3.16) is proved. Concerning the second part of the lemma, we first observe that by a translation argument we can suppose that $f_n(0) = g(0) = 0$ for every $n \in \mathbb{N}$. Let x_0 be the last point such that $g(x_0) = 0$ and define

$$\tilde{g}(x) := \begin{cases} 0 & \text{if } x \in [0, x_0], \\ g(x - x_0) & \text{if } x \geq x_0; \end{cases}$$

it is easy to see that $\tilde{g}$ is still convex and $\tilde{g}^\infty = g^\infty$. Moreover, taking into account the fact that $f_n \to g$ uniformly on compact subsets of $[0, +\infty)$ (the uniformity follows from the pointwise convergence and from the monotonicity of f_n), we have

$$\forall \delta \in (0,1),\ \forall K > 0,\ \exists\, \overline{n} \quad \text{s.t.} \quad f_n \geq (1-\delta)\tilde{g} \text{ in } [0, K] \qquad \forall n \geq \overline{n}. \tag{3.18}$$

Fix $\overline{y} > 0$ such that $\tilde{g}'(\overline{y}) > g^\infty(1)/2$; set $k := 2\sup_n F_n(u_n) + 1$ and let $\overline{x}$ be the first point such that $\tilde{g}'(\overline{x}+)/2 \geq \min\{g^\infty(1)/3, b(1)/(3k)\}$. Since either $\overline{x} = 0$ or

$$\frac{\tilde{g}'(\overline{x}-)}{2} \leq \min\left\{\frac{g^\infty(1)}{3}, \frac{b(1)}{3k}\right\} \leq \frac{\tilde{g}'(\overline{x}+)}{2}, \tag{3.19}$$

it is clear that $\overline{x} < \overline{y}$. So, by virtue of (3.18), (3.19), (3.8), and (3.6), we can find $\overline{n}$ such that

(a) $f_n \geq \tilde{g}(x)/2$ in $[0, \overline{x}+1]$,
(b) $f_n'((\overline{x}+1)+) \geq \min\{g^\infty(1)/3, b(1)/(3k)\}$,
(c) $f_n(k/r_n)/(k/r_n) \geq b(k)/(3k)$

for every $n \geq \overline{n}$; we define $a(t) := \tilde{g}(\overline{x})/2 + \min\{g^\infty(1)/3, b(1)/(3k)\}(t-\overline{x})$. From the convexity of $\tilde{g}$ and by (3.19) and (a), we observe that

$$a(t) \leq \frac{\tilde{g}(\overline{x})}{2} \leq f_n(t) \qquad \text{in } [0, \overline{x}+1]. \tag{3.20}$$

If (st1) holds, then, taking into account (b) and (3.20), we also have

$$a(t) \leq f_n(\overline{x}+1) + f_n'\big((\overline{x}+1)+\big)(t-\overline{x}-1) \leq f_n(t) \qquad \forall t \geq \overline{x}+1.$$

Suppose now that (st2) holds; by replacing f_n with

$$\tilde{f}_n(t) := \begin{cases} f_n(t) + (r_n t)^2 & \text{if } t \leq x_n, \\ f_n(t) + (r_n x_n)^2 & \text{if } t > x_n, \end{cases}$$

if needed, we can assume that f_n is strictly convex in $[0, x_n]$ (recall that x_n is the point appearing in condition (st2)). Arguing as above, we obtain

$$a(t) \leq f_n(t) \qquad \forall t \in [\overline{x}+1, x_n \wedge k/r_n]. \tag{3.21}$$

We denote by y_n the first strictly positive point such that $f_n(y_n) = [f_n(k/r_n)/(k/r_n)]y_n$; by the strict convexity assumption we have that $0 < y_n \leq k/r_n$. If $x_n < y_n < k/r_n$, we can first observe that by concavity

$$f_n'(t\pm) \geq f_n'(y_n-) \geq \frac{r_n}{k} f_n\left(\frac{k}{r_n}\right) \qquad \forall t \in (x_n, y_n), \tag{3.22}$$

where the last inequality is a consequence of the following one:

$$f_n(t) \geq f_n(y_n) + \frac{r_n}{k} f_n\left(\frac{k}{r_n}\right)(t-y_n) = \frac{r_n}{k} f_n\left(\frac{k}{r_n}\right) t \qquad \text{in } [y_n, k/r_n]$$

(again the concavity of f_n in (x_n, y_n) is taken into account). Using (3.22) and (c) we then have

$$a(t) \leq f_n(x_n) + \min\left\{\frac{g^\infty(1)}{3}, \frac{b(1)}{3k}\right\}(t-x_n) \leq f_n(t) \qquad \text{in } [x_n, y_n] \tag{3.23}$$

and therefore

$$a(t) \leq f_n(y_n) + \frac{r_n}{k} f_n\left(\frac{k}{r_n}\right)(t-y_n) \leq f_n(t) \qquad \text{in } [y_n, k/r_n].$$

If $x_n < y_n = k/r_n$, then either

$$f_n(t) > \frac{r_n}{k} f_n\left(\frac{k}{r_n}\right) t \geq \frac{b(k)}{3k} t \qquad \text{in } [0, k/r_n]$$

or

$$f_n(t) < \frac{r_n}{k} f_n\left(\frac{k}{r_n}\right) t.$$

In the first case (3.17) follows immediately; in the second case we observe that (3.22) and therefore (3.23) are still true. Summarizing, we have proved that if $x_n < y_n$, then

$$a(t) \le f_n(t) \qquad \text{in } [0, k/r_n]; \tag{3.24}$$

arguing in a similar way we obtain the same estimate also if $y_n \le x_n$. Using the definition of $a(t)$ and the fact that $\overline{x} < \overline{y}$, from (3.24) we easily obtain

$$\min\left\{\frac{g^\infty(1)}{3}, \frac{b(1)}{3}\right\} t \le k f_n(t) + \frac{k g^\infty(1)}{3}\overline{y} \qquad \forall t \in [0, k/r_n],$$

from which, recalling the definition of k and (3.16), inequality (3.17) follows with

$$C := (6 + 2g^\infty(1)\overline{y}|I|)(\min\{g^\infty(1)/3, b(1)/3\})^{-1}. \qquad \square$$

Remark 3.9. Let us remark that if $u_n \to u$ in L^1 and $\sup_n F_n(u_n) < +\infty$, then $u_n \to u$ in L^p for every $p \ge 1$: indeed from (3.17) it easily follows that u_n is equibounded in L^∞. As a consequence we have that in one dimension the functionals F_n Γ-converge with respect to the L^1-norm if and only if they Γ-converge with respect to the L^p-norm for every $p \ge 1$.

LEMMA 3.10. *Assume also that condition* (iii) *of Theorem* 3.2 *holds and let* $(u_n)_{n\in\mathbb{N}} \subset SBV(I)$ *be such that* $r_n\|u_n'\|_\infty \to 0$ *as* $n \to \infty$. *Then there exists an increasing sequence* $(\psi_i)_{i\in\mathbb{N}}$ *of positive convex functions with the following properties:*

(i) $\psi_i(t) \uparrow g_1(t)$ *for every* $t > 0$ *as* $i \to \infty$ *(we recall that* g_1 *is the function appearing in* (3.4)*);*

(ii) $\psi_i^\infty(1) = g_1^\infty(1) = b^0(1) \wedge g^\infty(1)$ *for every* i*;*

(iii) *passing to a subsequence, still denoted by* $(u_n)_n$*, we have that for every* i *there exists* n_i *such that*

$$f_n(|u_n'|) \ge \psi_i(|u_n'|)$$

for every $n \ge n_i$.

Proof. We can assume that $\min\{g^\infty(1), b^0(1)\} \ne 0$; otherwise the statement is trivial. We will distinguish two cases.

Case 1. $g^\infty(1) > b^0(1)$.

In this case, by Remark 3.4, we have that (st2) holds true with $\lim_n x_n r_n = 0$. We can suppose without loss of generality that $f_n(0) = g(0) = 0$ for every $n \in \mathbb{N}$. We begin by assuming

$$g'(0+) < b^0(1); \tag{3.25}$$

let $\overline{x}$ be the last point such that $g'(\overline{x}-) \le b^0(1) \le g'(\overline{x}+)$ and set $\overline{y} := \sup\{y \ge 0 : g(t) \le b^0(1)t \ \forall t \in [0, y]\}$; then $\overline{x} < \overline{y}$. We also make the following assumption:

(3.26)
$$\forall \delta \in (0,1),\ \forall K > 0,\ \exists\, n_{\delta,K} \quad \text{s.t.} \quad f_n(t) \ge (1-\delta)g(t) \quad \forall t \in [0, K] \text{ and } \forall n \ge n_{\delta,K}.$$

It is clear that we can find $\delta_0 \in (0,1)$ such that for every $0 < \delta \le \delta_0$ we have $(1-\delta)g^\infty(1) > b^0(1)$ and $\overline{x}_\delta < \overline{y}$, where $\overline{x}_\delta$ is the last point satisfying $(1-\delta)g'(\overline{x}_\delta-) \le b^0(1) \le (1-\delta)g'(\overline{x}_\delta+)$. In particular for every $\delta \le \delta_0$ there exists $x_\delta \in [\overline{x}_\delta, \overline{y})$ such that

$$(1-\delta)g'(x_\delta) \ge b^0(1). \tag{3.27}$$

Let us choose now a sequence d_n increasing to $+\infty$ with the following properties:

(a) $d_n > \|u_n'\|_\infty$ and $d_n > x_n$ for every $n \in \mathbb{N}$, where x_n is the point appearing in assumption (st2);

(b) $d_n r_n \to 0$ so slowly that $r_n f_n(d_n)/b(d_n r_n) \to 1$ (this is possible thanks to (3.6)).

Setting $\tilde{s}_n := f_n(d_n)/d_n$, it follows from (b) that $\lim_{n\to\infty} \tilde{s}_n = b^0(1)$; we can now pass to a subsequence such that $s_n := \tilde{s}_n \wedge b^0(1)$ increases to $b^0(1)$. Finally, denoting by y_n the first strictly positive point such that $f_n(y_n) = s_n y_n$, we have $y_n \to \overline{y}$. Taking into account all of these facts and recalling (3.8), it is now evident that we can find $\overline{n}_\delta > 0$ such that

(*) $f_n(t) \geq (1-\delta)g(t)$ for every $t \in [0, x_\delta]$;

(**) $f_n'(x_\delta -) > b^0(1) \geq s_n$ and $x_\delta < y_n$ for every $n \geq \overline{n}_\delta$.

At this point, for $k > \overline{n}_\delta$ we define the function ψ_δ^k by induction in the following way:

$$\psi_\delta^k = [(1-\delta)g \wedge s_k t]^{**} \text{ in } [0, d_k] \quad \text{and} \quad \psi_\delta^k = \psi_\delta^k(d_j) + s_{j+1}(t - d_j) \text{ in } [d_j, d_{j+1}] \text{ for } j \geq k.$$

Since $s_n \uparrow b^0(1)$ the function ψ_δ^k is convex with $(\psi_\delta^k)^\infty(1) = b^0(1)$ and $\psi_\delta^k \uparrow [(1-\delta)g \wedge b^0(1)t]^{**}$ as k tends to infinity. Defining

$$\tilde{f}_n(t) := \begin{cases} f_n(t) & \text{if } t \in [0, x_\delta], \\ f_n(x_\delta) + s_n(t - x_\delta) & \text{otherwise,} \end{cases}$$

by (*) and (**), we have

$$\psi_\delta^k(t) \leq \tilde{f}_n(t) \qquad \text{in } [0, d_n]. \tag{3.28}$$

Moreover it turns out that

$$\tilde{f}_n(t) \leq f_n(t) \qquad \text{in } [0, y_n]. \tag{3.29}$$

The last inequality actually holds in $[0, x_n]$ by (**) and by convexity (since $x_n \to +\infty$ we have that $x_n > y_n$, provided n is large enough). Exploiting the concave or convex–concave structure of f_n in $[y_n, d_n]$ it is also easy to prove (see Figure 1 and the proof of Lemma 3.10 for the details of the argument) that

$$\tilde{f}_n(t) \leq s_n t \leq f_n(t) \qquad \text{in } [y_n, d_n]. \tag{3.30}$$

Combining (3.28), (3.29), and (3.30) we obtain that $\psi_\delta^k \leq f_n$ in $[0, d_n]$ for every $n > k$ and so, in particular, $f_n(|u_n'|) \geq \psi_\delta^k(|u_n'|)$ for $n > k$. Finally, choosing a sequence $\delta_n \downarrow 0$, we can extract by diagonalization from the family $(\psi_{\delta_n}^k)_{k,n}$ a subfamily $(\psi_i)_i$ having all the required properties. If g does not satisfy (3.26), then we can proceed in the following way: let x_0 be the last point where g vanishes and define

$$g_k(x) := \begin{cases} 0 & \text{if } x \in [0, x_0], \\ g\left(x - \frac{1}{k}\right) & \text{if } x \geq x_0. \end{cases}$$

It turns out that $g_k^\infty(1) = g^\infty(1)$, $g_k \uparrow g$ as $k \to \infty$, and g_k satisfies (3.26). Hence we can repeat the construction above for every g_k and conclude by diagonalization. If g does not satisfy (3.25), then in particular $g(t) \geq b^0(1)t$ for every $t > 0$ and therefore $g_1 = b^0(1)t$; moreover there exists $\overline{n} \in \mathbb{N}$ such that $f_n'(0+) > (1-\delta)b^0(1)$ for every $n \geq \overline{n}$. If (d_n) and (s_n) are as above, then for $k > \overline{n}$ we define

$$\psi_\delta^k(t) = (1-\delta)s_k t \qquad \text{in } [0, d_k]$$

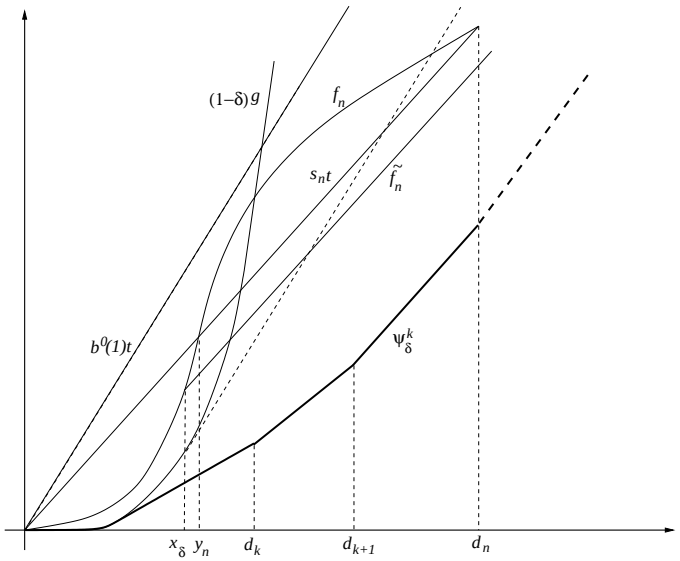


FIG. 1. *The construction of ψ_δ^k in the case $g^\infty(1) > b^0(1)$.*

and

$$\psi_\delta^k(t) = \psi_\delta^k(d_{k+j}) + \left(1 - \frac{\delta}{j+1}\right) s_{k+j+1}(t - d_{k+j}) \text{ in } [d_{k+j}, d_{k+j+1}] \qquad \text{for } j \geq 0,$$

and we argue as above.

Case 2. $b^0(1) \geq g^\infty(1)$. Note that in this case $g_1 = g$. As above it is not restrictive to suppose that $g(0) = f_n(0) = 0$ for every $n \in \mathbb{N}$ and that g satisfies (3.26). At first we choose a sequence $\delta_n \downarrow 0$ and a diverging sequence d_n such that $\|u_n'\|_\infty \leq d_n$ and

$$\lim_{n\to\infty} \frac{f_n(d_n)}{d_n} = b^0(1) \geq g^\infty(1).$$

Recalling (3.8) we can define for every $i \in \mathbb{N}$

$$n_{0,i} := \inf \Big\{ j \in \mathbb{N} : \, j > i,\, f_n(t) \geq (1-\delta_i) g(t) \text{ in } [0, d_i],\, f_n'(d_i-) > (1-\delta_i) g'(d_i-), \\ \text{and } \frac{f_n(d_n)}{d_n} \geq (1-\delta_i) g'(d_i-) \quad \forall n \geq j \Big\},$$

and, for $h \geq 1$,

$$n_{h,i} := \inf \Big\{ j > n_{h-1,i} : \, f_n(t) \geq (1-\delta_{i+h}) g(t) \text{ in } [0, d_{i+h}],\, f_n'(d_{i+h}-) > (1-\delta_{i+h}) g'(d_{i+h}-), \\ \text{and } \frac{f_n(d_n)}{d_n} \geq (1-\delta_{i+h}) g'(d_{i+h}-) \quad \forall n \geq j \Big\}.$$

We define the function ψ_i by induction on h:

$$\psi_i(t) := \begin{cases} (1-\delta_i) g(t) & \text{if } t \in [0, d_i], \\ (1-\delta_i)[g(d_i) + g'(d_i-)(t - d_i)] & \text{if } t \in (d_i, d_{n_{1,i}}] \end{cases}$$

and

$$\psi_i(t) := \psi_k(d_{n_{h,i}}) + (1-\delta_{i+h})g'(d_{i+h}-)(t-d_{n_{h,i}}) \quad \text{in} \quad (d_{n_{h,i}}, d_{n_{h+1,i}}].$$

Clearly $\psi_i^\infty = g^\infty(1)$ for every i and $\psi_i \uparrow g$ as $i \to \infty$. Set for every $h \geq 0$

$$\phi_{i+h}(t) := \begin{cases} (1-\delta_{i+h})g(t) & \text{if } t \in [0, d_{i+h}], \\ (1-\delta_{i+h})[g(d_{i+h}) + g'(d_{i+h}-)(t-d_{i+h})] & \text{if } t > d_{i+h}. \end{cases}$$

First of all, taking into account the definition of $n_{h,i}$ and exploiting the structure assumption on f_n exactly as we did before, one can prove that

$$\phi_{i+h} \leq f_n \qquad \forall\, n \geq n_{h,i}; \tag{3.31}$$

moreover we have

$$\psi_i \leq \phi_{i+h} \qquad \text{in } [0, d_{n_{h+1,i}}]. \tag{3.32}$$

The last inequality is an immediate consequence of

$$\psi_i \leq (1-\delta_{i+h})g \qquad \text{in } [0, d_{i+h}],$$

which can be proved easily by induction on h.

Take $n \geq n_{0,i}$ and let h be such that $n_{h,i} \leq n \leq n_{h+1,i}$; combining (3.31) and (3.32) we finally obtain that $\psi_i \leq f_n$ in $[0, d_n]$. □

LEMMA 3.11. *Suppose that* (3.10) *and condition* (iii) *of Theorem* 3.2 *hold and let* $(u_n)_{n\in\mathbb{N}} \subset W^{2,2}(I)$ *be such that* $\sup_n F_n(u_n) < +\infty$. *Then for every* $\delta > 0$ *there exists a sequence* $(v_n)_{n\in\mathbb{N}} \subseteq SBV(I)$ *such that* $\|u_n - v_n\|_1 \to 0$, $r_n\|v_n'\|_\infty \to 0$ *as* $n \to \infty$, $|v_n'| \leq |u_n'|$ *everywhere, and*

$$F_n(u_n) \geq (1-\delta) \sum_{x \in S_{v_n}} \varphi(v_n^+(x) - v_n^-(x))$$

for n *sufficiently large.*

Proof. By Lemma 3.8 there exists $K > 0$ such that

$$r_n\|u_n'\|_\infty \leq K. \tag{3.33}$$

For every $0 < s \leq K$ we define

$$\omega_n(s) := \sup_{t\in[s,K]} \left| r_n f_n\left(\frac{t}{r_n}\right) - b(t) \right|. \tag{3.34}$$

Recalling that the functions $r_n f_n(\cdot/r_n)$ are monotone, by (3.6) we have that $\omega_n \to 0$ pointwise. As a first step we choose a sequence $(c_n)_n$ of positive real numbers converging to 0 so slowly that

(a) $\frac{r_n}{(c_n)^{\frac{5}{2}}} \to 0$ as $n \to \infty$;

(b) $\lim_{n\to\infty} \frac{\omega_n(c_n)}{b(c_n)} = 0$.

We set $D_n := \{x \in I : |u_n'| > c_n/r_n\} = \bigcup_{k=1}^\infty I_n^k$, where (I_n^k) is the collection of the connected components of D_n; we also denote $I_n^k = (a_n^k, b_n^k)$. Arguing as in Lemma 3.8 and taking into account condition (b), we obtain

$$|D_n| \leq \left(\frac{2\sup_n F_n(u_n)}{b(c_n)}\right) r_n \tag{3.35}$$

for n large enough. For every $n \in \mathbb{N}$ we define

$$\tilde{v}_n(x) := \begin{cases} u_n(x) & \text{if } x \in I \setminus D_n, \\ u_n(a_n^k) & \text{if } x \in (a_n^k, b_n^k), \end{cases}$$

and we set $w_n := u_n - \tilde{v}_n$. Clearly $\tilde{v}_n \in L^1(I) \cap SBV(I)$, and since $w_n' = u_n'$ and $w_n'' = u_n''$ on D_n we have

$$F_n(u_n, I_n^k) = \int_{I_n^k} f_n(|w_n'|)\, dx + (r_n)^3 \int_{I_n^k} |w_n''|^2\, dx.$$

Summing over k and setting $\tilde{z}_n(x) := w_n(r_n x)$, we therefore obtain

$$\begin{aligned} F_n(u_n, D_n) &= \sum_k \left(\int_{I_n^k} f_n(|w_n'|)\, dx + (r_n)^3 \int_{I_n^k} |w_n''|^2\, dx \right) \\ &= \sum_k \left(\int_{I_n^k} f_n\left(\frac{1}{r_n} \left| \tilde{z}_n'\left(\frac{x}{r_n} \right) \right| \right) dx + \frac{1}{r_n} \int_{I_n^k} \left| \tilde{z}_n''\left(\frac{x}{r_n} \right) \right|^2 dx \right) \\ &= \sum_k \left(r_n \int_{\frac{I_n^k}{r_n}} f_n\left(\frac{1}{r_n} |\tilde{z}_n'| \right) dy + \int_{\frac{I_n^k}{r_n}} |\tilde{z}_n''|^2\, dy \right). \end{aligned} \tag{3.36}$$

By (3.33) we have

$$c_n \le |\tilde{z}_n'| \le K \qquad \text{in } D_n/r_n; \tag{3.37}$$

moreover for every $\delta > 0$ there exists $\overline{n}$ such that $r_n f_n(\frac{t}{r_n}) \ge (1-\delta) b(t)$ for every $t \in [c_n, K]$ and for every $n \ge \overline{n}$, and thus, by (3.37),

$$r_n f_n\left(\frac{|\tilde{z}_n'|}{r_n} \right) \ge (1-\delta) b\left(|\tilde{z}_n'| \right) \qquad \text{in } D_n/r_n. \tag{3.38}$$

Indeed, by condition (b), for every $\delta > 0$ we can find $\overline{n}$ such that $\omega_n(c_n) \le \delta b(c_n)$ for every $n \ge \overline{n}$, so that

$$r_n f_n\left(\frac{t}{r_n} \right) \ge b(t) - \omega_n(c_n) \ge b(t) - \delta b(c_n) \ge (1-\delta) b(t) \qquad \forall t \in [c_n, K].$$

Let us define the functions z_n as

$$z_n(x) := \begin{cases} \tilde{z}_n(x) & \text{if } x \in (I \setminus D_n)/r_n, \\ \tilde{z}_n(x) - \tilde{z}_n'\left(\frac{a_k^n}{r_n} \right) \left(x - \frac{a_k^n}{r_n} \right) & \text{in } I_n^k/r_n. \end{cases}$$

By (3.36) and (3.38), and by using the fact that $|z_n'| \le |\tilde{z}_n'|$, we have

$$\begin{aligned} F_n(u_n, D_n) &\ge (1-\delta) \sum_k \left(\int_{\frac{I_n^k}{r_n}} b\left(|\tilde{z}_n'| \right)\, dy + \int_{\frac{I_n^k}{r_n}} |\tilde{z}_n''|^2\, dy \right) \\ &\ge (1-\delta) \sum_k \left(\int_{\frac{I_n^k}{r_n}} b\left(|z_n'| \right)\, dy + \int_{\frac{I_n^k}{r_n}} |z_n''|^2\, dy \right) \\ &\ge (1-\delta) \sum_k \varphi\left(\left| z_n\left(\frac{b_n^k}{r_n} \right) \right| \right) = (1-\delta) \sum_k \varphi(v_n^+(b_n^k) - v_n^-(b_n^k)) = (**), \end{aligned} \tag{3.39}$$

where $v_n(x) := u_n(x) - z_n(x/r_n)$. Using the definition of z_n, it is easy to check that

$$v_n(x) = \begin{cases} u_n(x) & \text{if } x \in I \setminus D_n, \\ u_n(a_n^k) + u_n'(a_n^k)(x - a_n^k) & \text{if } x \in (a_n^k, b_n^k) \end{cases}$$

and $(**) = (1-\delta)\sum_{x\in S_{v_n}} \varphi(v_n^+(x) - v_n^-(x))$. This together with (3.39) gives the thesis of the lemma once we have shown that

$$\|v_n - u_n\|_1 \to 0 \qquad \text{as } n \to \infty. \tag{3.40}$$

If $t \in I_n^k$, by Hölder's inequality, we have

$$\begin{aligned} |v_n(t) - u_n(t)| &\le \int_{a_n^k}^t |v_n'(s) - u_n'(s)|\, ds \le \int_{a_n^k}^t \int_{a_n^k}^s |u_n''(z)|\, dz \\ &\le \left(\int_{I_n^k} |u_n''|^2\, dz \right)^{\frac12} \int_{a_n^k}^t (s - a_n^k)^{\frac12}\, ds \\ &= \frac{2}{3} \left(\int_{I_n^k} |u_n''|^2\, dz \right)^{\frac12} (t - a_n^k)^{\frac32}, \end{aligned}$$

and therefore

$$\int_{I_n^k} |v_n(t) - u_n(t)|\, dt \le \frac{4}{15} \left(\int_{I_n^k} |u_n''|^2\, dz \right)^{\frac12} |I_n^k|^{\frac52}.$$

Using (3.35), we can conclude

$$\begin{aligned} \|u_n - v_n\|_1 &\le \frac{4}{15} \sum_{k\in\mathbb{N}} \left(\int_{I_n^k} |u_n''|^2\, dz \right)^{\frac12} |I_n^k|^{\frac52} \\ &\le \frac{4}{15 (r_n)^{\frac32}} \left((r_n)^3 \int_{D_n} |u_n''|^2\, dz \right)^{\frac12} \left(\sum_{k\in\mathbb{N}} |I_n^k|^5 \right)^{\frac12} \\ &\le \frac{4(\sup_n F_n(u_n))^{\frac12}}{15 (r_n)^{\frac32}} |D_n|^{\frac52} \le \frac{4(\sqrt{2})^5 (\sup_n F_n(u_n))^3}{15} \frac{r_n}{(b(c_n))^{\frac52}}, \end{aligned}$$

which gives (3.40) thanks to condition (a) and (3.10). □

LEMMA 3.12. *Let $g : [0,\infty) \to [0,\infty)$ be a convex superlinear function and let $u \in SBV(I)$ be such that $\int_I g(|u'|)\, dx + \mathcal{H}^0(S_u) < +\infty$. Then there exists a sequence $(u_n) \in SBV(I)$ such that $S_{u_n} \subseteq S_u$, $u_n \in W^{2,2}(I \setminus S_u)$, $u_n'(t\pm) = 0$ on S_u, $u_n \to u$ in $L^\infty(I)$, $u_n^\pm(t) \to u^\pm(t)$ on S_u, and $\int_I g(|u_n'|)\, dx \to \int_I g(|u'|)\, dx$.*

Proof. Let $I = (a,b)$ and $S_u = \{x_1, \dots, x_N\}$ with $x_i < x_{i+1}$ and set $x_0 = a$ and $x_{N+1} = b$. We can construct a family (g_k) of strictly convex and superlinear functions of class C^2 such that

$$g_k'(0) = 0, \qquad g_k \downarrow g, \qquad \text{and} \qquad \lim_{t\to+\infty} \frac{g_k(t)}{g(t)} = 1.$$

For every $k \in \mathbb{N}$ and for every $i \in \{0, \dots, N\}$ let $u_{i,j}^k$ be the solution of the minimum problem

$$\min \left\{ \int_I g_k(|v'|)\, dx + j \int_I |v - u|^2\, dx : \ v \in W^{1,1}(x_i, x_{i+1}) \right\}.$$

Note that the existence of such a solution is guaranteed by the convexity and the superlinearity of g_k; moreover $u^k_{i,j}$ is a classical solution to the Euler equation $h_k''(w')w'' = j(w-u)$ with the Neumann conditions $w'(x_i) = w'(x_{i+1}) = 0$, where h_k is the function in $C^2(\mathbb{R})$ obtained by reflection of g_k. Therefore $u^k_{i,j} \in C^2([x_i, x_{i+1}])$ so that, denoting by u^k_j the function in $SBV(I)$ which coincides with $u^k_{i,j}$ on (x_i, x_{i+1}), we clearly have that the family $(u^k_j)_j$ satisfies all the required conditions except for the last one. By construction, we have $\int_I g_k(|(u^k_j)'|)\,dx \xrightarrow{j} \int_I g_k(|u'|)\,dx$, and since $\int_I g_k(|u'|)\,dx \xrightarrow{k} \int_I g(|u'|)\,dx$ the final approximating sequence can be obtained by diagonalization. □

We finally state a lemma which will be useful in what follows.

LEMMA 3.13. *Denote by $\mathcal{A}(\Omega)$ the family of all open subsets of Ω and let $\nu : \mathcal{A}(\Omega) \to [0, +\infty)$ be a superadditive set-function. Let λ be a positive measure on Ω and let $(\psi_i)_i$ be a family of positive Borel functions such that $\nu(A) \geq \int_A \psi_i\,d\lambda$ for all $A \in \mathcal{A}(\Omega)$ and for all $i \in \mathbb{N}$. Then $\nu(A) \geq \int_A \sup_i \psi_i\,d\lambda$ for all $A \in \mathcal{A}(\Omega)$.*

Proof. See Proposition 1.16 of [10]. □

Proof of Theorem 3.2: *the case* $b^0(1) = +\infty$.

(1) Γ-lim sup inequality.

Let us set $F'' := \Gamma\text{-}\limsup_n F_n$. We first remark that it is enough to show that $F''(u) \leq \int_I g(|u'|)\,dx + \sum_{S_u} \varphi(u^+ - u^-)$ for every $u \in SBV(I)$ with $\mathcal{H}^0(S_u) < +\infty$; indeed the thesis would follow from the semicontinuity of F'' and the fact that $\overline{\mathcal{F}}_{b,g}$ coincides with the relaxed functional of

$$H(u) := \begin{cases} \displaystyle\int_I g(|u'|)\,dx + \sum_{S_u} \varphi(u^+ - u^-) & \text{if } u \in SBV(I) \text{ and } \mathcal{H}^0(S_u) < +\infty, \\ +\infty & \text{otherwise.} \end{cases}$$

Claim 1. Let $u \in SBV(I)$ such that $\mathcal{H}^0(S_u) < +\infty$, $u \in W^{2,2}(I \setminus S_u)$, $F(u) < +\infty$, and $u'(t\pm) = 0$ for every $t \in S_u$. Then

$$F''(u) \leq \int_I g(|u'|)\,dx + \sum_{S_u} \varphi(u^+ - u^-).$$

Since the construction is local we may assume that $S_u = \{\bar t\}$ and $u(t\pm) = u^\pm(\bar t)$.

Fix $\delta > 0$ and choose an admissible pair (η, v) for problem (3.3) (with $z = u^+(\bar t) - u^-(\bar t)$) satisfying

$$\int_0^\eta b(|v'|)\,dx + \int_0^\eta |v''|^2\,dx < \varphi(u^+(\bar t) - u^-(\bar t)) + \delta.$$

We define the recovery sequence in the following way:

$$u_n(x) := \begin{cases} u(x) & \text{if } x \leq \bar t, \\ v\left(\dfrac{x - \bar t}{r_n}\right) + u^-(\bar t) & \text{if } x \in (\bar t, \bar t + r_n\eta), \\ u(x - r_n\eta) + u^+(\bar t) & \text{if } x \geq \bar t + r_n\eta. \end{cases}$$

Clearly $u_n \to u$ in L^1. We can now compute

$$\begin{aligned} F_n(u_n) &= F_n(u_n, I \setminus (\bar t, \bar t + r_n\eta)) + \int_{\bar t}^{\bar t + r_n\eta} f_n\left(\frac{1}{r_n}\left|v'\left(\frac{x-\bar t}{r_n}\right)\right|\right) dx \\ &\quad + (r_n)^3 \int_{\bar t}^{\bar t + r_n\eta} \frac{1}{(r_n)^4}\left|v''\left(\frac{x-\bar t}{r_n}\right)\right|^2 dx \\ &= F_n(u_n, I \setminus (\bar t, \bar t + r_n\eta)) + \underbrace{\int_0^\eta r_n f_n\left(\frac{|v'|}{r_n}\right) dy + \int_0^\eta |v''|^2\, dy}_{\substack{\| \\ (*)_n}}. \end{aligned} \tag{3.41}$$

Since

$$\begin{aligned} &r_n f_n\left(\frac{|v'|}{r_n}\right) \to b(|v'|) \text{ in } \{x \in I : |v'(x)| \neq 0\}, \\ &r_n f_n\left(\frac{|v'|}{r_n}\right) \le r_n f_n\left(\frac{\|v'\|_\infty}{r_n}\right) \to b(\|v'\|_\infty), \end{aligned}$$

by the dominated convergence theorem and the fact that

$$\lim_{n\to\infty} \int_{\{x\in I:\, |v'(x)|\neq 0\}} r_n f_n\left(\frac{|v'|}{r_n}\right) dx = \lim_{n\to\infty} |\{x \in I : |v'(x)| \neq 0\}| r_n f_n(0) = 0,$$

we have

$$\limsup_{n\to\infty} (*)_n \le \int_0^\eta b(|v'|)\, dx + \int_0^\eta |v''|^2\, dx \quad \text{as } n \to \infty;$$

moreover, again by the dominated convergence theorem, we easily see that

$$\lim_{n\to\infty} F_n(u_n, I \setminus (\bar t, \bar t + r_n\eta)) = \int_I g(|u'|)\, dx.$$

From (3.41) we therefore obtain

$$\begin{aligned} \limsup_{n\to\infty} F_n(u_n) &\le \int_I g(|u'|)\, dx + \int_0^\eta b(|v'|)\, dx + \int_0^\eta |v''|^2\, dx \\ &\le \int_I g(|u'|)\, dx + \varphi(u^+(\bar t) - u^-(\bar t)) + \delta. \end{aligned}$$

By the arbitrariness of δ, Claim 1 is proved. By a standard density argument based on the use of Lemma 3.12 we get the same inequality for every $u \in SBV(I)$ with $\mathcal{H}^0(S_u) < +\infty$, and this concludes the proof of the Γ-lim sup inequality, as we remarked above.

(2) Γ-lim inf inequality.

By assumption, g is convex and b is convex, concave, or convex–concave; the functional $F_{b,g}$ is then well defined and coincides with $\overline{\mathcal{F}_{b,g}}$, thanks to Theorem 2.3. We distinguish two cases.

Case 1. $g^\infty(1) = +\infty$; i.e., g is superlinear.

Note that in this case $F_{b,g}(u)$ is finite only if $u \in SBV(I)$ and

$$F_{b,g}(u) = \int_I g(|u'|)\, dx + \sum_{S_u} \varphi(u^+ - u^-).$$

Let $u_n \to u$ in L^1 be such that $\sup_n F_n(u_n) < +\infty$; let $(v_n)_n$ be the sequence constructed in Lemma 3.11 and $(\psi_i)_i$ the approximating family of convex superlinear functions provided by Lemma 3.10. For δ and $\mu \in (0,1)$, and for every open subset $J \subseteq I$, by Lemmas 3.11 and 3.10, we have

$$\begin{aligned} F_n(u_n, J) &= (1-\delta)\int_J f_n(|u_n'|)\,dx + \delta\left[\int_J f_n(|u_n'|)\,dx + ((1/\sqrt[3]{\delta})r_n)^3\int_J |u_n''|^2\,dx\right] \\ &\geq (1-\delta)\int_J \psi_i(|v_n'|)\,dx + \delta(1-\mu)\sum_{x\in S_u}\varphi(v_n^+ - v_n^-) \end{aligned} \tag{3.42}$$

for n sufficiently large. Therefore, by the Ambrosio semicontinuity theorem (recall also Lemma 3.6), we obtain that $u \in SBV(I)$ and

$$\liminf_{n\to\infty} F_n(u_n, J) \geq (1-\delta)\int_J \psi_i(|u'|)\,dx + \delta(1-\mu)\sum_{x\in S_u}\varphi(u^+ - u^-) \qquad \forall i;$$

letting $i \uparrow \infty$ and $\mu \downarrow 0$, we obtain

$$\begin{aligned} (\Gamma\text{-}\liminf_{n\to\infty} F_n)(u, J) &\geq (1-\delta)\int_J g(|u'|)\,dx + \delta\sum_{x\in S_u}\varphi(u^+ - u^-) \\ &= \int_J h^\delta(x)\,d\lambda \quad \forall \text{ open } J \subseteq I,\ \forall \delta \in (0,1), \end{aligned} \tag{3.43}$$

where we have set $\lambda := g(|u'|)\mathcal{L}^1 + \varphi(u^+ - u^-)\mathcal{H}^0$ and $h^\delta := (1-\delta)\,(1-\chi_{S_u}) + \delta\chi_{S_u}$. Let δ_n be a dense sequence in $(0,1)$; since $\sup_i h^{\delta_i} = 1$, from (3.43) we finally deduce that

$$(\Gamma\text{-}\liminf_{n\to\infty} F_n)(u) \geq \int_I \sup_i h^{\delta_i}\,d\lambda = \int_I g(|u'|)\,dx + \sum_{x\in S_u}\varphi(u^+ - u^-),$$

where we applied Lemma 3.13 (with $\nu := (\Gamma\text{-}\liminf_{n\to\infty} F_n)(u,\cdot)$).

Case 2. $g^\infty(1) < +\infty$. Let v_n be as above. According to Lemma 3.10, let $(\psi_i)_i$ be a family of convex functions such that $\psi_i^\infty(1) = g^\infty(1)$ for every $i \in \mathbb{N}$, $\psi_i \uparrow g$ as $i \to \infty$, and $\psi_i(|v_n'|) \leq f_n(|v_n'|)$ for every i and for n sufficiently large. Therefore, by using Lemma 3.11, we can write

$$\begin{aligned} F_n(u_n, I) &= F_n(u_n, D_n) + F_n(u_n, I\setminus D_n) \\ &\geq (1-\delta)\left(\int_I \psi_i(|v_n'|)\,dx + \sum_{x\in S_{v_n}}\varphi(v_n^+ - v_n^-)\right) - \int_{D_n}\psi_i(|v_n'|)\,dx. \end{aligned} \tag{3.44}$$

Using the inequality $\psi_i(t) \leq g(0) + g^\infty(1)t$ and recalling (3.35) and the fact that $\lim_n b(c_n)/c_n = +\infty$, we can estimate

$$\int_{D_n}\psi_i(|v_n'|)\,dx = \psi_i\left(\frac{c_n}{r_n}\right)|D_n| \leq \left(g(0) + g^\infty(1)\frac{c_n}{r_n}\right)\left(\frac{2\sup_n F_n(u_n)}{b(c_n)}\right)r_n = O(1).$$

Invoking the relaxation theorem, Theorem 2.3, from (3.44) we obtain

$$\liminf_{n\to\infty} F_n(u_n) \geq (1-\delta)\left(\int_I \psi_i(|u'|)\,dx + \sum_{S_u}(\varphi\mathbin{\square} g^\infty)(u^+ - u^-) + g^\infty(1)|D^c u|\right).$$

Letting $i \uparrow +\infty$ and $\delta \downarrow 0$ we complete the proof of the Γ-lim inf inequality.

Concerning the last part of the theorem, we first observe that, thanks to (3.17), the approximating functionals are equicoercive; the conclusion then follows from Remark 3.9. □

In order to treat the case

$$b^0(1) = \lim_{t\to 0^+} \frac{b(t)}{t} < +\infty,$$

we first need the following lemma.

LEMMA 3.14. *Suppose that b satisfies* (3.14), *and for every $\delta > 0$ let $\varphi^\delta : (0,\infty) \to (0,\infty)$ be the function defined by*

$$\varphi^\delta(z) := \inf_{\eta>0} \inf \left\{ \int_0^\eta b(|u'|)\,dx + \int_0^\eta |u''|^2\,dx : \, u \in W^{2,2}(0,\eta), \right.$$
$$\left. u(0) = 0,\, u(\eta) = z, u'(0) = u'(\eta) = \delta \right\}.$$

Then the following properties hold true:

(i) $\lim_{\delta\to 0^+} \varphi^\delta(z) = \varphi(z)$ *uniformly in* $[k,+\infty)$ *for every* $k > 0$*;*

(ii) *for every* $\varepsilon \in (0,1)$ *there exists* $\overline{\delta}$ *such that* $\varphi^\delta(z) \geq (1-\varepsilon)\varphi(z)$ *for every* $\delta \leq \overline{\delta}$ *and for every* $z > 0$.

Proof. Fix $k > 0$, let $\phi \in C^2([0,1])$ be such that $\phi(0) = \phi'(0) = 0$ and $\phi(1) = \phi'(1) = 1$, and choose $0 < \overline{\delta} < k/2$ such that

$$\int_0^1 b(\delta|\phi'|)\,dx + \delta^2 \int_0^1 |\phi''|^2\,dx \leq \frac{\varepsilon}{4} \tag{3.45}$$

for every $\delta \leq \overline{\delta}$. Now fix $\delta \in (0,\overline{\delta})$, $z \geq k$, and set $z' := z - 2\delta$. Pick an admissible pair (v,η) for the minimum problem defining $\varphi(z')$ in such a way that

$$\int_0^\eta b(|v'|)\,dx + \int_0^\eta |v''|^2\,dx \leq \varphi(z') + \frac{\varepsilon}{2} \leq \varphi(z) + \frac{\varepsilon}{2}. \tag{3.46}$$

We now define $\tilde{\eta} := \eta + 2$ and $\tilde{v} \in W^{2,2}(0,\tilde{\eta})$ by

$$\tilde{v}(t) := \begin{cases} \delta - \delta\phi(1-t) & \text{if } t \in [0,1), \\ v(t-1) + \delta & \text{if } t \in [1,\eta+1), \\ z - \delta + \delta\phi(t-\eta-1) & \text{if } t \in [\eta+1,\tilde{\eta}]. \end{cases}$$

It is clear that $(\tilde{v},\tilde{\eta})$ is an admissible pair for the minimum problem defining $\varphi^\delta(z)$ so that we have

$$\begin{aligned} \varphi^\delta(z) &\leq \int_0^{\tilde{\eta}} b(|\tilde{v}'|)\,dx + \int_0^{\tilde{\eta}} |\tilde{v}''|^2\,dx \\ &= 2\left(\int_0^1 b(\delta|\phi'|)\,dx + \delta^2 \int_0^1 |\phi''|^2\,dx \right) + \int_0^\eta b(|v'|)\,dx + \int_0^\eta |v''|^2\,dx \\ &\leq \varphi(z) + \varepsilon, \end{aligned} \tag{3.47}$$

where the last inequality follows from (3.45) and (3.46). Now let (v,η) be an admissible pair for $\varphi^\delta(z)$ satisfying

$$\int_0^\eta b(|v'|)\,dx + \int_0^\eta |v''|^2\,dx \leq \varphi^\delta(z) + \frac{\varepsilon}{2}. \tag{3.48}$$

Define $\tilde\eta := \eta + 2$ and $\tilde v$ by

$$\tilde v(t) := \begin{cases} \delta\phi(t) & \text{if } t \in [0,1), \\ v(t-1) + \delta & \text{if } t \in [1, \eta+1), \\ z + 2\delta - \delta\phi(\tilde\eta - t) & \text{if } t \in [\eta+1, \tilde\eta]. \end{cases}$$

As above, we have

$$\begin{aligned} \varphi(z) \le \varphi(z+2\delta) &\le \int_0^{\tilde\eta} b(|\tilde v'|)\,dx + \int_0^{\tilde\eta} |\tilde v''|^2\,dx \\ &= 2\left(\int_0^1 b(\delta|\phi'|)\,dx + \delta^2 \int_0^1 |\phi''|^2\,dx\right) + \int_0^\eta b(|v'|)\,dx + \int_0^\eta |v''|^2\,dx \\ &\le \varphi^\delta(z) + \varepsilon, \end{aligned}$$

thanks to (3.45) and (3.48); recalling (3.47), (i) is proved.

For the last part we suppose by contradiction that there exist $\varepsilon \in (0,1)$, a sequence $\delta_n \downarrow 0$, and a sequence x_n such that

$$\varphi^{\delta_n}(x_n) < (1-\varepsilon)\varphi(x_n) \tag{3.49}$$

for every $n \in \mathbb{N}$. Testing with the pair $(v(t) := \delta t, z/\delta)$, we easily obtain

$$\varphi^\delta(z) \le \frac{b(\delta)}{\delta} z \le C' z \qquad \forall \delta < 1. \tag{3.50}$$

Taking into account (i) we see that (3.49) and (3.50) imply

$$x_n \to 0 \qquad \text{and} \qquad \varphi^{\delta_n}(x_n) \to 0. \tag{3.51}$$

Let (v_n, η_n) be an admissible pair for the minimum problem defining $\varphi^{\delta_n}(x_n)$ such that

$$\int_0^{\eta_n} b(|v_n'|)\,dx + \int_0^{\eta_n} |v_n''|^2\,dx \le \varphi^{\delta_n}(x_n) + (\varphi^{\delta_n}(x_n))^2; \tag{3.52}$$

arguing as in the proof of Lemma 3.6 we deduce that $\|v_n'\|_\infty \to 0$. Choose $\sigma > 0$ such that

$$b(t) \ge \left(1 - \frac{\varepsilon}{2}\right) Ct \qquad \forall t \le \sigma \tag{3.53}$$

and let $\overline n$ be such that $\|v_n'\|_\infty \le \sigma$ for every $n \ge \overline n$. Then, using (3.52), (3.53), (3.50), and (3.51) and recalling that $\varphi(z) \le Cz$ for every $z > 0$ (see (3.15)), we estimate

$$\begin{aligned} \varphi^{\delta_n}(x_n) &\ge \int_0^{\eta_n} b(|v_n'|)\,dx - (\varphi^{\delta_n}(x_n))^2 \ge \left(1 - \frac{\varepsilon}{2}\right) Cx_n - (C'x_n)^2 \\ &\ge \left(1 - \frac{3}{4}\varepsilon\right) Cx_n \ge \left(1 - \frac{3}{4}\varepsilon\right) \varphi(x_n) \end{aligned}$$

for n large enough, in contradiction to (3.49). ∎

We are now in a position to conclude the proof of Theorem 3.2.

Proof of Theorem 3.2: *the case* $b^0(1) < +\infty$.

The Γ-lim sup inequality can be proved as in the other case. We may suppose that $g \not\equiv 0$; otherwise the Γ-lim inf inequality is trivial. Let $\varepsilon_n \to 0$ and $u_{\varepsilon_n} \to u$ in L^1 and such that $\exists \lim_{n\to\infty} F_{\varepsilon_n}(u_{\varepsilon_n}) < +\infty$. Choose now an infinitesimal sequence c_n with the same properties as those in the proof of Lemma 3.11. Set

$$D_n := \left\{ x \in I : \ |u'_{\varepsilon_n}| > \frac{c_n}{r(\varepsilon_n)} \right\} = \bigcup_{k=1}^{\infty} (a_n^k, b_n^k) = \bigcup_{k=1}^{\infty} I_n^k$$

and define

$$v_{\varepsilon_n}(x) := \begin{cases} u_{\varepsilon_n}(x) & \text{if } x \in I \setminus D_n, \\ u_{\varepsilon_n}(a_n^k) & \text{if } x \in (a_n^k, b_n^k). \end{cases}$$

Finally set $w_{\varepsilon_n} := u_{\varepsilon_n} - v_{\varepsilon_n}$ and $z_{\varepsilon_n}(x) := w_{\varepsilon_n}(r(\varepsilon_n)x)$. For $\delta \in (0,1)$ the same arguments used in Lemma 3.11 yield

$$\begin{aligned} F_{\varepsilon_n}(u_{\varepsilon_n}, D_n) &\geq (1-\delta) \sum_k \left(\int_{I_n^k/(r(\varepsilon_n))} b(|z'_{\varepsilon_n}|)\, dx + \int_{I_n^k/(r(\varepsilon_n))} |z''_{\varepsilon_n}|^2\, dx \right) \\ &\geq (1-\delta) \sum_k \inf_{\eta>0} \inf \left\{ \int_0^\eta b(|z'|)\, dx + \int_0^\eta |z''|^2\, dx : \ z \in W^{2,2}(0,\eta), \right. \\ &\qquad \left. z(0) = 0,\, z(\eta) = |w_{\varepsilon_n}(b_n^k)|,\ z'(0) = z'(\eta) = c_n \right\} \\ &= (1-\delta) \sum_k \varphi^{c_n}(|w_{\varepsilon_n}(b_n^k)|) = (1-\delta) \sum_{S_{v_{\varepsilon_n}}} \varphi^{c_n}(v^+_{\varepsilon_n} - v^-_{\varepsilon_n}) \end{aligned} \tag{3.54}$$

for n large enough (φ^{c_n} is the function defined in Lemma 3.14 with $\delta = c_n$). Using (ii) of Lemma 3.14, from (3.54) we deduce

$$F_{\varepsilon_n}(u_{\varepsilon_n}, D_n) \geq (1-\delta)^2 \sum_{S_{v_{\varepsilon_n}}} \varphi(v^+_{\varepsilon_n} - v^-_{\varepsilon_n})$$

for n large enough. Combining the estimate above with Lemma 3.10, we therefore obtain (passing to a subsequence, if needed)

$$\begin{aligned} F_{\varepsilon_n}(u_{\varepsilon_n}) &\geq (1-\delta)^2 \left(\int_{I \setminus D_n} \psi_i(|v'_{\varepsilon_n}|)\, dx + \sum_{S_{v_{\varepsilon_n}}} \varphi(v^+_{\varepsilon_n} - v^-_{\varepsilon_n}) \right) \\ &= (1-\delta)^2 \left(\int_I \psi_i(|v'_{\varepsilon_n}|)\, dx + \sum_{S_{v_{\varepsilon_n}}} \varphi(v^+_{\varepsilon_n} - v^-_{\varepsilon_n}) \right), \end{aligned} \tag{3.55}$$

where ψ_i is the sequence constructed in Lemma 3.10. Since, by Lemma 3.8, $\sup_n \operatorname{Var} v_{\varepsilon_n} \leq \sup_n \operatorname{Var} u_{\varepsilon_n} < +\infty$, Rellich's theorem implies that v_{ε_n} is precompact in L^1, and since $v_{\varepsilon_n} \to u$ in measure (recall that $|D_n| \to 0$), we deduce $v_{\varepsilon_n} \to u$ in L^1. Applying Theorem 2.3 (recall that $b^0(1) = \varphi^0(1)$), from (3.55) we deduce

$$\liminf_{n\to\infty} F_{\varepsilon_n}(u_{\varepsilon_n}) \geq (1-\delta)^2 \left(\int_I \psi_i(|u'|)\, dx + \sum_{S_u} \varphi_1(u^+ - u^-) + (g^\infty(1) \wedge b^0(1)) |D^c u| \right);$$

letting $i \uparrow +\infty$ and $\delta \downarrow 0$ we finally obtain the desired Γ-lim inf inequality. □

Remark 3.15. The structure assumption (iii) of Theorem 3.2 can be slightly weakened without changing the result; more precisely it is sufficient to suppose that there exists a family $(g_n^k)_{n,k}$ of positive continuous nondecreasing functions with the following properties:

(i) $f_n \geq g_n^k$ for every n, $k \in \mathbb{N}$;
(ii) for every $k \in \mathbb{N}$ the family $(g_n^k)_n$ satisfies either (st1) or (st2);
(iii) $g_n^k(t) \to g^k(t)$ for every $t \geq 0$ and $r_n g_n^k(t/r_n) \to b^k(t)$ for every $t > 0$, with g^k and b^k satisfying

$$g^k \uparrow g \qquad \text{and} \qquad (b^k)^0(1) \uparrow b^0(1) \qquad \text{as } k \to \infty.$$

Indeed if G_n^k is the functional associated with g_n^k, then for every $k \in \mathbb{N}$ we have

$$\Gamma\text{-}\liminf_{n\to\infty} F_n \geq \Gamma\text{-}\lim_{n\to\infty} G_n^k = F_{b^k,g^k},$$

where $F_{b^k,g^k} \uparrow F_{b,g}$ as $k \to \infty$.

We want now to show that if $g : [0,+\infty) \to [0,+\infty)$ is any superlinear nondecreasing convex function and $b : [0,+\infty) \to [0,+\infty)$ is an arbitrary concave function with $b^0(1) = +\infty$, then $F_{b,g}$ can be reached by functionals of the form (3.1).

THEOREM 3.16. *Let $g : [0,+\infty) \to [0,+\infty)$ be nondecreasing, convex, and superlinear ($g^\infty(1) = +\infty$) and let $b : [0,+\infty) \to [0,+\infty)$ be nondecreasing and concave with $b(0) = 0$ and $b^0(1) = +\infty$. Then there exists a family (f_ε) of positive, continuous, and nondecreasing functions such that the functionals*

$$F_\varepsilon := \begin{cases} \displaystyle\int_I f_\varepsilon(|u'|)\,dx + \varepsilon^3 \int_I |u''|^2\,dx & \text{if } u \in W^{2,2}(I), \\ +\infty & \text{otherwise in } L^1(I) \end{cases}$$

Γ-converge with respect to the L^1-metric to $F_{b,g}$ as $\varepsilon \to 0^+$.

The theorem is an immediate consequence of Theorem 3.2 and of the following proposition, which is proved in [22] (see Lemmas 6.6 and 6.7 of that work).

PROPOSITION 3.17. *Let g and b be as in the previous theorem. Then the functions f_ε defined by*

$$f_\varepsilon(t) := \min\left\{ g(s) + \frac{1}{\varepsilon} b(\varepsilon(t-s)) : \ s \in [0,t] \right\}$$

are continuous, nondecreasing, and satisfy the following properties:

(i) *$f_\varepsilon(t) \to g(t)$ for every $t \geq 0$;*
(ii) *$\varepsilon f(t/\varepsilon) \to b(t)$ for every $t > 0$;*
(iii) *setting $x_\varepsilon := \sup\{x \geq 0 : \ f_\varepsilon(x) = g(x)\}$, there holds that $f_\varepsilon = g$ in $[0, x_\varepsilon]$ and f_ε is concave in $[x_\varepsilon, +\infty)$; moreover $x_\varepsilon \to +\infty$ as $\varepsilon \to 0^+$.*

We conclude this subsection with some considerations about the asymptotic behavior of the function φ defined in (3.3).

PROPOSITION 3.18.

(i) *Let $b(t) = ct^p$ with $c > 0$ and $p \in [0,1)$. Then $\varphi(z) = m(p) c^{\frac{3}{4-p}} z^{\frac{2+p}{4-p}}$, where*

(3.56)

$$m(p) := \min\left\{ \left[\left(\frac{3}{1-p}\right)^{\frac{1-p}{4-p}} + \left(\frac{1-p}{3}\right)^{\frac{3}{4-p}} \right] \left(\int_0^1 |v''|^2\,dt \right)^{\frac{1-p}{4-p}} \left(\int_0^1 |v'|^p\,dt \right)^{\frac{3}{4-p}} : \right.$$
$$\left. v \in W^{2,2}(0,1),\ v(0) = 0,\ v(1) = 1,\ v'(0) = v'(1) = 0 \right\}.$$

(ii) *Let $b : [0,+\infty) \to [0,+\infty)$ be a concave function with $b^0(1) \neq 0$. Then the function φ defined in* (3.3) *satisfies the growth condition*

$$C_1(\sqrt{z}-1) \le \varphi(z) \le C_2(z+1) \qquad \forall z \ge 0, \tag{3.57}$$

for suitable C_1, $C_2 > 0$.

(iii) *For every $\gamma \in [1/2, 1)$ there exists a concave function b satisfying the hypotheses of Theorem* 3.16 *such that the associated function φ satisfies*

$$\lim_{z\to+\infty} \frac{\varphi(z)}{z^\gamma} = +\infty \qquad \textit{and} \qquad \lim_{z\to+\infty} \frac{\varphi(z)}{z^{\gamma+\varepsilon}} = 0 \qquad \forall \varepsilon > 0. \tag{3.58}$$

Proof. (i) Let us set

$$S_{\eta,z} := \{u \in W^{2,2}(0,\eta) :\ u(0) = 0,\ u(\eta) = z,\ u'(0) = u'(\eta) = 0\}.$$

Noting that for every $v \in S_{\eta,z}$ we can write $v(\cdot) = w(\cdot/\eta)$ with $w \in S_{1,z}$, we can use the definition of φ to compute

$$\begin{aligned}
\varphi(z) &= \inf_\eta \inf_{v\in S_{\eta,z}} \left(c\int_0^\eta |v'|^p\,dt + \int_0^\eta |v''|^2\,dt \right) \\
&= \inf_{w\in S_{1,z}} \inf_\eta \left(c\int_0^\eta \frac{1}{\eta^p} \left| w'\left(\frac{t}{\eta}\right)\right|^p dt + \int_0^\eta \frac{1}{\eta^4}\left|w''\left(\frac{t}{\eta}\right)\right|^2 dt \right) \\
&= \inf_{w\in S_{1,z}} \inf_\eta \left(c\eta^{1-p}\int_0^1 |w'|^p\,ds + \frac{1}{\eta^3}\int_0^1 |w''|^2\,ds \right) \\
&= \inf_{w\in S_{1,z}} \left\{ \left[\left(\frac{3}{1-p}\right)^{\frac{1-p}{4-p}} + \left(\frac{1-p}{3}\right)^{\frac{3}{4-p}} \right] c^{\frac{3}{4-p}} \left(\int_0^1 |w''|^2\,ds\right)^{\frac{1-p}{4-p}} \left(\int_0^1 |w'|^p\,ds\right)^{\frac{3}{4-p}} \right\} \\
&= \inf_{w\in S_{1,1}} \left\{ \left[\left(\frac{3}{1-p}\right)^{\frac{1-p}{4-p}} + \left(\frac{1-p}{3}\right)^{\frac{3}{4-p}} \right] \left(\int_0^1 |w''|^2\,ds\right)^{\frac{1-p}{4-p}} \left(\int_0^1 |w'|^p\,ds\right)^{\frac{3}{4-p}} \right\} c^{\frac{3}{4-p}} z^{\frac{2+p}{4-p}} \\
&= m(p) c^{\frac{3}{4-p}} z^{\frac{2+p}{4-p}}.
\end{aligned}$$

It is clear from the computations above that if v is a solution of problem (3.56), then the pair (η, w) defined by

$$\eta := \left[\frac{3}{c^p(1-p)}\right]^{\frac{1}{4-p}} \left(\frac{\int_0^1 |v''|^2\,ds}{\int_0^1 |v'|^p\,ds}\right)^{\frac{1}{4-p}} z^{\frac{2+p}{4-p}} \qquad \text{and} \qquad w(t) := zv\left(\frac{t}{\eta}\right) \tag{3.59}$$

is optimal.

(ii) Under our assumptions there exists $C > 0$ such that $b(t) \le C(1+t)$ for every $t \ge 0$. Take (η, v) such that $v \in S_{\eta,z}$, v is nondecreasing, and

$$C\eta + \int_0^\eta |v''|^2\,dx = m(0)C^{3/4}\sqrt{z}.$$

Then

$$\varphi(z) \le C\int_0^\eta |v'|\,dx + C\eta + \int_0^\eta |v''|^2\,dx = Cz + m(0)C^{3/4}\sqrt{z} \le C'(1+z).$$

Concerning the reverse inequality, since, under our hypotheses, there exist α, $\beta > 0$ such that $b(t) \ge \alpha t \wedge \beta$, it will be enough to prove the following claim.

Claim. Let $b(t) = \alpha t \wedge \beta$ with α, $\beta > 0$. Then

$$\lim_{z\to+\infty} \frac{\varphi(z)}{m(0)\beta^{3/4}\sqrt{z}} = 1.$$

First of all since $b(t) \le \beta$ we immediately obtain by the previous point that

$$\varphi(z) \le m(0)\beta^{3/4}\sqrt{z}. \tag{3.60}$$

Let $z_n \uparrow +\infty$ and let (η_n, z_n) be an admissible pair for $\varphi(z_n)$ such that v_n is nondecreasing and

$$\int_0^{\eta_n} (\alpha|v_n'| \wedge \beta)\,dx + \int_0^{\eta_n} |v_n''|^2\,dx < \varphi(z_n) + 1. \tag{3.61}$$

Let $\sigma_n \in (0,1)$ be such that $\int_{\{x\in I:\,|v_n'|\le\beta/\alpha\}} |v_n'|\,dx = \sigma_n z_n$; since, by (3.60) and (3.61),

$$m(0)\beta^{3/4}\sqrt{z_n} + 1 \ge \int_0^{\eta_n} (\alpha|v_n'| \wedge \beta)\,dx \ge \int_{\{x\in I:\,|v_n'|\le\beta/\alpha\}} \alpha|v_n'|\,dx = \alpha\sigma_n z_n,$$

it follows that $\sigma_n \to 0$. Consider the sets $D_n := \{x \in I : |v_n'| > \beta/\alpha\} = \cup_{k=1}^{\infty} I_n^k$, where $(I_n^k)_k$ is the collection of the connected components of D_n. We denote also $I_n^k := (a_n^k, b_n^k)$. Let $\Phi \in C^2([0,1])$ be such that $\Phi(0) = \Phi'(0) = 0$, $\Phi(1) = 1$, and $\Phi'(1) = \beta/\alpha$ and, for every $t \in [1, 1 + |D_n|]$, set

$$i_n(t) := \min\left\{ k : \sum_{j=1}^{k} |I_n^j| \ge t - 1 \right\}, \qquad \tau_n(t) := t - 1 - \sum_{j=1}^{i_n(t)-1} |I_n^j|.$$

We can now define the new sequence of admissible pairs $(\tilde\eta_n, \tilde v_n)$ by $\tilde\eta_n := |D_n| + 2$ and $\tilde v_n(t) = \int_0^t \tilde v_n'(s)\,ds$, where

$$\tilde v_n' := \begin{cases} \Phi'(s) & \text{if } s \in [0,1], \\ v_n'(a_n^{i_n(s)} + \tau_n(s)) & \text{if } s \in [1, \tilde\eta_n - 1], \\ \Phi'(\tilde\eta_n - s) & \text{if } s \in [\tilde\eta_n - 1, \tilde\eta_n]. \end{cases}$$

Note that $\tilde v_n$ is constructed by gluing together the pieces of v_n defined by the sets I_n^k; since $v_n'(a_n^k) = v'(b_n^k) = \beta/\alpha$ for every k, we have $\tilde v_n \in W^{2,2}(0, \tilde\eta_n)$. Therefore $\tilde v_n \in S_{\tilde\eta_n, \tilde z_n}$ with $\tilde z_n := (1 - \sigma_n)z_n + 2$. Since by construction

$$\int_{D_n} (\alpha|v_n'|\wedge\beta)\,dx + \int_{D_n} |v_n''|^2\,dx = \beta\tilde\eta_n + \int_0^{\tilde\eta_n} |\tilde v_n''|^2\,dx - 2\left(\int_0^1 |\Phi'|\,dx + \int_0^1 |\Phi''|^2\,dx\right),$$

recalling (3.61) and (i) we can estimate

$$\begin{aligned} \varphi(z_n) + 1 &\ge \beta\tilde\eta_n + \int_0^{\tilde\eta_n} |\tilde v_n''|^2\,dx - 2\left(\int_0^1 |\Phi'|\,dx + \int_0^1 |\Phi''|^2\,dx\right) \\ &\ge \inf_{\eta>0} \inf_{S_{\eta,\tilde z_n}} \left(\beta\eta + \int_0^{\eta} |\tilde v''|^2\,dx\right) - 2\left(\int_0^1 |\Phi'|\,dx + \int_0^1 |\Phi''|^2\,dx\right) \\ &= m(0)\beta^{3/4}\sqrt{(1-\sigma_n)z_n + 2} - 2\left(\int_0^1 |\Phi'|\,dx + \int_0^1 |\Phi''|^2\,dx\right), \end{aligned}$$

whence, taking into account that $\sigma_n \to 0$,

$$\liminf_{z\to+\infty} \frac{\varphi(z)}{m(0)\beta^{3/4}\sqrt{z}} \geq 1.$$

The claim is proved.

(iii) For simplicity we treat in detail only the case $\gamma = 1/2$. We take $b(t) := 1+\log(1+t)$ for $t > 0$ and $b(0) = 0$. Fix $p \in (0,1)$ and take (η, w) with $w \in S_{\eta,z}$ such that

$$\int_0^\eta |w'|^p\,dx + \int_0^\eta |w''|^2\,dx = m(p)z^{\frac{2-p}{4-p}} \quad \text{and} \quad \eta \leq c(p)z^{\frac{2-p}{4-p}}.$$

This is possible thanks to (i) (see (3.59)). Then since $b(t) \leq 1 + t^p$ we have

$$\varphi(z) \leq (m(p) + c(p))z^{\frac{2-p}{4-p}}; \tag{3.62}$$

since as p varies in $(0,1)$ the exponent $(2-p)/(4-p)$ varies in $(1/2,1)$, from (3.62) we deduce that

$$\lim_{z\to+\infty} \frac{\varphi(z)}{z^{(1/2)+\varepsilon}} = 0 \qquad \forall \varepsilon > 0.$$

Now take two positive sequences (α_n) and (β_n) with $\beta_n \to +\infty$ such that $b(t) \geq b_n(t) := \alpha_n t \wedge \beta_n$ for every $t \geq 0$ and for every $n \in \mathbb{N}$. Denoting by φ_n the function associated with b_n, the previous claim yields

$$\liminf_{z\to+\infty} \frac{\varphi(z)}{\sqrt{z}} \geq \lim_{z\to+\infty} \frac{\varphi_n(z)}{\sqrt{z}} = m(0)\beta_n^{3/4}$$

for every $n \in \mathbb{N}$; letting $n \to \infty$ we eventually complete the proof of (3.58). If γ is any number in $(1/2,1)$, take $b(t) = t^p \log(1+t)$, where p satisfies $\gamma = (2-p)/(4-p)$, and argue as above. □

4. Some applications. Given a positive function $p(\varepsilon)$ such that $\lim_{\varepsilon\to0^+} p(\varepsilon) = 0$, we consider the functionals

$$F_\varepsilon(u) = \begin{cases} \dfrac{1}{\varepsilon}\displaystyle\int_I f(\varepsilon^{1/q}|u'|)\,dx + (p(\varepsilon))^3 \int_I |u''|^2\,dx & \text{if } u \in W^{2,2}(I), \\ +\infty & \text{otherwise in } L^1(I). \end{cases} \tag{4.1}$$

Our purpose is to classify all the possible Γ-limits generated by the family (F_ε) depending on the asymptotic behavior of the "rescaling" function p. Let us begin by considering the case $q > 1$. Let $f : [0,+\infty) \to [0,+\infty)$ be a nondecreasing continuous function satisfying the following properties:

(H1) f is concave in $(x_1, +\infty)$ for some $x_1 > 0$;

(H2) $\lim_{x\to0^+} \frac{f(x)}{x^q} = \alpha > 0$, with $q > 1$;

(H3) $\lim_{x\to+\infty} \frac{f(x)}{x} = 0$.

We will show that there exists a unique (up to asymptotic equivalence) rescaling function $r(\varepsilon)$ such that the corresponding Γ-limit is a nontrivial "free-discontinuity" functional. Setting $h(x) := f(x)/x$, such a rescaling function is defined as

$$r(\varepsilon) := \frac{\varepsilon^{1/q}}{h^{-1}\left(\varepsilon^{1/q'}\right)}, \tag{4.2}$$

where q is the exponent appearing in (H2) and q' denotes its Lebesgue conjugate exponent (i.e., $1/q + 1/q' = 1$).

Remark 4.1. Since h is decreasing for x large enough (this is a consequence of the concavity of f), the function r is well defined for ε small enough. Moreover the sublinearity of f yields

$$\lim_{n\to\infty} \frac{\varepsilon_n^{1/q}}{r(\varepsilon_n)} = \lim_{n\to\infty} h^{-1}(\varepsilon_n^{1/q'}) = +\infty$$

since $h \downarrow 0$ as $x \to +\infty$.

The main result of this section is stated in the following theorem.

THEOREM 4.2. *Let $I \subset \mathbb{R}$ be a bounded interval and let f and $p(\varepsilon)$ be as before. Finally let $(\varepsilon_n)_{n\in\mathbb{N}}$ be an infinitesimal sequence such that*

$$\lim_{n\to\infty} \frac{p(\varepsilon_n)}{r(\varepsilon_n)} = a > 0 \qquad \text{and} \qquad \exists \lim_{n\to+\infty} \frac{f\left(t\frac{{\varepsilon_n}^{1/q}}{r(\varepsilon_n)}\right)}{f\left(\frac{{\varepsilon_n}^{1/q}}{r(\varepsilon_n)}\right)} =: b(t) \qquad \forall t > 0. \tag{4.3}$$

If $b^0(1) = +\infty$, then the functionals F_{ε_n} (defined in (4.1)) Γ-converge with respect to the L^1-metric to

$$F(u) := \begin{cases} \alpha \displaystyle\int_I |u'|^q \, dx + \sum_{x\in S_u} \varphi^{(a)}(u^+(x) - u^-(x)) & \text{if } u \in SBV(I), \\ +\infty & \text{otherwise in } L^1(I), \end{cases} \tag{4.4}$$

where $\varphi^{(a)}$ is defined by (3.3) with $b^{(a)}(t) := ab(t/a)$ instead of $b(t)$.

If $b^0(1) = C < +\infty$, then $\Gamma\text{-}\lim_{n\to\infty} F_{\varepsilon_n} = F$ with F given by

$$F(u) := \begin{cases} \displaystyle\int_I g(|u'|)\, dx + \sum_{x\in S_u} \varphi^{(a)}(u^+(x) - u^-(x)) + C|D^c u| & \text{if } u \in BV(I), \\ +\infty & \text{if } x \in L^1(I) \setminus BV(I), \end{cases} \tag{4.5}$$

where $g := (\alpha x^q \wedge Cx)^{**}$. Moreover, in both cases, every sequence u_n such that $\sup_n (F_n(u_n) + \|u_n\|_1) < +\infty$ is strongly precompact in L^p for every $p \geq 1$.

An easy consequence of the theorem is the fact that, up to asymptotic equivalence, the function r defined in (4.2) is the unique nontrivial rescaling function; this is made precise by the following corollary, whose easy proof is left to the reader (see [2]).

COROLLARY 4.3. *Let I, f, and r be as in Theorem* 3.2. *Let $(\varepsilon_n)_{n\in\mathbb{N}}$ and $(a_n)_{n\in\mathbb{N}}$ be two sequences converging to 0 and set*

$$F_n(u) = \begin{cases} \dfrac{1}{\varepsilon_n} \displaystyle\int_I f({\varepsilon_n}^{1/q}|u'|)\, dx + (a_n)^3 \int_I |u''|^2 \, dx & \text{if } u \in W^{2,2}(I), \\ +\infty & \text{otherwise in } L^1(I). \end{cases}$$

If $\lim_{n\to\infty} a_n/r(\varepsilon_n) = 0$, then $\Gamma\text{-}\lim_{n\to\infty} F_n = 0$ with respect to the L^1-metric; if $\lim_{n\to\infty} a_n/r(\varepsilon_n) = +\infty$, then the functionals F_n Γ-converge to

$$F(u) := \begin{cases} \alpha \displaystyle\int_I |u'|^q \, dx & \text{if } u \in W^{1,q}(I), \\ +\infty & \text{otherwise in } L^1(I). \end{cases}$$

Remark 4.4. Since f is concave for x large, it follows that $b(\cdot)$ is in turn concave. Moreover, taking into account also the sublinear growth of f, we deduce the existence of $x_2 \geq x_1$ such that

$$f(a+b) \leq f(a) + f(b) \qquad \forall a, b > x_2; \tag{4.6}$$

if f is unbounded, then we have

$$\begin{aligned} b(t) &\leq \limsup_{x\to+\infty} \frac{f(tx)}{f(x)} \leq \limsup_{x\to+\infty} \frac{f(([t]+1)x)}{f(x)} \\ &\leq \limsup_{x\to+\infty} \frac{([t]+1)f(x)}{f(x)} = [t]+1, \end{aligned} \tag{4.7}$$

where $[t]$ denotes the integer part of t; if f is bounded, we get trivially $b(t) \equiv 1$. Finally note that $b(t) > 0$ for any $t > 0$.

Proof of Theorem 4.2. Setting $r_n := p(\varepsilon_n)$ and $f_n(t) := \frac{1}{\varepsilon_n} f(\varepsilon_n^{1/q} t)$, by (H2) we get immediately that $f_n(t) \to \alpha t^q$ for every $t \geq 0$; moreover, using the identity $f(\frac{\varepsilon_n^{1/q}}{r(\varepsilon_n)}) = \frac{r(\varepsilon_n)}{\varepsilon_n}$, which follows easily from the definition of r (see (4.2)), for $t > 0$ we have

$$\begin{aligned} r_n f_n\left(\frac{t}{r_n}\right) &= \frac{p(\varepsilon_n)}{r(\varepsilon_n)} \frac{r(\varepsilon_n)}{\varepsilon_n} f\left(\frac{\varepsilon_n^{1/q}}{r(\varepsilon_n)} \frac{r(\varepsilon_n)}{p(\varepsilon_n)} t\right) \\ &= \frac{p(\varepsilon_n)}{r(\varepsilon_n)} \frac{f\left(\frac{\varepsilon_n{}^{1/q}}{r(\varepsilon_n)} \frac{r(\varepsilon_n)}{p(\varepsilon_n)} t\right)}{f\left(\frac{\varepsilon_n{}^{1/q}}{r(\varepsilon_n)}\right)} \overset{n\to\infty}{\longrightarrow} ab\left(\frac{t}{a}\right) = b^{(a)}(t), \end{aligned} \tag{4.8}$$

where we used (4.3). By the first part of Theorem 3.2 we therefore obtain the Γ-lim sup inequality. By Theorem 3.2 and Remark 3.15, the Γ-lim inf inequality will be proved if for every $\delta > 0$ we construct a family of functions (f_n^δ) such that $f_n \geq f_n^\delta$, f_n^δ satisfies the structure condition (st2), and finally

$$f_n^\delta(t) \to (1-\delta)\alpha t^q \qquad \forall t \geq 0, \qquad\qquad r_n f_n^\delta\left(\frac{t}{r_n}\right) \to b^{(a)}(t) \qquad \forall t > 0. \tag{4.9}$$

It is also clear that if f^δ verifies

(a) $f \geq f^\delta$,

(b) $\lim_{t\to 0^+} \frac{f^\delta(t)}{t^q} = (1-\delta)\alpha$,

(c) $\lim_{t\to+\infty} \frac{f^\delta(t)}{f(t)} = 1$,

(d) there exists $\overline{x}$ such that f^δ is convex in $[0, \overline{x}]$ and concave in $[\overline{x}, +\infty)$,

then the family $f_n^\delta(t) := \frac{1}{\varepsilon_n} f^\delta(\varepsilon_n^{1/q} t)$ meets all the required conditions. Therefore we are left only with the construction of such an f^δ. By assumption we know that there exist $x' < x''$ such that $f(t) \geq (1-\delta)\alpha t^q$ for every $t \in [0, x']$ and f is concave in $[x'', +\infty)$. Define $a(t) := (1-\delta)\alpha(x'^q/x'')t$, $g := [\min\{(1-\delta)\alpha t^q, a(t)\}]^{**}$, and finally

$$f^\delta(t) := \begin{cases} g(t) & \text{if } t \leq x'', \\ f(t) + g(x'') - f(x'') & \text{if } t \geq x''; \end{cases}$$

it is easy to see that f^δ satisfies all conditions (a), ..., (d) (see Figure 2). Finally the equicoerciveness of the family (F_{ε_n}) follows again from Theorem 3.2. ∎

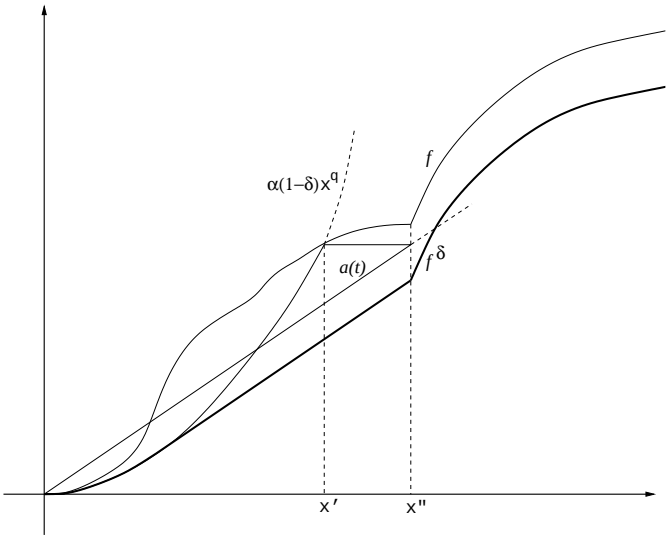


FIG. 2. *The construction of* f^δ.

An easy consequence of Theorem 4.2 is the following compactness result.

THEOREM 4.5. *Let* I, f, *and* r *be as above and consider the family of functionals* F_ε *defined in* (4.1) *with* $p(\varepsilon)$ *satisfying* $0 < \liminf_{\varepsilon\to 0^+} \frac{p(\varepsilon)}{r(\varepsilon)} \le \limsup_{\varepsilon\to 0^+} \frac{p(\varepsilon)}{r(\varepsilon)} < +\infty$. *Then for every infinitesimal sequence* $(\varepsilon_n)_n$ *there exist a subsequence, still denoted by* $(\varepsilon_n)_n$, *and a concave nondecreasing function* b *such that* (F_{ε_n}) Γ*-converges to a functional* F *which is either as in* (4.4) *or as in* (4.5).

Proof. It is sufficient to extract a subsequence such that (4.3) holds and then apply Theorem 4.2. The existence of such a subsequence is an easy consequence of Helly's theorem. □

PROPOSITION 4.6. *Let* f *be a function satisfying* (H1), (H2), *and* (H3) *and let us suppose in addition that*

$$\exists \lim_{x\to+\infty} \frac{f(tx)}{f(x)} =: b(t) \qquad \forall t > 0. \tag{4.10}$$

Let F_ε *be the functional defined in* (4.1), *with* p *satisfying* $\lim_{\varepsilon\to 0^+} \frac{p(\varepsilon)}{r(\varepsilon)} = a > 0$ *(*r *is the rescaling function defined in* (4.2)*). If* $b^0(1) = +\infty$, *then the family* F_ε Γ*-converges to*

$$F(u) := \begin{cases} \displaystyle\int_I |u'|^q\,dx + m(\gamma)a^{\frac{3(1-\gamma)}{4-\gamma}} \sum_{x\in S_u} (u^+ - u^-)^{\frac{2+\gamma}{4-\gamma}} & \text{if } u \in SBV(I), \\ +\infty & \text{in } L^1(I)\setminus SBV(I), \end{cases}$$

where $\gamma = \log b(\mathrm{e})$ *and* $m(\gamma)$ *is the constant defined in* (3.56). *If* $b^0(1) < +\infty$, *the family* (F_ε) Γ*-converges to*

$$F(u) := \begin{cases} \displaystyle\int_I g_\alpha(|u'|)\,dx + |D^s u| & \text{if } u \in BV(I), \\ +\infty & \text{in } L^1(I)\setminus BV(I), \end{cases}$$

where $g_\alpha = (\alpha x^q \wedge x)^{**}$.

Proof. From Theorem 4.2 it is clear that $\Gamma\text{-}\lim_{\varepsilon\to 0^+} F_\varepsilon = F$, where F is the functional defined either in (4.4) or in (4.5). It remains only to prove that if (4.3) holds, then

$$\varphi^{(a)}(z) = m(\gamma) a^{\frac{3(1-\gamma)}{4-\gamma}} z^{\frac{2+\gamma}{4-\gamma}} \qquad \forall z > 0, \tag{4.11}$$

and

$$\varphi^{(a)}(z) = z \qquad \forall z > 0 \tag{4.12}$$

otherwise. First of all note that from (4.10) it follows immediately that $b(st) = b(s)b(t)$ for t, $s > 0$ and therefore $b(t) = t^\gamma$ for $t > 0$, with $\gamma = \log b(\mathrm{e})$; by Remark 4.4 (and in particular by (4.7)), we have that $\gamma \in [0,1]$. If $\gamma < 1$, then (4.11) follows from Proposition 3.18 since $b^{(a)}(t) = ab(t/a) = a^{1-\gamma} t^\gamma$. If $\gamma = 1$, then by Lemma 3.5 we get (4.12). □

Let us now look at some examples. We will use the following notation: given two functions r_1 and r_2 we will write $r_1 \simeq r_2$ if $\lim_{\varepsilon\to 0^+} r_1(\varepsilon)/r_2(\varepsilon) = 1$.

Example 4.7. Let γ belong to $[0,1)$ and set $f(x) := \alpha x^2/(1 + x^{2-\gamma})$; using the definitions (see (4.2) and (4.3)), it is easy to see that $r(\varepsilon) \simeq \varepsilon^{\frac{2-\gamma}{2-2\gamma}}$ and $b(t) = t^\gamma$. Therefore, setting

$$F_\varepsilon(u) := \begin{cases} \displaystyle\int_I \frac{\alpha |u'|^2}{1 + \varepsilon^{\frac{2-\gamma}{2}} |u'|^{2-\gamma}}\, dx + a^3 \varepsilon^{\frac{6-3\gamma}{2-2\gamma}} \int_I |u''|^2\, dx & \text{if } u \in W^{2,2}(I), \\ +\infty & \text{otherwise in } L^1(I), \end{cases}$$

we have that the functionals F_ε Γ-converge to

$$F^\gamma(u) := \begin{cases} \displaystyle \alpha \int_I |u'|^2\, dx + m(\gamma) a^{\frac{3(1-\gamma)}{4-\gamma}} \sum_{x \in S_u} (u^+ - u^-)^{\frac{2+\gamma}{4-\gamma}} & \text{if } u \in SBV(I), \\ +\infty & \text{in } L^1(I) \setminus SBV(I). \end{cases} \tag{4.13}$$

We recover in this way the result by Bouchitté, Dubs, and Seppecher (see [9]). Note that as γ varies in $[0,1)$ the exponent $\frac{2+\gamma}{4-\gamma}$ varies in $\left[\frac{1}{2}, 1\right)$.

Example 4.8. Let $f(x) := (1 + x^\gamma)\log(1 + \alpha x^2)$ with $\gamma \in [0,1)$. We claim that

$$r(\varepsilon) \simeq (1-\gamma)^{\frac{1}{1-\gamma}} \frac{\varepsilon^{\frac{2-\gamma}{2-2\gamma}}}{\left(\log \frac{1}{\varepsilon}\right)^{\frac{1}{1-\gamma}}}. \tag{4.14}$$

Indeed, with the same notation as that of Theorem 3.2, we have

$$\lim_{\varepsilon\to 0^+} r(\varepsilon) \frac{\left(\log \frac{1}{\varepsilon}\right)^{\frac{1}{1-\gamma}}}{\varepsilon^{\frac{2-\gamma}{2-2\gamma}}} = \lim_{\varepsilon\to 0^+} \frac{\sqrt{\varepsilon}}{h^{-1}(\sqrt{\varepsilon})} \frac{\left(\log \frac{1}{\varepsilon}\right)^{\frac{1}{1-\gamma}}}{\varepsilon^{\frac{2-\gamma}{2-2\gamma}}} = \lim_{y\to+\infty} \frac{\left(\log \frac{1}{h^2}\right)^{\frac{1}{1-\gamma}}}{y h^{\frac{1}{1-\gamma}}} = (1-\gamma)^{\frac{1}{1-\gamma}},$$

where we used the change of variable $y = h^{-1}(\sqrt{\varepsilon})$.

We finally observe that $b(t) = \lim_{x\to+\infty} f(tx)/f(x) = t^\gamma$ for all $t > 0$; therefore, setting

$$F^\gamma_\varepsilon(u) := \begin{cases} \displaystyle \frac{1}{\varepsilon} \int_I (1 + \varepsilon^{\frac{\gamma}{2}} |u'|^\gamma) \log(1 + \varepsilon\alpha |u'|^2)\, dx \\ \qquad\qquad + \left(a^{1-\gamma} (1-\gamma) \frac{\varepsilon^{\frac{2-\gamma}{2}}}{\log \frac{1}{\varepsilon}} \right)^{\frac{3}{1-\gamma}} \displaystyle\int_I |u''|^2\, dx & \text{if } u \in W^{2,2}(I), \\ +\infty & \text{otherwise in } L^1(I), \end{cases}$$

we obtain that the sequence F_ε^γ Γ-converges to the functional F^γ defined in (4.13). In particular, taking $\gamma = 0$, we see that the family of singular perturbations of the rescaled Perona–Malik energy given by

$$F_\varepsilon(u) := \begin{cases} \dfrac{1}{\varepsilon}\displaystyle\int_I \log(1+\varepsilon\alpha|u'|^2)\,dx + \left(\frac{a\varepsilon}{\log\frac{1}{\varepsilon}}\right)^3 \int_I |u''|^2\,dx & \text{if } u \in W^{2,2}(I), \\ +\infty & \text{otherwise in } L^1(I) \end{cases}$$

Γ-converge to F^0, as stated in the introduction.

Remark 4.9. Let f_1 and f_2 be two functions satisfying the hypotheses of Theorem 3.2 and let r_1 and r_2 be the rescaling functions associated with f_1 and f_2, respectively, according to (4.2). For $\varepsilon > 0$ and $i = 1, 2$ set

$$F_{i,\varepsilon}(u) := \begin{cases} \dfrac{1}{\varepsilon}\displaystyle\int_I f_i(\sqrt{\varepsilon}|u'|)\,dx + (r_i(\varepsilon))^3 \int_I |u''|^2\,dx & \text{if } u \in W^{2,2}(I), \\ +\infty & \text{otherwise in } L^1(I). \end{cases}$$

Suppose in addition that

$$\lim_{x\to+\infty} \frac{f_1(x)\log^\gamma(x)}{f_2(x)} = 1. \tag{4.15}$$

Then for any infinitesimal sequence $(\varepsilon_n)_n$ $\Gamma\text{-}\lim_{n\to\infty} F_{1,\varepsilon_n} = F \iff \Gamma\text{-}\lim_{n\to\infty} F_{2,\varepsilon_n} = F$; in other words, functions which differ asymptotically by a logarithmic factor generate the same Γ-limits. To prove this fact we pass to a subsequence such that

$$\exists \lim_{n\to+\infty} \frac{f_1\left(t\frac{\sqrt{\varepsilon_n}}{r_1(\varepsilon_n)}\right)}{f_1\left(\frac{\sqrt{\varepsilon_n}}{r_1(\varepsilon_n)}\right)} =: b_1(t) \qquad \forall t > 0$$

and

$$\exists \lim_{n\to+\infty} \frac{f_2\left(t\frac{\sqrt{\varepsilon_n}}{r_2(\varepsilon_n)}\right)}{f_2\left(\frac{\sqrt{\varepsilon_n}}{r_2(\varepsilon_n)}\right)} =: b_2(t) \qquad \forall t > 0,$$

and we observe that (4.15) yields $b_1 \equiv b_2$; we conclude by applying Theorem 4.2. Note that the results of Example 4.8 can be derived from those of Example 4.7 by using the present remark.

Remark 4.10. The hypothesis (H3) is in some sense necessary; indeed, suppose that f is an increasing function satisfying (H1), (H2), and $\lim_{x\to+\infty} \frac{f(x)}{x} = C > 0$. Then it is easy to see that the functionals

$$G_\varepsilon(u) := \begin{cases} \dfrac{1}{\varepsilon}\displaystyle\int_I f(\sqrt{\varepsilon}|u'|)\,dx & \text{if } u \in C^1(I), \\ +\infty & \text{otherwise in } L^1(I) \end{cases}$$

Γ-converge to the functional G given by

$$G(u) := \begin{cases} \alpha\displaystyle\int_I |u'|^q dx & \text{if } u \in W^{1,q}(I), \\ +\infty & \text{otherwise in } L^1(I). \end{cases}$$

We leave the details to the reader.

Let us treat the case where the function f in (4.1) has a finite strictly positive derivative at the origin so that $q = 1$.

THEOREM 4.11. *Let $f : [0, +\infty) \to [0, +\infty)$ be continuous, nondecreasing, differentiable in 0 with $f'(0) > 0$, and concave in $(x_1, +\infty)$ for a suitable $x_1 > 0$. Then the family*

$$F_\varepsilon := \begin{cases} \dfrac{1}{\varepsilon}\displaystyle\int_I f(\varepsilon|u'|)\,dx + \varepsilon^3 \int_I |u''|^2\,dx & \text{if } u \in W^{2,2}(I), \\ +\infty & \text{otherwise in } L^1(I) \end{cases}$$

Γ-converges to the functional

$$F(u) := \begin{cases} f'(0)\displaystyle\int_I |u'|\,dx + \sum_{S_u} \varphi(u^+ - u^-) + f'(0)|D^c u| & \text{if } u \in BV(I), \\ +\infty & \text{otherwise in } L^1(I), \end{cases}$$

where φ is the function defined in (3.3) with $b = f$. Moreover, every sequence u_ε such that $\sup_\varepsilon (F_\varepsilon(u_\varepsilon) + \|u_\varepsilon\|_1) < +\infty$ is strongly precompact in L^p for every $p \geq 1$.

Proof. Take an infinitesimal sequence (ε_n) and consider the family of functions $f_n := (1/\varepsilon_n) f(\varepsilon_n \cdot)$: we clearly have that $f_n(t) \to f'(0)t$ for every $t > 0$ and $\varepsilon_n f_n(t/\varepsilon_n) = f(t)$ for every $t \geq 0$ and every $n \in \mathbb{N}$, so that (f_n) verifies (3.5) and (3.6) with $g = f'(0)t$ and $b = f$. Now construct a sequence of functions (f^k) such that $f \geq f^k$ for every k, $(f^k)'(0) \uparrow f'(0)$ as $k \to \infty$, and f^k is linear in $[0, y_k]$ and concave in $[y_k, +\infty)$ for a suitable $y_k > 0$ (it is clear that under our assumptions such a construction is possible); then, setting $f^k_n(t) := (1/\varepsilon_n) f^k(\varepsilon_n t)$, we have that the family $(f^k_n)_{n,k}$ satisfies the weaker structure assumption introduced in Remark 3.15. At this point we can conclude by applying Theorem 3.2. ∎

The following example is in the spirit of Proposition 3.17.

Example 4.12. Given a convex nondecreasing positive function g and a concave positive function b satisfying $b^0(1) = g^\infty(1) = C \in (0, +\infty)$, we have that the family

$$F_\varepsilon := \begin{cases} \displaystyle\int_I \left[g(|u'|) \wedge \left(\frac{1}{\varepsilon} b(\varepsilon|u'|) + g(0) \right) \right] dx + \varepsilon^3 \int_I |u''|^2\,dx & \text{if } u \in W^{2,2}(I), \\ +\infty & \text{otherwise} \end{cases}$$

Γ-converges to the functional $F_{b,g}$ defined in (3.4), i.e., to

$$\int_I g(|u'|)\,dx + \sum_{S_u} \varphi(u^+ - u^-) + C|D^c u| \qquad u \in BV(I),$$

where φ is the function associated with b according to (3.3). It is enough to apply Theorem 3.2 to the family $f_\varepsilon := g(t) \wedge \left(\frac{1}{\varepsilon} b(\varepsilon t) + g(0)\right)$ after noting that

$$f_\varepsilon(t) \to g(t) \wedge (b^0(t) + g(0)) = g(t) \qquad \text{and} \qquad \varepsilon f_\varepsilon\left(\frac{t}{\varepsilon}\right) \to b(t) \wedge g^\infty(t) = b(t).$$

5. The *N*-dimensional case. In this section we seek to extend the results of the previous sections to the N-dimensional case. Let us fix first some notation: for $u \in W^{2,2}(\Omega)$ we denote its hessian matrix by $\nabla^2 u$ and, given a matrix A, we consider the norm defined by

$$\|A\| := \sup_{|\xi|=1} A\,\xi \cdot \xi\,.$$

It is convenient to introduce the following definition.

DEFINITION 5.1. *Given $X \subseteq L^1(\Omega)$ we say that the sequence of functionals $F_n : X \to \mathbb{R} \cup \{+\infty\}$* steadily Γ-converges *in X to $F : X \to \mathbb{R} \cup \{+\infty\}$ (and we will write $\Gamma^s\text{-}\lim_{n\to\infty} F_n = F$ or $F_n \overset{\Gamma^s}{\to} F$) if, for every $p \geq 1$, $F_n|_{X\cap L^p(\Omega)}$ Γ-converges to $F|_{X\cap L^p(\Omega)}$ with respect to the L^p-convergence. Equivalently we have that $\Gamma^s\text{-}\lim_{n\to\infty} F_n = F$ if and only if the two following conditions are satisfied:*

(i) *for every $(u_n)_n \subset X$ such that $u_n \to u \in X$ in L^1 there holds*

$$\liminf_{n\to\infty} F_n(u_n) \geq F(u);$$

(ii) *for every $u \in X \cap L^p(\Omega)$ there exists a sequence $(u_n)_n \subset X \cap L^p(\Omega)$ such that*

$$u_n \to u \text{ in } L^p \qquad \text{and} \qquad \limsup_{n\to\infty} F_n(u_n) \leq F(u).$$

We will also say that G is the steady relaxed functional *of F if G is the Γ^s-limit of the constant sequence $F_n = F$.*

The main result of this section is the following theorem.

THEOREM 5.2. *Let $\Omega \subset \mathbb{R}^N$ be an open bounded set with Lipschitz boundary and let f_n, r_n satisfy hypotheses* (i), (ii), *and* (iii) *of Theorem* 3.2. *For every $n \in \mathbb{N}$ consider the following functional F_n:*

$$F_n^N(u) = \begin{cases} \displaystyle\int_\Omega f_n(|\nabla u|)\,dx + (r_n)^3 \int_\Omega \|\nabla^2 u\|^2\,dx & \text{if } u \in W^{2,2}(\Omega), \\ +\infty & \text{otherwise in } L^1(\Omega). \end{cases} \tag{5.1}$$

Then

$$\Gamma\text{-}\liminf_{n\to\infty} F_n^N \geq F_{b,g}^N, \tag{5.2}$$

with respect to the L^1-convergence, where $F_{b,g}^N$ is the N-dimensional version of $F_{b,g}$ (see (3.4)*) given by*

$$F_{b,g}^N(u) := \begin{cases} \displaystyle\int_\Omega g_1(|\nabla u|)\,dx + \int_{S_u} \varphi_1(u^+(x) - u^-(x))\,d\mathcal{H}^{N-1} \\ \qquad + (g^\infty(1) \wedge b^0(1))|D^c u| & \text{if } u \in GBV(\Omega), \\ +\infty & \text{otherwise.} \end{cases}$$

Suppose now that f_n satisfies the following additional growth conditions:

(gr1) *there exist C_1, $C_0 > 0$ and $q \geq 1$ such that $f_n(t) \leq C_1(1 + t^q)$ and $C_0 t^q \leq g(t) \leq C_1(1 + t^q)$ for every $t \geq 0$;*

(gr2) *for every $\alpha > 0$ there exists $c(\alpha) > 0$ such that $f_n(\alpha t) \leq c(\alpha) f_n(t)$ for every $t \geq 0$,*

where C_1, C_0, q, and $c(\alpha)$ are independent of n. Then the sequence (F_n^N) Γ^s-converges in $GSBV^q$ to $F_{b,g}^N$. Moreover, if $g^\infty(1) \wedge b^0(1) < +\infty$, then $\Gamma^s\text{-}\lim_{n\to\infty} F_n^N = F_{b,g}^N$ in the whole space $L^1(\Omega)$.

Remark 5.3. Note that if g satisfies both (gr1) and (gr2), then also the family (f_ε) constructed in Proposition 3.17 verifies the same growth conditions.

Proof. Let us prove (5.2). The inequality will be proved by means of the so-called slicing method, which relies on the use of Theorem 2.1.

Let us suppose for simplicity that $g^\infty(1) \wedge b^0(1) = +\infty$; the other case can be treated in an analogous way. First of all we observe that for $\xi \in S^{n-1}$, for $u \in W^{2,2}(\Omega)$,

and for $A \in \mathcal{A}(\Omega)$ we have

$$\begin{aligned} F_n^N(u,A) &= \int_{\Pi_\xi} \int_{A_\xi^y} \left(f_n(|\nabla u(y+t\xi)|) + (r_n)^3 \|\nabla^2 u(y+t\xi)\|^2 \right) dt\, d\mathcal{H}^{N-1}(y) \\ &\geq \int_{\Pi_\xi} \int_{A_\xi^y} \left(f_n(|(u_\xi^y)'|) + (r_n)^3 |(u_\xi^y)''|^2 \right) dt\, d\mathcal{H}^{N-1}(y) \\ &= \int_{\Pi_\xi} F_n(u_\xi^y, A_\xi^y)\, d\mathcal{H}^{N-1}(y), \end{aligned}$$

where Π_ξ is the hyperplane orthogonal to ξ, while A_ξ^y and u_ξ^y are the one-dimensional sections defined in subsection 2.1. Let $u_n \to u$ in $L^1(A)$ be such that $\sup_n F_n^N < +\infty$, and note that for every $\xi \in S^{n-1}$ and for almost every $y \in A_\xi^y$ the one-dimensional sections $(u_n)_\xi^y$ belong to $W^{2,2}(A_\xi^y)$ and $(u_n)_\xi^y \to (u)_\xi^y$ in $L^1(A_\xi^y)$. Hence, by using the results of the previous sections and Fatou's lemma, we have

$$\begin{aligned} \liminf_{n\to\infty} F_n^N(u_n,A) &\geq \int_{\Pi_\xi} \liminf_{n\to\infty} F_n((u_n)_\xi^y, A_\xi^y)\, d\mathcal{H}^{N-1}(y) \\ &\geq \int_{\Pi_\xi} \left(\int_{A_\xi^y} g(|(u_\xi^y)'|) + \sum_{(S_u)_\xi^y \cap A_\xi^y} \varphi((u_\xi^y)^+ - (u_\xi^y)^-) \right) \mathcal{H}^{N-1}(y). \end{aligned} \tag{5.3}$$

From (5.3), by virtue of Theorem 2.1, we deduce $u \in GSBV(\Omega)$ and

$$\begin{aligned} \Gamma\text{-}\liminf_{n\to\infty} F_n^N(u,A) &\geq \alpha \int_A g(|\nabla u \cdot \xi|)\, dx + \int_{S_u \cap A} \varphi(u^+ - u^-) |\nu_u \cdot \xi|\, d\mathcal{H}^{N-1} \\ &= \int_A \psi_\xi(x)\, \lambda, \end{aligned} \tag{5.4}$$

where we set

$$\lambda := \mathcal{L}^N + \varphi(u^+ - u^-)\mathcal{H}^{N-1} \lfloor S_u$$

and

$$\psi_\xi := g(|\nabla u \cdot \xi|)(1 - \chi_{S_u}) + |\nu_u \cdot \xi| \chi_{S_u};$$

since (5.4) holds true for every ξ and for every $A \in \mathcal{A}(\Omega)$, we can choose a dense sequence $(\xi_i)_{i\in\mathbb{N}}$ in S^{n-1} and apply Lemma 3.13 (with $\nu(\cdot) := \Gamma\text{-}\liminf_{n\to\infty} F_n^N(u,\cdot)$) to finally obtain

$$\Gamma\text{-}\liminf_{n\to\infty} F_n^N(u) \geq \int_\Omega \sup_i \psi_{\xi_i}\, d\lambda = \int_\Omega g(|\nabla u|)\, dx + \int_{S_u} \varphi(u^+ - u^-)\, d\mathcal{H}^{N-1},$$

as desired.

Concerning the Γ-lim sup inequality, we adapt the proof given in [3]. In what follows we will assume that (gr1) and (gr2) hold true. For every $p \geq 1$ we denote

$$\begin{aligned} G_p &: L^p(\Omega) \times \mathcal{A}(\Omega) \to [0, +\infty], \\ G_p(u,A) &:= \inf \left\{ \limsup_{n\to\infty} F_n^N(u_n, A) :\ u_n \to u \text{ in } L^p(A) \right\}; \end{aligned}$$

our thesis is then equivalent to proving that $G_p(u,\Omega) \le F^N_{b,g}(u,\Omega)$ for every $u \in GSBV^q(\Omega)\cap L^p(\Omega)$. It is clear that

$$G_{p_1}(u,A) \le G_{p_2}(u,A) \tag{5.5}$$

for $1 \le p_1 < p_2$.

Step 1. Let Π be an affine hyperplane, let Π^+ and Π^- denote the two open half-spaces whose union gives $\mathbb{R}^N \setminus \Pi$, and let ν be the unit normal vector to Π which points toward Π^+. Then, for every $A \in \mathcal{A}(\Omega)$ and for every $z \in \mathbb{R}$, we have

$$G_p(z\chi_{\Pi^+},A) \le \varphi(|z|)\mathcal{H}^{N-1}(\Pi\cap A) = F^N_{b,g}(z\chi_{\Pi^+},A) \qquad \forall p \ge 1.$$

First of all, since

$$\lim_{t\to 0}\mathcal{H}^{N-1}(\{x\in A:\, d(x)=t\}) = \mathcal{H}^{N-1}(\Pi\cap A)$$

for $\delta \in (0,1)$ we can choose $\eta > 0$ such that

$$\sup_{t\in(-\eta,\eta)} \mathcal{H}^{N-1}(\{x\in A:\, d(x)=t\}) \le (1+\delta)\mathcal{H}^{N-1}(\Pi\cap A). \tag{5.6}$$

Let $u_n \to z\chi_{(0,+\infty)}$ be the one-dimensional recovery sequence constructed before and satisfying $\|u_n\|_\infty \le |z|$, $u_n \equiv z\chi_{(0,+\infty)}$ in $\mathbb{R}\setminus(-\eta,\eta)$, and

$$\lim_{n\to\infty} F_n(u_n,(-\eta,\eta)) = F_{b,g}(z\chi_{(0,+\infty)},(-\eta,\eta)) = \varphi(|z|); \tag{5.7}$$

we recall also that

$$r_n\|u_n'\|_\infty \le K \tag{5.8}$$

for a suitable $K > 0$. For every $x \in \Omega$ we define $v_n(x) := u_n(d(x))$, where d is the signed distance function from Π, positive in Π^+ and negative in Π^-. Clearly $v_n \in W^{2,2}(\Omega)$ and $v_n \to z\chi_{\Pi^+}$ in $L^p(\Omega)$ for every $p \ge 1$. Moreover, using the coarea formula (see (2.1)), (5.6), and (5.8), we can estimate

$$\begin{aligned}
F^N_n(v_n,A) &= \int_{A\cap(\Pi)_\eta} f_n(|u'(d)|)\,dx + (r_n)^3\int_{A\cap(\Pi)_\eta}\|u_n''(d)\nabla d\otimes\nabla d + u_n'(d)\nabla^2 d\|^2\,dx\\
&\le \int_{-\eta}^{\eta}\int_{\{x\in A:\, d(x)=t\}} f_n(|u_n'(t)|)\,d\mathcal{H}^{N-1}\,dt\\
&\quad + \int_{-\eta}^{\eta}(r_n)^3\int_{\{x\in A:\, d(x)=t\}}((1+\varepsilon)|u_n''(t)|^2 + c_\varepsilon|u_n'(t)|^2\|\nabla^2 d\|)\,d\mathcal{H}^{N-1}\,dt\\
&\le \int_{-\eta}^{\eta} F_n(u_n,(-\eta,\eta))\mathcal{H}^{N-1}(\{x\in A:\, d(x)=t\})\,dt\\
&\quad + c_\varepsilon K r_n\|\nabla^2 d\|_\infty\int_{-\eta}^{\eta}\mathcal{H}^{N-1}(\{x\in A:\, d(x)=t\})\,dt\\
&\le (1+\varepsilon)(1+\delta)\mathcal{H}^{N-1}(\Pi\cap A)F_n(u_n,(-\eta,\eta))\\
&\quad + c_\varepsilon(1+\delta)\mathcal{H}^{N-1}(\Pi\cap A)2K\eta r_n\|\nabla^2 d\|_\infty,
\end{aligned}$$

where $(\Pi)_\eta$ denotes the η-neighborhood of Π. From the last inequality, taking into account (5.7), we deduce

$$\limsup_{n\to\infty} F^N_n(v_n,A) \le (1+\varepsilon)(1+\delta)\mathcal{H}^{N-1}(\Pi\cap A)\varphi(|z|);$$

since δ and ε are arbitrary, Step 1 is proved.

Step 2. Let $u = \sum_{i=1}^k z_i \chi_{E_i}$ with E_i closed polyhedra such that $\overset{\circ}{E}_i \cap \overset{\circ}{E}_j = \emptyset$ for $i \neq j$. Then

$$G_p(u, A) \leq F^N_{b,g}(u, A)$$

for all $A \in \mathcal{A}(\Omega)$ and for every $p \geq 1$.

The proof combines Step 1 with a standard partition of unity argument; we refer to Proposition 2.6 of [3] for the details.

Step 3. Let A', A, $B \in \mathcal{A}(I)$ such that $A' \subset\subset A$ and let ϕ be a cut-off function between A' and A. Then there exists a positive constant $C > 0$ such that, for every u, $v \in W^{2,2}(\Omega) \cap L^q(\Omega)$, we have

$$\begin{aligned}(5.9)\quad F^N_n(\phi u + (1-\phi)v, A' \cup B) &\leq F^N_n(u, A) + F^N_n(v, B)\\ &+ C(F^N_n(u, S) + F^N_n(v, S)) + C\|\nabla\phi\|^q_\infty \|u - v\|^q_{L^q(S)} + C\mathcal{L}^N(S)\\ &+ C(r_n)^3(\|\nabla\phi\|^2_\infty \|\nabla u - \nabla v\|^2_{L^2(S)} + \|\nabla^2\phi\|^2_\infty \|u - v\|^2_{L^2(S)}),\end{aligned}$$

where $S := (A \setminus \overline{A'}) \cap B$.

Using the monotonicity of f, we can estimate

$$\begin{aligned}&F^N_n(\phi u + (1-\phi)v, A' \cup B)\\ &\leq F^N_n(u, A) + F^N_n(v, B) + \int_S f_n(|(u-v)\nabla\phi| + \phi|\nabla u| + (1-\phi)|\nabla v|)\,dx\\ &+ C_1(r_n)^3 \int_S (|\nabla\phi|^2|\nabla u - \nabla v|^2 + |u-v|^2\|\nabla^2\phi\|^2 + \|\nabla^2 u - \nabla^2 v\|^2)\,dx\\ &\leq F^N_n(u, A) + F^N_n(v, B) + \int_S (f_n(3|(u-v)\nabla\phi|) + f_n(3\phi|\nabla u|) + f_n(3(1-\phi)|\nabla v|))\,dx\\ &+ C_1(r_n)^3 \int_S (|\nabla\phi|^2|\nabla u - \nabla v|^2 + |u-v|^2\|\nabla^2\phi\|^2 + \|\nabla^2 u\|^2 + \|\nabla^2 v\|^2)\,dx =: (*).\end{aligned}$$

Using (gr2) we can continue our estimate:

$$\begin{aligned}(*) &\leq F^N_n(u, A) + F^N_n(v, B) + (C_1 + c(3))(F^N_n(u, S) + F^N_n(v, S))\\ &+ \int_S f_n(3|(u-v)\nabla\phi|)\,dx\\ &+ C_1(r_n)^3 \int_S (|\nabla\phi|^2|\nabla u - \nabla v|^2 + |u-v|^2\|\nabla^2\phi\|^2)\,dx.\end{aligned}$$

Recalling (gr1), from the last inequality we easily get (5.9).

Step 4. Let A', A, B, and S be as in Step 3 and $p \geq q$. Then for every u, $v \in L^p(\Omega)$ and for every $K \in \mathbb{N}$ there exists a cut-off function ϕ_K between A and A' such that

$$\begin{aligned}(5.10)\qquad G_p(\phi_K u + (1-\phi_K)v, A' \cup B) &\leq \left(1 + \frac{C}{K}\right)(G_p(u, A) + G_p(v, B))\\ &+ C\frac{K^{q-1}}{d^q}\|u - v\|^q_{L^q(S)} + \frac{C}{K}\mathcal{L}^N(S),\end{aligned}$$

where $d := \operatorname{dist}(A', \Omega \setminus A)$.

First of all choose u_n, $v_n \in W^{2,2}(\Omega)$ such that $u_n \to u$, $v_n \to v$ in $L^p(\Omega)$ and

$$G_p(u, A) = \lim_{n\to\infty} F^N_n(u_n, A) \qquad \text{and} \qquad G_p(u, B) = \lim_{n\to\infty} F^N_n(v_n, B).$$

For $j \in \{0, 1, \dots, K\}$ we consider the set

$$A_j^K = \left\{ x \in A : \ \operatorname{dist}(x, A') < j\frac{d}{K} \right\}.$$

For any $j \in \{0, 1, \dots, K-1\}$ we choose a cut-off function ϕ_j^K between A_j^K and A_{j+1}^K such that

$$\|\nabla \phi_j^K\|_\infty \le 2\frac{K}{d}; \tag{5.11}$$

finally we set $S_j^K := \left(A_{j+1}^K \setminus \overline{A}_j^K\right) \cap B$. By using (5.9) and (5.11), we get

$$\begin{aligned}
&F_n^N(\phi_j^K u_n + (1-\phi_j^K)v_n, A' \cup B) \\
&\le F_n^N(u_n, A) + F_n^N(v_n, B) + C(F_n^N(u_n, S_j^K) + F_n^N(v_n, S_j^K)) + C\frac{K^q}{d^q}\|u - v\|^q_{L^q(S_j^K)} \\
&\quad + C(r_n)^3(\|\nabla \phi_j^K\|_\infty^2 \|\nabla u_n - \nabla v_n\|^2_{L^2(S_j^K)} + \|\nabla^2 \phi_j^K\|_\infty^2 \|u_n - v_n\|^2_{L^2(S_j^K)}).
\end{aligned}$$

Passing to a subsequence if needed, it follows that there exists $j_K \in \{0, 1, \dots, K-1\}$ such that

$$\begin{aligned}
F_n^N(\phi_{j_K}^K u_n + (1-\phi_{j_K}^K)v_n, A' \cup B) &\le \frac{1}{K}\sum_{j=0}^{K-1} F_n^N(\phi_j^K u_n + (1-\phi_j^K)v_n, A' \cup B) \\
&\le \left(1 + \frac{C}{K}\right)(F_n^N(u_n, A) + F_n^N(v_n, B)) + C\frac{K^{q-1}}{d^q}\|u_n - v_n\|^q_{L^q(S)} \\
&\quad + \frac{C}{K}\mathcal{L}^N(S) + C(K)(r_n)^3(\|\nabla u_n - \nabla v_n\|^2_{L^2(S)} + \|u_n - v_n\|^2_{L^2(S)})
\end{aligned} \tag{5.12}$$

for every $n \in \mathbb{N}$. Recall that by the Nirenberg inequality (see [25]) there exists $M > 0$ such that

$$\|\nabla u\|_{L^2(S)} \le M(\|\nabla^2 u\|^{1/2}_{L^2(S)}\|u\|^{1/2}_{L^2(S)} + \|u\|_{L^2(S)})$$

for all $u \in W^{2,2}(S)$. Hence, from the equiboundedness of

$$(\|u_n - v_n\|_{L^2(S)} + (r_n)^3\|\nabla^2 u_n - \nabla^2 v_n\|^2_{L^2(S)})_n,$$

we get

$$(r_n)^3\|\nabla u_n - \nabla v_n\|^2_{L^2(S)} \to 0 \qquad \text{as } n \to \infty.$$

Thus (5.10) follows letting n tend to $+\infty$ in (5.12).

Step 5. For every $u \in GSBV^q(\Omega) \cap L^\infty(\Omega)$ we have

$$G_p(u, \Omega) \le \int_\Omega g(|\nabla u|)\, dx + \int_{S_u} \varphi(u^+ - u^-)\, d\mathcal{H}^{N-1} \qquad \forall p \ge 1. \tag{5.13}$$

We start with $u \in \mathcal{W}(\Omega)$ (see subsection 2.3) and, for every $h \in \mathbb{N}$, we consider the sets

$$B_h := (S_u)_{1/h} \cap \Omega = \left\{ x \in \Omega : \ \operatorname{dist}(x, S_u) < \frac{1}{h} \right\};$$

by the regularity assumptions on S_u we have that $\mathcal{L}^N(B_h) = O(1/h)$, and therefore, setting

$$\rho_h := h^{-\frac{1}{2}} \left(\int_{B_h} |\nabla u|^2 \right)^{\frac{1}{4}},$$

we have

$$\lim_{h\to\infty} \frac{1}{\rho_h} \int_{B_h} |\nabla u| \, dx = 0. \tag{5.14}$$

By a standard argument based on the use of the coarea formula (2.1) (see, for example, [12]) it is possible to find a sequence u_h satisfying the hypotheses of Step 2 such that

$$\|u - u_h\|_{L^\infty(B_h)} \le \rho_h, \qquad \mathcal{H}^{N-1}((S_{u_h} \cap B_h) \setminus S_u) \le \frac{1}{\rho_h} \int_{B_h} |\nabla u| \, dx + O(1). \tag{5.15}$$

We apply Step 4 with $A = B_h$, $A' = B_{2h}$, $B = \Omega \setminus \overline{B}_{3h}$ to obtain the existence of a cut-off function ϕ_K^h such that

$$\begin{aligned} &G_p(\phi_K^h u_h + (1-\phi_K^h)u, \Omega) \\ &\le \left(1 + \frac{C}{K}\right) (G_p(u, \Omega \setminus \overline{S}_u) + G_p(u_h, B_h)) + C \frac{K^{q-1}}{d^q} \|u - u_h\|^q_{L^q(S)} + \frac{C}{K} \mathcal{L}^N(B_h) \\ &\le \left(1 + \frac{C}{K}\right) (G_p(u, \Omega \setminus \overline{S}_u) + G_p(u_h, B_h)) + \left(C K^{q-1} h^q \rho_h^q + \frac{C}{K} \right) \mathcal{L}^N(B_h), \end{aligned} \tag{5.16}$$

where $p \ge q$. By Step 2 we have

$$\begin{aligned} G_p(u_h, B_h) &\le \int_{S_{u_h} \cap B_h} \varphi(u_h^+ - u_h^-) \, d\mathcal{H}^{N-1} \\ &\le \int_{S_u} \varphi(u^+ - u^-) \, d\mathcal{H}^{N-1} + \varphi(2\|u\|_\infty) \mathcal{H}^{N-1}((S_{u_h} \cap B_h) \setminus S_u) \\ &\quad + \int_{S_u} (\varphi(u_h^+ - u_h^-) - \varphi(u^+ - u^-)) \, d\mathcal{H}^{N-1}; \end{aligned}$$

using (5.15) and (5.14) and noting that, by the dominated convergence theorem,

$$\lim_{h\to\infty} \int_{S_u} (\varphi(u_h^+ - u_h^-) - \varphi(u^+ - u^-) \, d\mathcal{H}^{N-1} = 0,$$

we therefore obtain

$$\limsup_{h\to\infty} G_p(u_h, B_h) \le \int_{S_u} \varphi(u^+ - u^-) \, d\mathcal{H}^{N-1}. \tag{5.17}$$

Moreover, taking as approximating sequence $u_n = u$ for every $n \in \mathbb{N}$, we discover that

$$G_p(u, \Omega \setminus \overline{S}_u) \le \int_\Omega g(|\nabla u|) \, dx. \tag{5.18}$$

Combining (5.17) and (5.18), letting $h \to +\infty$ in (5.16), and taking into account the lower semicontinuity of G_p, we finally get (5.13) for $p \ge q$ and therefore for every

$p \geq 1$, by virtue of (5.5). For a general $u \in GSBV^q(\Omega) \cap L^\infty(\Omega)$ we conclude by using a standard density argument based on Theorem 2.5.

We are now in a position to complete the proof of Theorem 5.2. Take $u \in GSBV^q(\Omega) \cap L^p(\Omega)$ and set $u_k := (-k \vee u) \wedge k$. Then, by (5.13) and the monotone convergence theorem, we have

$$\begin{aligned} G_p(u,\Omega) &\leq \liminf_{k\to\infty} G_p(u_k,\Omega) \\ &\leq \lim_{k\to\infty} \int_\Omega g(|\nabla u_k|)\,dx + \int_{S_u} \varphi(u_k^+ - u_k^-)\,d\mathcal{H}^{N-1} \\ &= \int_\Omega g(|\nabla u|)\,dx + \int_{S_u} \varphi(u^+ - u^-)\,d\mathcal{H}^{N-1}. \end{aligned}$$

If $g^\infty(1) \wedge b^0(1) = +\infty$, then we are done; if that is not the case, then the conclusion follows from the fact that, thanks to Theorem 2.3 and an easy truncation argument, $F^N_{b,g}$ coincides with the steady relaxed functional (see Definition 5.1) of

$$H(u) := \begin{cases} \displaystyle\int_\Omega g(|\nabla u|)\,dx + \int_{S_u} \varphi(u^+ - u^-)\,d\mathcal{H}^{N-1} & \text{if } u \in GSBV^q(\Omega), \\ +\infty & \text{if } u \in L^1(\Omega) \setminus GSBV^q(\Omega). \end{cases} \qquad \square$$

The following two corollaries are an immediate consequence of Theorems 5.2, 4.2, and 4.11.

COROLLARY 5.4. *Let $\Omega \subset \mathbb{R}^N$ be an open bounded set with Lipschitz boundary and let f, r, p be as in Theorem* 4.2. *Let $(\varepsilon_n)_n$ be an infinitesimal sequence such that* (4.3) *holds. If $b^0(1) = +\infty$, then the functionals*

$$F^N_n(u) = \begin{cases} \displaystyle\frac{1}{\varepsilon_n}\int_\Omega f(\sqrt{\varepsilon_n}|\nabla u|)\,dx + (p(\varepsilon_n))^3 \int_\Omega \|\nabla^2 u\|^2\,dx & \text{if } u \in W^{2,2}(\Omega), \\ +\infty & \text{otherwise in } L^1(\Omega) \end{cases}$$

Γ^s-converge in $GSBV^q(\Omega)$ to

$$F^N(u) := \begin{cases} \displaystyle\alpha\int_\Omega |\nabla u|^q\,dx + \int_{S_u} \varphi^{(a)}(u^+(x) - u^-(x))\,d\mathcal{H}^{N-1} & \text{if } u \in GSBV(\Omega), \\ +\infty & \text{otherwise in } L^1(\Omega), \end{cases}$$

where $\varphi^{(a)}$ is as in Theorem 4.2. *If $b^0(1) = C$, then the sequence (F^N_n) Γ^s-converges in $L^1(\Omega)$ to*

$$F^N(u) := \begin{cases} \displaystyle\int_\Omega g(|\nabla u|)\,dx + \int_{S_u} \varphi^{(a)}(u^+(x) - u^-(x))\,d\mathcal{H}^{N-1} + C|D^c u| & \text{if } u \in GBV(\Omega), \\ +\infty & \text{if } u \in L^1(\Omega) \setminus GBV(\Omega), \end{cases}$$

with $\varphi^{(a)}$ still given by (3.3) *and $g = (\alpha x^q \wedge Cx)^{**}$.*

COROLLARY 5.5. *Let $\Omega \subset \mathbb{R}^N$ be an open bounded set with Lipschitz boundary and let f be as in Theorem* 4.11. *Then the family*

$$F^N_\varepsilon := \begin{cases} \displaystyle\frac{1}{\varepsilon}\int_\Omega f(\varepsilon|\nabla u|)\,dx + \varepsilon^3 \int_\Omega \|\nabla^2 u\|^2\,dx & \text{if } u \in W^{2,2}(\Omega), \\ +\infty & \text{otherwise in } L^1(\Omega) \end{cases}$$

Γ^s-converges in $L^1(\Omega)$ to the functional

$$F^N(u) := \begin{cases} f'(0)\displaystyle\int_\Omega |\nabla u|\,dx + \int_{S_u} \varphi(u^+ - u^-)\,d\mathcal{H}^{N-1} + f'(0)|D^c u| & \text{if } u \in BV(\Omega), \\ +\infty & \text{otherwise in } L^1(\Omega), \end{cases}$$

where φ is the function defined in (3.3) *with $b = f$.*

To conclude the N-dimensional analysis, it remains to prove the equicoerciveness of the approximating functionals: this is done in the following proposition.

PROPOSITION 5.6. *Under the same hypotheses as Theorem* 5.2, *let $(u_n)_n \subset L^1(\Omega)$ be equi-integrable and such that*

$$\sup_n F_n^N(u_n) < M < +\infty.$$

Then $(u_n)_n$ is strongly precompact in $L^1(\Omega)$. Suppose in addition that $F_n^N \overset{\Gamma^s}{\to} G$ in $L^1(\Omega)$. Then for every $g \in L^p(\Omega)$ $(p > 1)$ and $\beta > 0$ the solutions u_n of

$$\min\left\{F_n^N(v) + \beta\int_\Omega |v-g|^p\,dx :\ v \in W^{2,2}(\Omega)\right\}$$

converge, up to a subsequence, in the $L^p(\Omega)$-norm to a solution of

$$\min\left\{G(v) + \beta\int_\Omega |v-g|^p\,dx :\ v \in L^1(\Omega)\right\}.$$

Proof. As at the beginning of the proof of Theorem 5.2, we fix $\xi \in S^{n-1}$ and get

$$M \geq F_n^N(u_n) \geq \int_{\Omega_\xi} g_n(y)\,d\mathcal{H}^{N-1}(y), \tag{5.19}$$

where $g_n(y) := F_n((u_n)_\xi^y, \Omega_\xi^y)$. Using the equi-integrability assumption, for $\delta > 0$ we find $\sigma_\delta > 0$ such that

$$\mathcal{L}^N(B) \leq \sigma_\delta \Rightarrow \int_B |u_n|\,dx < \delta \qquad \forall n \in \mathbb{N}. \tag{5.20}$$

Choose $k > 0$ such that

$$\frac{M\mathrm{diam}(\Omega)}{k} \leq \sigma_\delta; \tag{5.21}$$

set $A_{n,k} := \{y \in \Omega_\xi :\ g_n(y) > k\}$ and denote by P_ξ the orthogonal projection on Π_ξ. We now define the new sequence v_n in the following way:

$$v_n(x) := \begin{cases} u_n(x) & \text{if } P_\xi(x) \in \Omega_\xi \setminus A_{n,k}, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\|u_n - v_n\|_{L^1(\Omega)} = \int_{\{x\in\Omega:\, P_\xi(x)\in A_{n,k}\}} |u_n|\,dx$; since by Chebyshev's inequality, (5.19), and (5.21) we have

$$\begin{aligned}\mathcal{L}^N(\{x \in \Omega :\ P_\xi(x) \in A_{n,k}\}) &\leq \mathcal{H}^{N-1}(A_{n,k})\mathrm{diam}(\Omega) \\ &\leq \frac{\|g_n\|_{L^1(\Omega_\xi)}}{k}\mathrm{diam}(\Omega) \leq \frac{M}{k}\mathrm{diam}(\Omega) \leq \sigma_\delta,\end{aligned}$$

recalling (5.20), we obtain $\|u_n - v_n\|_{L^1(\Omega)} \leq \delta$.

Moreover, $F_n((v_n)^y_\xi, \Omega^y_\xi) \le g_n(y)(1-\chi_{A_{n,k}}(y)) \le k$, and therefore, by the one-dimensional results, $(v_n)^y_\xi$ is precompact in $L^1(\Omega^y_\xi)$ for every $y \in \Omega_\xi$. Since the construction can be repeated for every $\delta > 0$ and for every $\xi \in S^{n-1}$, the thesis follows by applying Lemma 2.2.

Concerning the second part, we first observe that

$$\sup_n \left(F_n^N(u_n) + \beta \int_\Omega |u_n - g|^p \, dx \right) \le \sup_n f_n(0)|\Omega| + \beta \int_\Omega |g|^p \, dx < +\infty, \tag{5.22}$$

and therefore, by the first part of the theorem, there exist $u \in L^1(\Omega)$ and a subsequence, still denoted by u_n, such that $u_n \to u$ in L^1. Note that by (5.22) $\sup_n \|u_n\|_{L^p} < +\infty$, which implies that $u_n \rightharpoonup u$ weakly in L^p. Since $F_n^N \overset{\Gamma^s}{\to} G$, there exists $v_n \to v$ in L^p such that $F_n^N(v_n) \to G(u)$, and therefore, by the minimality of u_n,

$$\begin{aligned} G(u) + \beta \int_\Omega |u-g|^p \, dx &\le \liminf_{n\to\infty} \left(F_n^N(u_n) + \beta \int_\Omega |u_n - g|^p \, dx \right) \\ &\le \limsup_{n\to\infty} \left(F_n^N(u_n) + \beta \int_\Omega |u_n - g|^p \, dx \right) \\ &\le \lim_{n\to\infty} \left(F_n^N(v_n) + \beta \int_\Omega |v_n - g|^p \, dx \right) = G(u) + \beta \int_\Omega |u-g|^p \, dx, \end{aligned}$$

whence

$$\begin{aligned} G(u) + \beta \int_\Omega |u-g|^p \, dx &= \lim_{n\to\infty} \left(F_n^N(u_n) + \beta \int_\Omega |u_n - g|^p \, dx \right) \\ &\ge G(u) + \limsup_{n\to\infty} \beta \int_\Omega |u_n - g|^p \, dx \\ &\ge G(u) + \liminf_{n\to\infty} \beta \int_\Omega |u_n - g|^p \, dx \ge G(u) + \beta \int_\Omega |u-g|^p \, dx. \end{aligned}$$

We deduce that $\int_\Omega |u_n - g|^p \, dx \to \int_\Omega |u-g|^p \, dx$, and since $u_n - g \rightharpoonup u - g$ weakly in L^p we conclude that $u_n \to u$ in L^p. The minimality of u follows now from the properties of Γ-convergence. □

We conclude this section by remarking that all the examples contained in section 4 can be generalized to the N-dimensional case by means of Theorem 5.2. In particular let us highlight the following ones.

Example 5.7 (Perona–Malik energy). By Example 4.8 and Theorem 5.2 the functionals

$$F_\varepsilon^N(u) := \begin{cases} \dfrac{1}{\varepsilon} \displaystyle\int_\Omega \log(1 + \varepsilon\alpha|\nabla u|^2) \, dx + \left(\dfrac{a\varepsilon}{\log \frac{1}{\varepsilon}} \right)^3 \int_\Omega \|\nabla^2 u\|^2 \, dx & \text{if } u \in W^{2,2}(\Omega), \\ +\infty & \text{otherwise in } L^1(\Omega) \end{cases}$$

Γ^s-converge in $GSBV^2(\Omega)$ to

$$F^N(u) := \begin{cases} \alpha \displaystyle\int_\Omega |\nabla u|^2 \, dx + m(0) a^{\frac{3}{4}} \int_{S_u} \sqrt{u^+ - u^-} \, d\mathcal{H}^{N-1} & \text{if } u \in GSBV(\Omega), \\ +\infty & \text{in } L^1(\Omega) \setminus GSBV(\Omega). \end{cases}$$

Example 5.8. Let b and g be as in Example 4.12 (and suppose for simplicity $g(0) = 0$); then the family

$$F_\varepsilon^N := \begin{cases} \displaystyle\int_\Omega \left(g(|\nabla u|) \wedge \frac{1}{\varepsilon} b(\varepsilon |\nabla u|) \right) dx + \varepsilon^3 \int_\Omega \|\nabla^2 u\|^2 \, dx & \text{if } u \in W^{2,2}(\Omega), \\ +\infty & \text{otherwise} \end{cases}$$

Γ^s-converges in $L^1(\Omega)$ to the functional

$$\begin{cases} \displaystyle\int_\Omega g(|\nabla u|) \, dx + \int_{S_u} \varphi(u^+ - u^-) \, d\mathcal{H}^{N-1} + C|D^c u| & \text{if } u \in GBV(\Omega), \\ +\infty & \text{otherwise,} \end{cases}$$

where φ is the function defined in (3.3).

Acknowledgments. The author wishes to thank Andrea Braides for suggesting the study of this problem and for stimulating discussions, and Gianni Dal Maso for some useful advise.

REFERENCES

[1] G. Alberti, G. Bouchitté, and P. Seppecher, *Phase transition with the line-tension effect*, Arch. Ration. Mech. Anal., 144 (1998), pp. 1–46.
[2] R. Alicandro, A. Braides, and M. S. Gelli, *Free-discontinuity problems generated by singular perturbations*, Proc. Roy. Soc. Edinburgh Sect. A, 128 (1998), pp. 1115–1129.
[3] R. Alicandro and M. S. Gelli, *Free-discontinuity problems generated by singular perturbations: The n-dimensional case*, Proc. Roy. Soc. Edinburgh Sect. A, 130 (2000), pp. 449–469.
[4] L. Ambrosio, *A compactness theorem for a new class of variational problems*, Boll. Un. Mat. Ital. B (7), 3 (1989), pp. 857–881.
[5] L. Ambrosio, *Existence theory for a new class of variational problems*, Arch. Ration. Mech. Anal., 111 (1990), pp. 291–322.
[6] L. Ambrosio, N. Fusco, and D. Pallara, *Special Functions of Bounded Variation and Free-Discontinuity Problems*, Oxford University Press, Oxford, 2000.
[7] L. Ambrosio and V. Tortorelli, *Approximation of functionals depending on jumps by elliptic functionals via* Γ*-convergence*, Comm. Pure Appl. Math., 43 (1990), pp. 999–1036.
[8] G. Bouchitté, A. Braides, and G. Buttazzo, *Relaxation results for some free-discontinuity problems*, J. Reine Angew. Math., 458 (1995), pp. 1–18.
[9] G. Bouchitté, C. Dubs, and P. Seppecher, *Regular approximation of free-discontinuity problems*, Math. Models Methods Appl. Sci., 10 (2000), pp. 1073–1097.
[10] A. Braides, *Approximation of Free-Discontinuity Problems*, Springer-Verlag, Berlin, 1998.
[11] A. Braides, *Gamma-Convergence for Beginners*, Oxford University Press, Oxford, 2002.
[12] A. Braides and A. Coscia, *The interaction between bulk energy and surface energy in multiple integrals*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 737–756.
[13] A. Braides and G. Dal Maso, *Non-local approximation of the Mumford-Shah functional*, Calc. Var. Partial Differential Equations, 5 (1997), pp. 293–322.
[14] A. Braides, G. Dal Maso, and A. Garroni, *Variational formulation of softening phenomena in fracture mechanics: The one-dimensional case*, Arch. Ration. Mech. Anal., 146 (1999), pp. 23–58.
[15] A. Braides and M. S. Gelli, *Limits of discrete systems with long-range interactions*, J. Convex Anal., 9 (2002), pp. 363–399.
[16] F. Catté, P.-L. Lions, J.-M. Morel, and T. Coll, *Image selective smoothing and edge detection by nonlinear diffusion*, SIAM J. Numer. Anal., 29 (1992), pp. 182–193.
[17] G. Cortesani, *Sequences of non-local functionals which approximate free-discontinuity problems*, Arch. Ration. Mech. Anal., 144 (1998), pp. 357–402.
[18] G. Cortesani and R. Toader, *A density result in SBV with respect to non isotropic energies*, Nonlinear Anal., 38B (1999), pp. 585–604.
[19] G. Dal Maso, *An Introduction to* Γ*-convergence*, Birkhäuser Boston, Cambridge, MA, 1993.
[20] E. De Giorgi and L. Ambrosio, *Un nuovo funzionale del calcolo delle variazioni*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8), 82 (1988), pp. 199–210.

[21] M. Gobbino, *Finite difference approximation of the Mumford-Shah functional*, Comm. Pure Appl. Math., 51 (1998), pp. 197–228.
[22] M. Gobbino and M. G. Mora, *Finite difference approximation of free discontinuity problems*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 567–595.
[23] S. Müller, *Singular perturbations as a selection criterion for periodic minimizing sequences*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 169–204.
[24] D. Mumford and J. Shah, *Optimal approximation by piecewise smooth functions and associated variational problems*, Comm. Pure Appl. Math., 42 (1989), pp. 577–685.
[25] L. Nirenberg, *On elliptic partial differential equations*, Ann. Scuola Norm. Sup. Pisa (3), 13 (1958), pp. 115–162.
[26] P. Perona and J. Malik, *Scale space and edge detection using anisotropic diffusion*, IEEE Trans. Pattern and Mach. Intell., 12 (1990), pp. 629–639.

SIAM J. MATH. ANAL.
Vol. 35, No. 3, pp. 806–822

CONVERGENCE AND TRAVELLING FRONTS IN FUNCTIONAL DIFFERENTIAL EQUATIONS WITH NONLOCAL TERMS: A COMPETITION MODEL*

STEPHEN A. GOURLEY† AND SHIGUI RUAN‡

Abstract. In this paper we consider a two-species competition model described by a reaction-diffusion system with nonlocal delays. In the case of a general domain, we study the stability of the equilibria of the system by using the energy function method. When the domain is one-dimensional and infinite, by employing linear chain techniques and geometric singular perturbation theory, we investigate the existence of travelling front solutions of the system.

Key words. competition-diffusion, equilibrium, stability, travelling front, energy function, geometric singular perturbation

AMS subject classifications. 92D25, 35K57, 35R20

DOI. 10.1137/S003614100139991

1. Introduction. Let $\mathbf{R} = (-\infty, \infty)$, and let Ω be some open bounded region in $\mathbf{R}^N$, $N \leq 3$, with a smooth boundary $\partial\Omega$. Let $\partial/\partial n$ denote the outward normal derivative on $\partial\Omega$ and let Δ be the Laplacian operator. For $1 \leq p \leq \infty$, let $L^p(\Omega)$ denote the Banach space of measurable functions u on Ω satisfying

$$\|u\|_p = \begin{cases} \left(\int_\Omega |u(x)|^p \, dx\right)^{1/p} < \infty & \text{if} \quad 1 \leq p < \infty, \\ \operatorname{ess\,sup}_{x\in\Omega} |u(x)| < \infty & \text{if} \quad p = \infty. \end{cases}$$

In particular, if $p = 2, L^2(\Omega)$ becomes a Hilbert space with the usual inner product $\langle \cdot, \cdot \rangle$ and $\|\cdot\|_2^2 = \langle \cdot, \cdot \rangle$. Also, let $\|\!| \cdot |\!\|_2$ denote the norm in $L^2((0,T); L^2(\Omega; R))$, i.e.,

$$|\!|\!| u |\!|\!|_2 = \left(\int_0^T \|u(s)\|_2^2 \, ds \right)^{1/2}.$$

Let $u_1(t,x)$ and $u_2(t,x)$ denote the population densities of two competitors at time t and location x, and let the diffusivities of the two competitors be d_1 and d_2, respectively. This paper is concerned with the following two-species Lotka–Volterra competition-diffusion model with distributed delays:

$$\begin{aligned} \frac{\partial u_1}{\partial t} &= d_1 \Delta u_1 + u_1 \left(r_1 - a_1 u_1 - b_1 \int_\Omega \int_{-\infty}^t K_1(x, y, t-s) u_2(s,y) \, ds \, dy \right), \\ \frac{\partial u_2}{\partial t} &= d_2 \Delta u_2 + u_2 \left(r_2 - b_2 \int_\Omega \int_{-\infty}^t K_2(x, y, t-s) u_1(s,y) \, ds \, dy - a_2 u_2 \right) \end{aligned} \tag{1.1}$$

*Received by the editors December 19, 2001; accepted for publication (in revised form) February 28, 2003; published electronically October 2, 2003.
http://www.siam.org/journals/sima/35-3/39991.html

†Department of Mathematics and Statistics, University of Surrey, Guildford, Surrey GU2 7XH, UK (s.gourley@surrey.ac.uk).

‡Department of Mathematics, University of Miami, P.O. Box 249085, Coral Gables, FL 33124-4250 (ruan@math.miami.edu). The research of this author was partially supported by the NSERC of Canada and the College of Arts and Sciences at the University of Miami. On leave from Dalhousie University, Halifax, NS, Canada.

for $t > 0$, $x \in \Omega$, under the homogeneous Neumann boundary conditions

$$\frac{\partial u_1}{\partial n} = \frac{\partial u_2}{\partial n} = 0, \quad x \in \partial\Omega, \tag{1.2}$$

and initial conditions

$$u_1(\theta, x) = \phi_1(\theta, x) \geq 0,\ u_2(\theta, x) = \phi_2(\theta, x) \geq 0, \quad (\theta, x) \in (-\infty, 0] \times \Omega, \tag{1.3}$$

where ϕ_1 and ϕ_2 are continuous functions. The parameters r_i, a_i, and b_i, $i = 1, 2$, are all positive constants.

The kernels $K_i(x, y, \sigma)$, $i = 1, 2$, are nonnegative functions which are continuous in $(x, y) \in \bar{\Omega} \times \bar{\Omega}$ for each $\sigma \in [0, \infty)$ and measurable in $\sigma \in [0, \infty)$ for each pair $(x, y) \in \bar{\Omega} \times \bar{\Omega}$. We assume that the kernels depend on both the spatial and the temporal variables. The delays in this type of model formulation are called *spatiotemporal delays* or *nonlocal delays*. This is a formulation that aims to account for the fact that, at previous times, individuals have not necessarily been at the same point in space. See Gourley and Britton [8] for a detailed discussion of this modelling issue on an infinite spatial domain and Gourley and So [9], who more recently have treated the finite domain case, explaining in detail why it leads to the type of delay term we are using in (1.1). See also Yamada [16] and the references cited therein. Gourley and So [9] concentrated on the one-dimensional domain $[0, \pi]$ and showed that on this domain a delayed variable $u(t, x)$, representing a population with diffusivity d, should be modelled in the equations by using a term of the form

$$\int_0^\pi \int_{-\infty}^t G(x, y, t - s) k(t - s) u(s, y)\, ds\, dy,$$

where $k(t)$ is the weight given to the population t time units ago and, in the homogeneous Neumann problem,

$$G(x, y, t) = \frac{1}{\pi} + \frac{2}{\pi} \sum_{n=1}^{\infty} e^{-dn^2 t} \cos nx \cos ny. \tag{1.4}$$

In our formulation we are, for convenience, absorbing the G and k of each delay term into a single kernel $K_i(x, y, t)$. Regarding these kernels K_i, we shall assume that

$$\int_\Omega K_i(x, y, \sigma)\, dx = \int_\Omega K_i(x, y, \sigma)\, dy = k_i(\sigma), \quad \sigma \geq 0, \tag{1.5}$$

and

$$\int_0^\infty k_i(\sigma)\, d\sigma = 1, \quad \sigma\, k_i(\sigma) \in L^1((0, \infty); \mathbf{R}). \tag{1.6}$$

Assumption (1.5), that integration of $K_i(x, y, s)$ with respect to either x or y removes *both* the x and the y dependence, is easily seen to be reasonable when we have in mind that $K_i(x, y, t)$ is a product of the form $G(x, y, t)k(t)$, with G given by (1.4) or the corresponding expression for whatever domain is under consideration.

The local existence of solutions $(u_1(t, x), u_2(t, x))$ to (1.1)–(1.3) follows from the results in Yamada [17] or Ruan and Wu [11]. The comparison theorem for parabolic differential equations implies that $(u_1(t, x), u_2(t, x))$ exists globally such that

$$\begin{aligned} 0 \leq u_1(t, x) &\leq \max\left\{ \frac{r_1}{a_1},\ \sup_{\theta \leq 0} \|\phi_1(\theta, \cdot)\|_{C(\bar{\Omega}; R)} \right\}, \\ 0 \leq u_2(t, x) &\leq \max\left\{ \frac{r_2}{a_2},\ \sup_{\theta \leq 0} \|\phi_2(\theta, \cdot)\|_{C(\bar{\Omega}; R)} \right\} \end{aligned} \tag{1.7}$$

for $x \in \bar{\Omega}$ and $t \in \mathbf{R}$. Also, by the strong maximum principle, if $\phi_1(0,x) \not\equiv 0$ and $\phi_2(0,x) \not\equiv 0$, then we have $u_1(t,x) > 0$, $u_2(t,x) > 0$ for all $x \in \bar{\Omega}$ and $t > 0$.

Notice that system (1.1) has a trivial equilibrium $E_0 = (0,0)$, two semitrivial spatially homogeneous equilibria

$$E_1 = \left(\frac{r_1}{a_1}, 0\right), \quad E_2 = \left(0, \frac{r_2}{a_2}\right),$$

and a positive spatially homogeneous equilibrium

$$E^* = \left(\frac{r_1 a_2 - r_2 b_1}{a_1 a_2 - b_1 b_2}, \ \frac{r_2 a_1 - r_1 b_2}{a_1 a_2 - b_1 b_2}\right), \tag{1.8}$$

provided that $a_1 a_2 \neq b_1 b_2$ and either (i) $r_2 b_1 < r_1 a_2$ and $r_1 b_2 < r_2 a_1$ or (ii) $r_2 b_1 > r_1 a_2$ and $r_1 b_2 > r_2 a_1$. The trivial equilibrium E_0 is of no interest here. The stability of the semitrivial equilibrium E_i means that the ith competitor ($i = 1, 2$) wins the competition. These semitrivial equilibria are of considerable interest ecologically because of the possibility of a transition between the two. In fact we shall prove in this paper that, when the coexistence equilibrium E^* is absent, a transition can occur between E_1 and E_2 in the form of a travelling wave-front solution.

Various special cases of system (1.1) have been studied by many researchers. When the delay kernels are independent of the spatial variable (i.e., when the delays are local), Ruan and Wu [11] studied the stability of the equilibria. See also Ruan and Zhao [12] for competition models with finite delays and Schiaffino and Tesei [13] for a nonlinear competition system. When there are no delays, the stability of the competition-diffusion model was investigated by Zhou and Pao [18]. Delayed competition models without diffusion have been studied by Cushing [2] and by Gopalsamy [6], and the monograph by Wu [15] provides a very comprehensive description of current research into delay-diffusion equations. When the domain $\Omega = (-\infty, \infty)$ and there are no delays, Conley and Gardner [1], Gardner [4], Kan-on [10], and Tang and Fife [14] have shown that the competition-diffusion model has travelling front solutions connecting the boundary equilibria

$$\left(\frac{r_1}{a_1}, \ 0\right) \qquad \text{and} \qquad \left(0, \ \frac{r_2}{a_2}\right). \tag{1.9}$$

The existence of such solutions even for the nondelay problem is a highly nontrivial matter because one is seeking a heteroclinic connection between equilibria in a four-dimensional phase space. The introduction of delays increases the dimension to eight (for the particular delays we consider). However, when the delays are small, considerable progress can be achieved by the use of geometric singular perturbation theory.

In this paper we shall first discuss the stability of the equilibria E_1, E_2, and E^* by using the energy function method (see Yamada [16, 17]). Then, for the case when $\Omega = (-\infty, \infty)$, we will study the existence of travelling front solutions of system (1.1) connecting the two boundary equilibria E_1 and E_2.

2. Convergence. The main result of this section is a theorem on the global stability of each of the equilibria. First, we shall derive an inequality that will be needed in the proof of the main theorem. The hypotheses of this lemma are not restricted to this application (see, in particular, Gourley and So [9]).

LEMMA 2.1. *Let $K(x,y,t) = G(x,y,t)k(t)$, $x, y \in \Omega \subset \mathbf{R}^N$, where $k(t) \geq 0$ and $G(x,y,t)$ is the solution of*

$$\frac{\partial G}{\partial t} = d\,\nabla^2 G, \qquad \frac{\partial G}{\partial n} = 0 \quad \text{on } \partial\Omega, \qquad G(x,y,0) = \delta(x-y). \tag{2.1}$$

Then

$$\left\| \int_\Omega \int_{-\infty}^t K(x,y,t-s)u(s,y)\,ds\,dy \right\|_2 \leq \int_{-\infty}^t k(t-s)\|u(s)\|_2\,ds$$

for any function $u(t,x)$ such that $\partial u/\partial n = 0$ on $\partial\Omega$.

Remark 2.2. Before we prove this lemma let us stress that x and y are both vectors in $\mathbf{R}^N$ here. For the purposes of computing G, ∇^2 is calculated with respect to either of these vectors (say x for definiteness) with the other one, y, held fixed. In the case considered in detail in [9], Ω is one-dimensional, $\nabla^2 = \partial^2/\partial x^2$, and $G(x,y,t)$ is given by (1.4).

Proof of Lemma 2.1. We have

$$\begin{aligned}
&\left\| \int_\Omega \int_{-\infty}^t K(x,y,t-s)u(s,y)\,ds\,dy \right\|_2 \\
&= \left\| \int_{-\infty}^t \int_\Omega K(x,y,t-s)u(s,y)\,dy\,ds \right\|_2 \\
&\leq \int_{-\infty}^t \left\| \int_\Omega K(x,y,t-s)u(s,y)\,dy \right\|_2 ds \\
&= \int_{-\infty}^t k(t-s) \left\| \int_\Omega G(x,y,t-s)u(s,y)\,dy \right\|_2 ds.
\end{aligned}$$

Therefore, we want to show that

$$\left\| \int_\Omega G(x,y,t-s)u(s,y)\,dy \right\|_2 \leq \|u(s)\|_2$$

for $s \leq t$. Let λ_k, $k = 0,1,2,\ldots$, be the eigenvalues of $-d\nabla^2$ under homogeneous Neumann boundary conditions, with corresponding normalized (in L^2) eigenfunctions $\phi_k(x)$ so that

$$-d\nabla^2\phi_k = \lambda_k\phi_k, \qquad \frac{\partial \phi_k}{\partial n} = 0 \quad \text{on } \partial\Omega.$$

Then $\lambda_0 = 0$ with $\phi_0 =$ constant, and $\lambda_k > 0$ for all other k. The solution $G(x,y,t)$ of (2.1) will be given by a Fourier series expansion in terms of these functions $\phi(x)$ with coefficients depending on y. In fact,

$$G(x,y,t) = \sum_{n=0}^\infty e^{-\lambda_n t}\phi_n(x)\phi_n(y).$$

Also, $u(t,x)$ satisfies the boundary conditions and therefore can be expanded in terms of the ϕ_n:

$$u(t,x) = \sum_{n=0}^\infty a_n(t)\phi_n(x).$$

Therefore, since the ϕ_k are orthonormal,

$$\int_\Omega G(x,y,t-s)u(s,y)\,dy = \sum_{n=0}^{\infty} a_n(s)e^{-\lambda_n(t-s)}\phi_n(x),$$

and hence, by Parseval's identity,

$$\begin{aligned}\left\|\int_\Omega G(x,y,t-s)u(s,y)\,dy\right\|_2 &= \left(\sum_{n=0}^{\infty} a_n^2(s)e^{-2\lambda_n(t-s)}\right)^{1/2}\\ &\le \left(\sum_{n=0}^{\infty} a_n^2(s)\right)^{1/2} = \|u(s)\|_2\end{aligned}$$

as desired. The proof is complete.

Next, we state our main theorem of this section.

THEOREM 2.3. *Let $(u_1(t,x),u_2(t,x))$ satisfy (1.1) with boundary conditions (1.2) and initial conditions (1.3), with $\phi_1(0,x)\not\equiv 0$ and $\phi_2(0,x)\not\equiv 0$.*

(i) *If $r_1/r_2 > a_1/b_2 > b_1/a_2$, then $\lim_{t\to\infty}(u_1(t,x),u_2(t,x)) = (r_1/a_1,0)$ uniformly for $x\in\bar\Omega$.*

(ii) *If $r_1/r_2 < b_1/a_2 < a_1/b_2$, then $\lim_{t\to\infty}(u_1(t,x),u_2(t,x)) = (0,r_2/a_2)$ uniformly for $x\in\bar\Omega$.*

(iii) *If $b_1/a_2 < r_1/r_2 < a_1/b_2$, then $\lim_{t\to\infty}(u_1(t,x),u_2(t,x)) = (u_1^*,u_2^*)$ uniformly for $x\in\bar\Omega$, where u_1^* and u_2^* are the components of the equilibrium E^* given by (1.8).*

Proof. We prove only (i); the proofs of (ii) and (iii) are similar. To study the stability of the semitrivial equilibrium $E_1 = (r_1/a_1,0)$, define

$$E(u_1) = \int_\Omega \left[u_1 - \frac{r_1}{a_1} - \frac{r_1}{a_1}\log\frac{u_1}{r_1/a_1}\right]dx, \quad F(u_2) = \int_\Omega u_2\,dx. \tag{2.2}$$

Then $E(u_1)\ge 0$ and $F(u_2)\ge 0$. For some constant $\alpha>0$ to be found later, we have

$$\begin{aligned}&\frac{d}{dt}[\alpha E(u_1)+F(u_2)]\\ &= \alpha\int_\Omega \frac{\partial u_1}{\partial t}\left(1-\frac{r_1/a_1}{u_1}\right)dx + \int_\Omega \frac{\partial u_2}{\partial t}\,dx\\ &= -\alpha d_1\frac{r_1}{a_1}\int_\Omega \frac{|\nabla u_1|^2}{u_1^2}\,dx - \alpha a_1\int_\Omega\left(u_1-\frac{r_1}{a_1}\right)^2 dx\\ &\quad - \alpha b_1\int_\Omega\left(\int_\Omega\int_{-\infty}^t K_1(x,y,t-s)u_2(s,y)\,ds\,dy\right)\left(u_1-\frac{r_1}{a_1}\right)dx - a_2\int_\Omega u_2^2(t,x)\,dx\\ &\quad + r_2\int_\Omega u_2(t,x)\,dx - b_2\int_\Omega\left(\int_\Omega\int_{-\infty}^t K_2(x,y,t-s)u_1(s,y)\,ds\,dy\right)u_2(t,x)\,dx\\ &= -\alpha d_1\frac{r_1}{a_1}\int_\Omega \frac{|\nabla u_1|^2}{u_1^2}\,dx - \alpha a_1\int_\Omega\left(u_1-\frac{r_1}{a_1}\right)^2 dx\\ &\quad - \alpha b_1\int_\Omega\left(\int_\Omega\int_{-\infty}^t K_1(x,y,t-s)u_2(s,y)\,ds\,dy\right)\left(u_1-\frac{r_1}{a_1}\right)dx - a_2\int_\Omega u_2^2(t,x)\,dx\\ &\quad - b_2\int_\Omega\left(\int_\Omega\int_{-\infty}^t K_2(x,y,t-s)\left(u_1(s,y)-\frac{r_2}{b_2}\right)ds\,dy\right)u_2(t,x)\,dx,\end{aligned}$$

where we have used (1.5) and (1.6). By hypothesis, $r_2/b_2 < r_1/a_1$, so

$$\begin{aligned}
&\frac{d}{dt}[\alpha E(u_1) + F(u_2)] \\
&\le -\alpha d_1 \frac{r_1}{a_1} \int_\Omega \frac{|\nabla u_1|^2}{u_1^2}\,dx - \alpha a_1 \int_\Omega \left(u_1 - \frac{r_1}{a_1}\right)^2 dx \\
&\quad - \alpha b_1 \int_\Omega \left(\int_\Omega \int_{-\infty}^t K_1(x,y,t-s) u_2(s,y)\,ds\,dy\right)\left(u_1 - \frac{r_1}{a_1}\right) dx - a_2 \int_\Omega u_2^2(t,x)\,dx \\
&\quad - b_2 \int_\Omega \left(\int_\Omega \int_{-\infty}^t K_2(x,y,t-s)\left(u_1(s,y) - \frac{r_1}{a_1}\right) ds\,dy\right) u_2(t,x)\,dx.
\end{aligned} \tag{2.3}$$

Let $\langle \cdot, \cdot \rangle$ denote the standard inner product on $L^2(\Omega; \mathbf{R}), \|\cdot\|_2^2 = \langle \cdot, \cdot \rangle$. Then we have the following inequality:

$$\begin{aligned}
&\frac{d}{dt}[\alpha E(u_1) + F(u_2)] + \alpha d_1 \frac{r_1}{a_1} \int_\Omega \frac{|\nabla u_1|^2}{u_1^2}\,dx + \alpha a_1 \left\|u_1 - \frac{r_1}{a_1}\right\|_2^2 + a_2 \|u_2\|_2^2 \\
&\quad \le -\alpha b_1 \left\langle \int_\Omega \int_{-\infty}^t K_1(x,y,t-s) u_2(s,y)\,ds\,dy,\ u_1(t) - \frac{r_1}{a_1} \right\rangle \\
&\qquad -b_2 \left\langle \int_\Omega \int_{-\infty}^t K_2(x,y,t-s)\left(u_1(s,y) - \frac{r_1}{a_1}\right) ds\,dy,\ u_2(t) \right\rangle.
\end{aligned} \tag{2.4}$$

By Lemma 2.1, we have

$$\begin{aligned}
\left\| \int_\Omega \int_{-\infty}^t K_1(x,y,t-s) u_2(s,y)\,ds\,dy \right\|_2 &\le \int_{-\infty}^t k_1(t-s) \|u_2(s)\|_2\,ds \\
&\le \sup_{s\le 0} \|u_2(s)\|_2 \int_t^\infty k_1(s)\,ds + \int_0^t k_1(t-s)\|u_2(s)\|_2\,ds
\end{aligned} \tag{2.5}$$

and

$$\begin{aligned}
&\left\| \int_\Omega \int_{-\infty}^t K_2(x,y,t-s)\left(u_1(s,y) - \frac{r_1}{a_1}\right) ds\,dy \right\|_2 \le \int_{-\infty}^t k_2(t-s) \left\|u_1(s) - \frac{r_1}{a_1}\right\|_2 ds \\
&\quad \le \sup_{s\le 0} \left\|u_1(s) - \frac{r_1}{a_1}\right\|_2 \int_t^\infty k_2(s)\,ds + \int_0^t k_2(t-s) \left\|u_1(s) - \frac{r_1}{a_1}\right\|_2 ds.
\end{aligned} \tag{2.6}$$

Thus, for any $T > 0$,

$$\begin{aligned}
&\left| \int_0^T \left\langle \int_\Omega \int_{-\infty}^t K_1(x,y,t-s) u_2(s,y)\,ds\,dy, u_1(t) - \frac{r_1}{a_1} \right\rangle dt \right| \\
&\le \int_0^T \left\|u_1(t) - \frac{r_1}{a_1}\right\|_2 \left\| \int_\Omega \int_{-\infty}^t K_1(x,y,t-s) u_2(s,y)\,ds\,dy \right\|_2 dt \\
&\le \sup_{s\le 0} \|u_2(s)\|_2 \sup_{0\le t\le T} \left\|u_1(t) - \frac{r_1}{a_1}\right\|_2 \int_0^\infty s\,k_1(s)\,ds \\
&\quad + \int_0^T \left\|u_1(t) - \frac{r_1}{a_1}\right\|_2 \int_0^t k_1(t-s)\|u_2(s)\|_2\,ds\,dt.
\end{aligned}$$

We now estimate the second term in the above as follows:

$$\begin{aligned}
&\int_0^T \left\| u_1(t) - \frac{r_1}{a_1} \right\|_2 \int_0^t k_1(t-s)\|u_2(s)\|_2 \, ds \, dt \\
&\le \left\| \left\| u_1 - \frac{r_1}{a_1} \right\| \right\|_2 \left(\int_0^T \left(\int_0^t k_1(t-s)\|u_2(s)\|_2 \, ds \right)^2 dt \right)^{1/2} \\
&\le \left\| \left\| u_1 - \frac{r_1}{a_1} \right\| \right\|_2 \left(\int_0^T \left(\underbrace{\int_0^t k_1(t-s)\, ds}_{\le 1} \right) \left(\int_0^t k_1(t-s)\|u_2(s)\|_2^2 \, ds \right) dt \right)^{1/2} \\
&\le \left\| \left\| u_1 - \frac{r_1}{a_1} \right\| \right\|_2 \left(\int_0^T \int_0^t k_1(t-s)\|u_2(s)\|_2^2 \, ds \, dt \right)^{1/2} \\
&= \left\| \left\| u_1 - \frac{r_1}{a_1} \right\| \right\|_2 \left(\int_0^T \|u_2(s)\|_2^2 \underbrace{\int_s^T k_1(t-s)\, dt}_{\le 1} \, ds \right)^{1/2} \\
&\le \left\| \left\| u_1 - \frac{r_1}{a_1} \right\| \right\|_2 \, \| u_2 \|_2,
\end{aligned}$$

where $\| \cdot \|_2$ denotes the norm in $L^2((0,T); L^2(\Omega; R))$, i.e.,

$$\| u \|_2 = \left(\int_0^T \|u(s)\|_2^2 \, ds \right)^{1/2}.$$

Therefore, for any $T > 0$,

$$\begin{aligned}
&\left| \int_0^T \left\langle \int_\Omega \int_{-\infty}^t K_1(x,y,t-s) u_2(s,y) \, ds \, dy, u_1(t) - \frac{r_1}{a_1} \right\rangle dt \right| \\
&\quad \le \sup_{s \le 0} \|u_2(s)\|_2 \sup_{0 \le t \le T} \|u_1(t) - \tfrac{r_1}{a_1}\|_2 \int_0^\infty s \, k_1(s) \, ds + \left\| \left\| u_1 - \frac{r_1}{a_1} \right\| \right\|_2 \| u_2 \|_2.
\end{aligned}$$

In a similar way, we have

$$\begin{aligned}
&\left| \int_0^T \left\langle \int_\Omega \int_{-\infty}^t K_2(x,y,t-s) \left(u_1(s,y) - \frac{r_1}{a_1} \right) ds \, dy, u_2(t) \right\rangle dt \right| \\
&\quad \le \sup_{s \le 0} \left\| u_1(s) - \frac{r_1}{a_1} \right\|_2 \sup_{0 \le t \le T} \|u_2(t)\|_2 \int_0^\infty s \, k_2(s) \, ds + \left\| \left\| u_1 - \frac{r_1}{a_1} \right\| \right\|_2 \| u_2 \|_2.
\end{aligned} \tag{2.7}$$

Integrating (2.4) over $[0,T]$ and noting that $\sup_{0 \le t \le T} \|u_2(t)\|_2$ and $\sup_{0 \le t \le T} \|u_1(t) - \frac{r_1}{a_1}\|_2$ can be bounded independently of T (by (1.7)), we obtain that there exists a positive constant C independent of T such that

$$\frac{\alpha d_1 r_1}{a_1} \left\| \left\| \frac{\nabla u_1}{u_1} \right\| \right\|_2^2 + \alpha a_1 \left\| \left\| u_1 - \frac{r_1}{a_1} \right\| \right\|_2^2 + a_2 \| u_2 \|_2^2 \le C + (\alpha b_1 + b_2) \left\| \left\| u_1 - \frac{r_1}{a_1} \right\| \right\|_2 \| u_2 \|_2 \tag{2.8}$$

or, by using Young's inequality,

$$\begin{aligned}(2.9)\qquad &\frac{\alpha d_1 r_1}{a_1}\left|\!\left|\!\left|\frac{\nabla u_1}{u_1}\right|\!\right|\!\right|_2^2+\alpha a_1\left|\!\left|\!\left|u_1-\frac{r_1}{a_1}\right|\!\right|\!\right|_2^2+a_2\left\|u_2\right\|_2^2\\ &\le C+(\alpha b_1+b_2)\left(\frac{1}{2}\lambda\left|\!\left|\!\left|u_1-\frac{r_1}{a_1}\right|\!\right|\!\right|_2^2+\frac{1}{2\lambda}\left\|u_2\right\|_2^2\right)\end{aligned}$$

for any $\lambda>0$. If we choose

$$\lambda=\frac{\alpha b_1+b_2}{2a_2},$$

then (2.9) reads as

$$(2.10)\qquad \frac{\alpha d_1 r_1}{a_1}\left|\!\left|\!\left|\frac{\nabla u_1}{u_1}\right|\!\right|\!\right|_2^2+\alpha a_1\left|\!\left|\!\left|u_1-\frac{r_1}{a_1}\right|\!\right|\!\right|_2^2\le C+\frac{(\alpha b_1+b_2)^2}{4a_2}\left|\!\left|\!\left|u_1-\frac{r_1}{a_1}\right|\!\right|\!\right|_2^2.$$

From (2.10) we can conclude that

$$(2.11)\qquad \left|\!\left|\!\left|\frac{\nabla u_1}{u_1}\right|\!\right|\!\right|_2\le C_1$$

and

$$(2.12)\qquad \left|\!\left|\!\left|u_1-\frac{r_1}{a_1}\right|\!\right|\!\right|_2\le C_2$$

for some constants C_1, C_2 independent of T, provided that $\alpha>0$ can be chosen such that

$$2\sqrt{\alpha a_1 a_2}>\alpha b_1+b_2,$$

which is possible by the assumption $a_1a_2>b_1b_2$.

Because of (1.7) we may deduce from (2.11) that, for some constant C_3 independent of T,

$$(2.13)\qquad \left\|\nabla u_1\right\|_2\le C_3.$$

Since all this is for any $T>0$, (2.13) and (2.12) imply that $u_1-r_1/a_1\in L^2((0,\infty);W^{1,2}(\Omega;R))$ and thus

$$(2.14)\qquad \lim_{t\to\infty}\left\|u_1(t)-\frac{r_1}{a_1}\right\|_{W^{1,2}}=0.$$

Therefore,

$$\lim_{t\to\infty}\left\|u_1(t)-\frac{r_1}{a_1}\right\|_{C(\bar\Omega;R)}=0.$$

We deduce $\lim_{t\to\infty}\|u_2(t)\|_{C(\bar\Omega;R)}=0$ in a similar way (for example, λ in (2.9) would be chosen differently). This completes the proof.

Remark 2.4. Theorem 2.3 indicates that if $r_1/r_2>a_1/b_2>b_1/a_2$, then the competitor with density u_1 wins the competition; if $r_1/r_2<b_1/a_2<a_1/b_2$, then the

competitor with density u_2 overcompetes the one with density u_1; and if $b_1/a_2 < r_1/r_2 < a_1/b_2$, then the two competing species coexist in the sense of existence and stability of a positive steady state. Theorem 2.3 extends Propositions 7.5–7.7 in Ruan and Wu [11] on competition-diffusion systems with infinite time delays, Theorem 3.1 in Zhou and Pao [18] on competition-diffusion systems, and the results in Gopalsamy [6] on competition systems with finite delays.

Remark 2.5. It is known (see Yamada [16]) that in the case of the single-species delay equation

$$\frac{\partial u}{\partial t} = \Delta u + u\left(a - bu - \int_{-\infty}^{t} f(t-s)u(x,s)\,ds\right)$$

on homogeneous Neumann boundary conditions, where a and b are nonnegative constants, bifurcations can occur from the nonzero homogeneous equilibrium state for certain kernels and for suitable values of the parameters a and b, which include the requirement that b be sufficiently small. However, in the competition model (1.1), bifurcations to spatially patterned or to spatiotemporal structures are not expected to occur from the equilibrium E^*, given by (1.8). Let us explain why this is so. A standard linearized analysis about the boundary equilibrium $(r_1/a_1, 0)$ shows that, regardless of the delay kernels, this equilibrium is unstable to perturbations in which $u_2 > 0$ if $r_1/r_2 < a_1/b_2$. Similarly, the equilibrium $(0, r_2/a_2)$ is unstable to perturbations in which $u_1 > 0$ if $r_1/r_2 > b_1/a_2$. Now, if the interior equilibrium E^* were to lose stability and bifurcate to a spatial or spatiotemporal structure, we would expect that both boundary equilibria would remain unstable throughout this process so that they act as repellers. Yet the conditions for both boundary equilibria to be linearly unstable can be summarized as

$$\frac{b_1}{a_2} < \frac{r_1}{r_2} < \frac{a_1}{b_2},$$

which is precisely the condition for global convergence to E^* given in (iii) of Theorem 2.3. Hence, bifurcations from E^* cannot occur if the boundary equilibria are to remain unstable.

Remark 2.6. If we assume that, in the absence of the other competitor, each competitor's growth is governed by a Volterra integrodifferential equation with both instantaneous and delay self-regulatory terms (see Cushing [2], Schiaffino and Tesei [13], and Yamada [17]), then we have a more general model of the following form:

$$\begin{aligned}
\frac{\partial u_1}{\partial t} &= d_1\Delta u_1 + u_1\Bigg(r_1 - a_1u_1 - c_1\int_\Omega\int_{-\infty}^{t} H_1(x,y,t-s)u_2(s,y)\,ds\,dy \\
&\qquad -b_1\int_\Omega\int_{-\infty}^{t} K_1(x,y,t-s)u_2(s,y)\,ds\,dy\Bigg), \\
\frac{\partial u_2}{\partial t} &= d_2\Delta u_2 + u_2\Bigg(r_2 - b_2\int_\Omega\int_{-\infty}^{t} K_2(x,y,t-s)u_1(s,y)\,ds\,dy \\
&\qquad -a_2u_2 - c_2\int_\Omega\int_{-\infty}^{t} H_2(x,y,t-s)u_1(s,y)\,ds\,dy\Bigg),
\end{aligned} \tag{2.15}$$

where $a_i \geq 0$, $b_i > 0$, $c_i \geq 0$, $i = 1, 2$, are constants and the kernels H_i, $i = 1, 2$, satisfy similar properties as the K_i of (1.1). Notice that system (1.1) is a special case of (2.15) with $c_i = 0$. When $a_i = 0$, even when there is no diffusion (i.e., the ordinary delay competition model), both stability (see Gomatam and MacDonald [5])

and bifurcation (see Gopalsamy and Aggarwala [7]) are possible. We anticipate that system (2.15) will exhibit more complex dynamics, such as Hopf bifurcations, and we leave this for future consideration.

3. Travelling front solutions. In this section we discuss the modifications necessary to system (1.1) for the case of an infinite one-dimensional domain $\Omega = (-\infty, \infty)$, and travelling front solutions of the resulting system. The infinite domain case is in some respects slightly simpler from a modelling point of view since there are no boundaries for individuals to interact with as they drift from their past to their present positions. Because of this, in contrast to the finite domain case, the nonlocal averaging associated with the delay takes the form of a spatial convolution, so that the model assumes the form

$$\begin{aligned}\frac{\partial u_1}{\partial t} &= d_1\frac{\partial^2 u_1}{\partial x^2} + u_1\left(r_1 - a_1u_1 - b_1\int_{-\infty}^{\infty}\int_{-\infty}^{t} G_1(x-y,t-s)k_1(t-s)u_2(s,y)\,ds\,dy\right),\\ \frac{\partial u_2}{\partial t} &= d_2\frac{\partial^2 u_2}{\partial x^2} + u_2\left(r_2 - b_2\int_{-\infty}^{\infty}\int_{-\infty}^{t} G_2(x-y,t-s)k_2(t-s)u_1(s,y)\,ds\,dy - a_2u_2\right),\end{aligned} \tag{3.1}$$

where the k_i satisfy $\int_0^\infty k_i(s)\,ds = 1$, $i = 1, 2$, and the G_i satisfy diffusion equations as in Lemma 2.1 but without the boundary conditions. To be more precise, G_1 is a weighting function describing the distribution at past times of the individuals of the species u_2 who are at position x at time t. The u_2 individuals diffuse at diffusivity d_2; thus G_1 must satisfy

$$\frac{\partial G_1}{\partial t} = d_2\frac{\partial^2 G_1}{\partial x^2}, \qquad G_1(x,0) = \delta(x),$$

and similarly, G_2 satisfies

$$\frac{\partial G_2}{\partial t} = d_1\frac{\partial^2 G_2}{\partial x^2}, \qquad G_2(x,0) = \delta(x),$$

so that G_1, G_2 are both fundamental solutions of heat equations. With these assumptions, system (3.1) still preserves the same equilibria E_0, E_1, E_2, and (possibly) E^* enumerated earlier.

Our interest in this section is in the possibility of a transition between the boundary equilibria E_1 and E_2 in the form of a travelling wave-front solution. This is of ecological interest since it corresponds to a situation where an environment is initially inhabited only by the weaker of the two competitors at its carrying capacity, and some of the stronger competitor are introduced and then invade the domain, dominate, and drive the weaker to extinction so that the end result is that only the stronger species is present, at its carrying capacity.

In this section the assumptions we shall make on the parameters are those which ensure that the corresponding system without diffusion and without delay (removal of delay can be effected by setting each $k_i(t) = \delta(t)$ in (3.1)) has E_1 unstable and E_2 asymptotically stable. Elementary analysis shows that the conditions for this to happen are

$$r_1b_2 < r_2a_1 \quad \text{and} \quad r_1a_2 < r_2b_1. \tag{3.2}$$

Note that if (3.2) is satisfied, then the coexistence equilibrium E^* is absent. In the two-dimensional (u_1, u_2) phase plane, the diffusionless undelayed ODEs possess a heteroclinic connection from E_1 to E_2. It is known from the papers referred to in the introduction that under these circumstances the (undelayed) reaction-diffusion system has travelling-front solutions connecting these equilibria. Our intention now is to prove, for certain choices of the kernels k_i, that these travelling fronts persist under the introduction of delay, at least for small delays.

We shall consider the situation when the kernels k_i are given by

$$k_1(t) = \frac{1}{\tau_1} e^{-t/\tau_1}, \qquad k_2(t) = \frac{1}{\tau_2} e^{-t/\tau_2}, \tag{3.3}$$

where the delays $\tau_1, \tau_2 > 0$, and we shall prove the following.

THEOREM 3.1. *Let k_1 and k_2 be given by* (3.3) *and assume that* (3.2) *holds. Then, for sufficiently small delays τ_1, τ_2, system* (3.1) *possesses travelling front solutions connecting the semitrivial equilibria $E_1 = (r_1/a_1, 0)$ and $E_2 = (0, r_2/a_2)$.*

Proof. With the kernels given by (3.3), it is straightforward to see that system (3.1) is equivalent to

$$\begin{aligned}
\frac{\partial u_1}{\partial t} &= d_1 \frac{\partial^2 u_1}{\partial x^2} + u_1(r_1 - a_1 u_1 - b_1 w_1), \\
\frac{\partial u_2}{\partial t} &= d_2 \frac{\partial^2 u_2}{\partial x^2} + u_2(r_2 - b_2 w_2 - a_2 u_2), \\
\frac{\partial w_1}{\partial t} &= d_2 \frac{\partial^2 w_1}{\partial x^2} + \frac{1}{\tau_1} u_2 - \frac{1}{\tau_1} w_1, \\
\frac{\partial w_2}{\partial t} &= d_1 \frac{\partial^2 w_2}{\partial x^2} + \frac{1}{\tau_2} u_1 - \frac{1}{\tau_2} w_2.
\end{aligned} \tag{3.4}$$

Converting to travelling wave form, by writing

$$u_1(t, x) = u_1(z), \qquad z = x + ct,$$

and similarly for the other state variables, gives

$$\begin{aligned}
c u_1' &= d_1 u_1'' + u_1(r_1 - a_1 u_1 - b_1 w_1), \\
c u_2' &= d_2 u_2'' + u_2(r_2 - b_2 w_2 - a_2 u_2), \\
c w_1' &= d_2 w_1'' + \frac{1}{\tau_1} u_2 - \frac{1}{\tau_1} w_1, \\
c w_2' &= d_1 w_2'' + \frac{1}{\tau_2} u_1 - \frac{1}{\tau_2} w_2,
\end{aligned}$$

where prime denotes differentiation with respect to z. Let us introduce

$$v_1 = d_1 u_1', \qquad v_2 = d_2 u_2', \qquad v_3 = d_2 w_1', \qquad v_4 = d_1 w_2'.$$

Also, we shall replace τ_1 and τ_2 with $\varepsilon^2 \tau_1$ and $\varepsilon^2 \tau_2$, respectively, since we are interested

in the situation when the delays are small. The system becomes

$$
\begin{aligned}
u_1' &= \frac{1}{d_1} v_1, \\
v_1' &= \frac{c}{d_1} v_1 - u_1(r_1 - a_1 u_1 - b_1 w_1), \\
u_2' &= \frac{1}{d_2} v_2, \\
v_2' &= \frac{c}{d_2} v_2 - u_2(r_2 - b_2 w_2 - a_2 u_2), \\
w_1' &= \frac{1}{d_2} v_3, \\
\varepsilon^2 v_3' &= \frac{\varepsilon^2 c}{d_2} v_3 - \frac{1}{\tau_1} u_2 + \frac{1}{\tau_1} w_1, \\
w_2' &= \frac{1}{d_1} v_4, \\
\varepsilon^2 v_4' &= \frac{\varepsilon^2 c}{d_1} v_4 - \frac{1}{\tau_2} u_1 + \frac{1}{\tau_2} w_2.
\end{aligned}
\tag{3.5}
$$

If we introduce the new state variables

$$\tilde{u}_1 = u_1, \quad \tilde{v}_1 = v_1, \quad \tilde{u}_2 = u_2, \quad \tilde{v}_2 = v_2, \quad \tilde{w}_1 = w_1, \quad \tilde{v}_3 = \varepsilon v_3, \quad \tilde{w}_2 = w_2, \quad \tilde{v}_4 = \varepsilon v_4$$

and then drop the tildes, we have

$$
\begin{aligned}
u_1' &= \frac{1}{d_1} v_1, \\
v_1' &= \frac{c}{d_1} v_1 - u_1(r_1 - a_1 u_1 - b_1 w_1), \\
u_2' &= \frac{1}{d_2} v_2, \\
v_2' &= \frac{c}{d_2} v_2 - u_2(r_2 - b_2 w_2 - a_2 u_2), \\
\varepsilon w_1' &= \frac{1}{d_2} v_3, \\
\varepsilon v_3' &= \frac{\varepsilon c}{d_2} v_3 - \frac{1}{\tau_1} u_2 + \frac{1}{\tau_1} w_1, \\
\varepsilon w_2' &= \frac{1}{d_1} v_4, \\
\varepsilon v_4' &= \frac{\varepsilon c}{d_1} v_4 - \frac{1}{\tau_2} u_1 + \frac{1}{\tau_2} w_2.
\end{aligned}
\tag{3.6}
$$

When $\varepsilon = 0$, system (3.6) reduces to the equations satisfied by travelling wave solutions of the undelayed problem studied by previous investigators [1, 4, 10, 14]. In this degenerate case the system is four-dimensional, but for $\varepsilon > 0$ (i.e., delay is present), existence of a travelling front solution of (3.1) between E_1 and E_2 is equivalent to existence of a heteroclinic connection between the equilibrium points of the eight-dimensional system (3.6) that correspond to E_1 and E_2 of (3.1). We shall still denote

these equilibria by E_1 and E_2; for system (3.6) they are given by

$$E_1 = \left(\frac{r_1}{a_1}, 0, 0, 0, 0, 0, \frac{r_1}{a_1}, 0\right), \qquad E_2 = \left(0, 0, \frac{r_2}{a_2}, 0, \frac{r_2}{a_2}, 0, 0, 0\right). \tag{3.7}$$

Our intention is to apply the geometric singular perturbation theory described in [3], in particular, Theorem 9.1 of that paper. System (3.6) above will henceforth be referred to as the *slow system*. By introducing a new independent variable η defined by

$$z = \varepsilon\eta,$$

system (3.6) transforms into

$$\begin{aligned}
\dot{u}_1 &= \frac{\varepsilon}{d_1} v_1, \\
\dot{v}_1 &= \varepsilon\left(\frac{c}{d_1} v_1 - u_1(r_1 - a_1 u_1 - b_1 w_1)\right), \\
\dot{u}_2 &= \frac{\varepsilon}{d_2} v_2, \\
\dot{v}_2 &= \varepsilon\left(\frac{c}{d_2} v_2 - u_2(r_2 - b_2 w_2 - a_2 u_2)\right), \\
\dot{w}_1 &= \frac{1}{d_2} v_3, \\
\dot{v}_3 &= \frac{\varepsilon c}{d_2} v_3 - \frac{1}{\tau_1} u_2 + \frac{1}{\tau_1} w_1, \\
\dot{w}_2 &= \frac{1}{d_1} v_4, \\
\dot{v}_4 &= \frac{\varepsilon c}{d_1} v_4 - \frac{1}{\tau_2} u_1 + \frac{1}{\tau_2} w_2,
\end{aligned} \tag{3.8}$$

where dots denote differentiation with respect to η. System (3.8) is called the *fast system*. Geometric singular perturbation theory uses both the slow and the fast systems. The two are equivalent when $\varepsilon > 0$, but when $\varepsilon = 0$, the slow system (3.6) does not define a dynamical system in the whole of $\mathbf{R}^8$ but rather the dynamics takes place only on

$$M_0 = \{(u_1, v_1, u_2, v_2, w_1, v_3, w_2, v_4) \in \mathbf{R}^8 : \quad v_3 = 0, \quad v_4 = 0, \quad w_1 = u_2, \quad w_2 = u_1\}, \tag{3.9}$$

which is a four-dimensional submanifold of $\mathbf{R}^8$. Note that M_0 consists of the equilibria of the fast system when $\varepsilon = 0$. If M_0 is normally hyperbolic then, for sufficiently small $\varepsilon > 0$, Theorem 9.1 in [3] provides us with a four-dimensional invariant manifold M_ε for the system (3.6). It will be shown that the equilibrium points E_1 and E_2 lie on M_ε. By studying the system (3.6) reduced to this manifold, the dimensionality is reduced back to four and the existence of the heteroclinic connection we are seeking can be established.

To verify normal hyperbolicity it is necessary to use the fast system (3.8). We need to verify that the linearization of (3.8), restricted to M_0, has exactly four ($= \dim M_0$)

eigenvalues on the imaginary axis with the remainder of the spectrum hyperbolic. The linearization of the fast system, when $\varepsilon = 0$, is given by

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \dfrac{1}{d_2} & 0 & 0 \\ 0 & 0 & -\dfrac{1}{\tau_1} & 0 & \dfrac{1}{\tau_1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dfrac{1}{d_1} \\ -\dfrac{1}{\tau_2} & 0 & 0 & 0 & 0 & 0 & \dfrac{1}{\tau_2} & 0 \end{pmatrix},$$

which has eigenvalues $\{0, 0, 0, 0, \pm 1/\sqrt{\tau_2 d_1}, \pm 1/\sqrt{\tau_1 d_2}\}$. Thus, normal hyperbolicity is verified and there exists an invariant manifold M_ε, close to M_0, for the perturbed system (3.6) for $\varepsilon > 0$ sufficiently small. In fact, M_ε can be expressed in the form

$$\begin{aligned} M_\varepsilon = \big\{(u_1, v_1, u_2, v_2, w_1, v_3, w_2, v_4) \in \mathbf{R}^8 : \quad & v_3 = h_1(u_1, v_1, u_2, v_2; \varepsilon), \\ v_4 = h_2(u_1, v_1, u_2, v_2; \varepsilon), \quad & w_1 = u_2 + h_3(u_1, v_1, u_2, v_2; \varepsilon), \\ & w_2 = u_1 + h_4(u_1, v_1, u_2, v_2; \varepsilon)\big\} \end{aligned} \tag{3.10}$$

with $h_i(u_1, v_1, u_2, v_2; 0) = 0$ for $i = 1, 2, 3, 4$. The h_i can be computed by substitution into (3.6).

The slow system (3.6), restricted to M_ε, is

$$\begin{aligned} u_1' &= \frac{1}{d_1} v_1, \\ v_1' &= \frac{c}{d_1} v_1 - u_1(r_1 - a_1 u_1 - b_1 u_2 - b_1\, h_3(u_1, v_1, u_2, v_2; \varepsilon)), \\ u_2' &= \frac{1}{d_2} v_2, \\ v_2' &= \frac{c}{d_2} v_2 - u_2(r_2 - a_2 u_2 - b_2 u_1 - b_2\, h_4(u_1, v_1, u_2, v_2; \varepsilon)). \end{aligned} \tag{3.11}$$

When $\varepsilon = 0$ (i.e., no delay), system (3.11) again reduces to the system satisfied by travelling wave solutions of the undelayed equations, which has been studied previously. What we now claim is that, for $\varepsilon > 0$ sufficiently small, system (3.11) still possesses as equilibrium points

$$E_1 = \left(\frac{r_1}{a_1}, 0, 0, 0\right), \qquad E_2 = \left(0, 0, \frac{r_2}{a_2}, 0\right) \tag{3.12}$$

and that it falls within the class of systems studied by Gardner [4]. Neither is immediately clear. Indeed, Gardner studied competition systems of the form

$$\begin{aligned} \partial u_1/\partial t &= d_1 \partial^2 u_1/\partial x^2 + u_1 M(u_1, u_2), \\ \partial u_2/\partial t &= d_2 \partial^2 u_2/\partial x^2 + u_2 N(u_1, u_2), \end{aligned}$$

which, in travelling wave form, read as

$$
\begin{aligned}
u_1' &= \frac{1}{d_1} v_1, \\
v_1' &= \frac{c}{d_1} v_1 - u_1 M(u_1, u_2), \\
u_2' &= \frac{1}{d_2} v_2, \\
v_2' &= \frac{c}{d_2} v_2 - u_2 N(u_1, u_2).
\end{aligned} \tag{3.13}
$$

Comparing (3.11) with (3.13) we see that, for Gardner's results to be applicable, the functions h_3 and h_4 in (3.11) would need to involve u_1 and u_2 only. We shall now show that this is indeed the case, up to order ε^2.

Indeed, straightforward but tedious calculations, utilizing the fact that M_ε is an invariant manifold for (3.6), yield that the h_i satisfy

$$
\begin{aligned}
&\varepsilon \left[\frac{1}{d_2} v_2 + \frac{1}{d_1} v_1 \frac{\partial h_3}{\partial u_1} + \frac{\partial h_3}{\partial v_1} \left(\frac{c}{d_1} v_1 - u_1 (r_1 - a_1 u_1 - b_1 u_2 - b_1 h_3) \right) \right. \\
&\left. + \frac{1}{d_2} v_2 \frac{\partial h_3}{\partial u_2} + \frac{\partial h_3}{\partial v_2} \left(\frac{c}{d_2} v_2 - u_2 (r_2 - a_2 u_2 - b_2 u_1 - b_2 h_4) \right) \right] = \frac{1}{d_2} h_1
\end{aligned}
$$

together with three other similar equations. Attempting solutions of the equations in the form

$$
h_1(u_1, v_1, u_2, v_2; \varepsilon) = \varepsilon h_1^{(1)}(u_1, v_1, u_2, v_2) + \varepsilon^2 h_1^{(2)}(u_1, v_1, u_2, v_2) + \cdots,
$$

and similarly for the other h_i, yields, after some further algebra, that

$$
h_1^{(1)} = v_2, \quad h_2^{(1)} = v_1, \quad h_3^{(1)} = 0, \quad h_4^{(1)} = 0
$$

and

$$
\begin{aligned}
h_1^{(2)} = 0, \quad h_2^{(2)} = 0, \\
h_3^{(2)} = -\tau_1 u_2 (r_2 - a_2 u_2 - b_2 u_1), \\
h_4^{(2)} = -\tau_2 u_1 (r_1 - a_1 u_1 - b_1 u_2).
\end{aligned}
$$

Thus, system (3.11) becomes, to order ε^2,

$$
\begin{aligned}
u_1' &= \frac{1}{d_1} v_1, \\
v_1' &= \frac{c}{d_1} v_1 - u_1 (r_1 - a_1 u_1 - b_1 u_2 + \varepsilon^2 b_1 \tau_1 u_2 (r_2 - a_2 u_2 - b_2 u_1)), \\
u_2' &= \frac{1}{d_2} v_2, \\
v_2' &= \frac{c}{d_2} v_2 - u_2 (r_2 - a_2 u_2 - b_2 u_1 + \varepsilon^2 b_2 \tau_2 u_1 (r_1 - a_1 u_1 - b_1 u_2)),
\end{aligned} \tag{3.14}
$$

which has the structure of the system (3.13). Also, note that E_1 and E_2, given by (3.12), are indeed equilibria of (3.14). Therefore, the results in [4] are applicable,

yielding a heteroclinic connection between the equilibria E_1 and E_2 of (3.14). We have shown that travelling fronts exist for system (3.1) when the kernels k_1 and k_2 are given by (3.3), so the proof of Theorem 3.1 is complete.

Remark 3.2. Let us briefly discuss the role of the terms of order ε^2. If the system (3.14) is linearized about the equilibrium E_2, we find that the eigenvalues λ of the linearization satisfy an equation that does not involve ε, namely

$$(d_2\lambda^2 - c\lambda - r_2)(d_1a_2\lambda^2 - ca_2\lambda + r_1a_2 - r_2b_1) = 0. \tag{3.15}$$

About the equilibrium E_1, the eigenvalue equation becomes

$$(d_1\lambda^2 - c\lambda - r_1)(d_2a_1\lambda^2 - ca_1\lambda + r_2a_1 - r_1b_2) = 0, \tag{3.16}$$

which again does not involve ε. These observations suggest that the manner in which the front approaches the equilibria E_1 and E_2 as $z \to -\infty$ and $z \to \infty$, respectively, is independent of ε for $\varepsilon > 0$ sufficiently small and therefore that the front's qualitative profile is not sensitive to the delays, provided they are both sufficiently small. Of course, system (3.14) is itself the result of a perturbation analysis for small ε, and therefore no conclusions can be drawn for larger ε. In conclusion, we may state that the travelling front solutions of the corresponding undelayed competition model appear to be very *robust*, not only in the sense that they persist under the introduction of delays, but also in that they are not sensitive to small delays in the sense that, if the delays are small, they look qualitatively the same as they do with no delay at all.

REFERENCES

[1] C. Conley and R. Gardner, *An application of the generalized Morse index to travelling wave solutions of a competitive reaction-diffusion model,* Indiana Univ. Math. J., 33 (1984), pp. 319–343.

[2] J. M. Cushing, *Integrodifferential Equations and Delay Models in Population Dynamics*, Springer-Verlag, Heidelberg, 1977.

[3] N. Fenichel, *Geometric singular perturbation theory for ordinary differential equations,* J. Differential Equations, 31 (1979), pp. 53–98.

[4] R. Gardner, *Existence and stability of travelling wave solutions of competition models: A degree theoretic approach,* J. Differential Equations, 44 (1982), pp. 343–364.

[5] J. Gomatam and N. MacDonald, *Time delays and stability of two competing species,* Math. Biosci., 24 (1975), pp. 247–255.

[6] K. Gopalsamy, *Time lags and global stability in two-species competition,* Bull. Math. Biol., 42 (1980), pp. 729–737.

[7] K. Gopalsamy and B. D. Aggarwala, *Recurrence and non-stationary coexistence in two-species competition,* Ecol. Model., 9 (1980), pp. 153–163.

[8] S. A. Gourley and N. F. Britton, *A predator prey reaction diffusion system with nonlocal effects,* J. Math. Biol., 34 (1996), pp. 297–333.

[9] S. A. Gourley and J. W.-H. So, *Dynamics of a food-limited population model incorporating nonlocal delays on a finite domain,* J. Math. Biol., 44 (2002), pp. 49–78.

[10] Y. Kan-on, *Parameter dependence of propagation speed of travelling waves for competition-diffusion equations,* SIAM J. Math. Anal., 26 (1995), pp. 340–363.

[11] S. Ruan and J. Wu, *Reaction-diffusion systems with infinite delay,* Canad. Appl. Math. Quart., 2 (1994), pp. 485–550.

[12] S. Ruan and X.-Q. Zhao, *Persistence and extinction in two species reaction-diffusion systems with delays,* J. Differential Equations, 156 (1999), pp. 71–92.

[13] A. Schiaffino and A. Tesei, *Competition systems with Dirichlet boundary conditions,* J. Math. Biol., 15 (1982), pp. 93–105.

[14] M. M. Tang and P. C. Fife, *Propagating fronts for competing species equations with diffusion,* Arch. Ration. Mech. Anal., 73 (1980), pp. 69–77.

[15] J. Wu, *Theory and Applications of Partial Functional Differential Equations*, Springer-Verlag, New York, 1996.

[16] Y. Yamada, *On a certain class of semilinear Volterra diffusion equations*, J. Math. Anal. Appl., 88 (1982), pp. 433–451.
[17] Y. Yamada, *Asymptotic stability for some systems of semilinear Volterra diffusion equations,* J. Differential Equations, 52 (1984), pp. 295–326.
[18] L. Zhou and C. V. Pao, *Asymptotic behavior of a competition-diffusion system in population dynamics,* Nonlinear Anal., 6 (1982), pp. 1163–1184.

SIAM J. MATH. ANAL.
Vol. 35, No. 4, pp. 823–843

NONLINEAR SCHRÖDINGER EQUATIONS WITH REPULSIVE HARMONIC POTENTIAL AND APPLICATIONS*

RÉMI CARLES†

Abstract. We study the Cauchy problem for Schrödinger equations with repulsive quadratic potential and powerlike nonlinearity. The local problem is well-posed in the same space as that used when a confining harmonic potential is involved. For a defocusing nonlinearity, it is globally well-posed, and a scattering theory is available, with no long range effect for any superlinear nonlinearity. When the nonlinearity is focusing, we prove that choosing the harmonic potential sufficiently strong prevents blow-up in finite time. Thanks to quadratic potentials, we provide a method to anticipate, delay, or prevent wave collapse; this mechanism is explicit for critical nonlinearity.

Key words. nonlinear Schrödinger equation, finite-time blow-up, scattering

AMS subject classifications. Primary, 35Q55; Secondary, 35B05, 35B40, 35P25

DOI. 10.1137/S0036141002416936

1. Introduction.

Consider the Schrödinger equation

$$i\partial_t u + \frac{1}{2}\Delta u = V(x)u + \lambda|u|^{2\sigma}u, \quad (t,x) \in \mathbb{R}\times\mathbb{R}^n, \tag{1.1}$$

with $\sigma > 0$, $\sigma < 2/(n-2)$ if $n \geq 3$, $\lambda \in \mathbb{R}$, and V being a real-valued potential $V : \mathbb{R}^n \to \mathbb{R}$. If $V \in L^\infty + L^p$, for some $p \geq 1$, $p > n/2$, then the Cauchy problem in $H^1(\mathbb{R}^n)$ associated with (1.1) is known to be locally well-posed; it may also be globally well-posed or lead to blow-up in finite time (see, e.g., [6]).

If the potential is smooth, $V \in C^\infty$, nonnegative, and if its derivatives of order at least two are bounded, then the same holds in the domain of $\sqrt{-\Delta + V}$ (see [18], [6]). When $n = 1$ and V is nonnegative with superquadratic growth, then the fundamental solution for (1.1) with $\lambda = 0$ is nowhere C^1 [25], but smoothing properties make it possible to solve the nonlinear problem (1.1) in some cases [26].

If V is nonpositive, then $-\Delta + V$ is essentially self-adjoint on $C_0^\infty(\mathbb{R}^n)$, provided that there exist some constants a, b such that $V(x) \geq -a|x|^2 - b$ (see [20, p. 199]). If $-V$ has superquadratic growth, then it is not possible to define $e^{-it(-\Delta+V)}$ (see [9, Chap. 13, Sect. 6, Cor. 22]). In this paper, we study the Cauchy problem

$$\begin{cases} i\partial_t u + \dfrac{1}{2}\Delta u = -\omega^2\dfrac{|x|^2}{2}u + \lambda|u|^{2\sigma}u, \quad (t,x) \in \mathbb{R}\times\mathbb{R}^n, \\ u_{|t=0} = u_0, \end{cases} \tag{1.2}$$

with $\omega, \sigma > 0$, $\sigma < 2/(n-2)$ if $n \geq 3$, $\lambda \in \mathbb{R}$, and

$$u_0 \in \Sigma := \left\{ f \in H^1(\mathbb{R}^n) \ ; \ |x|f \in L^2(\mathbb{R}^n) \right\}.$$

The Hilbert space Σ is equipped with the norm

$$\|f\|_\Sigma = \|f\|_{L^2} + \|\nabla f\|_{L^2} + \|xf\|_{L^2}.$$

*Received by the editors October 31, 2002; accepted for publication (in revised form) March 7, 2003; published electronically October 14, 2003.
http://www.siam.org/journals/sima/35-4/41693.html
†MAB, UMR CNRS 5466, Université Bordeaux 1, 351 cours de la Libération, 33 405 Talence cedex, France (carles@math.u-bordeaux.fr).

Another motivation for studying (1.2) lies in the study of finite-time blow-up for the Cauchy problem

$$\begin{cases} i\partial_t u + \dfrac{1}{2}\Delta u = \lambda |u|^{2\sigma} u \ , \quad (t,x) \in \mathbb{R} \times \mathbb{R}^n, \\ u_{|t=0} = u_0. \end{cases} \tag{1.3}$$

It is well known that if $u_0 \in \Sigma$, $\lambda < 0$, $\sigma \geq 2/n$, and

$$E(u_0) := \frac{1}{2}\|\nabla u_0\|_{L^2}^2 + \frac{\lambda}{\sigma+1}\|u_0\|_{L^{2\sigma+2}}^{2\sigma+2} < 0\,, \tag{1.4}$$

then u blows up in finite time; that is, there exists $T > 0$ such that

$$\lim_{t \to T} \|\nabla_x u(t)\|_{L^2} = +\infty \ .$$

This is proven by the general approach of Zakharov–Glassey [12], [6]. There exists numerical evidence suggesting that the introduction of a stochastic white noise in (1.3) may amplify or prevent blow-up formation (see [8]); in this article, we confine ourselves to a deterministic framework.

It is shown in [7] that if the initial datum u_0 is replaced by $u_0(x)e^{-ib|x|^2}$ with $b > 0$ sufficiently large, then the blow-up time is anticipated (and is $O(b^{-1})$). On the other hand, if u_0 is replaced by $u_0(x)e^{ib|x|^2}$ with $b > 0$ sufficiently large, then no blow-up occurs.

Our approach is suggested by the semiclassical régime for the linear Schrödinger equation with potential. Consider the initial value problem

$$\begin{cases} i\varepsilon\partial_t u^\varepsilon + \dfrac{1}{2}\varepsilon^2\Delta u^\varepsilon = V(x)u^\varepsilon, \quad (t,x) \in \mathbb{R} \times \mathbb{R}^n, \\ u^\varepsilon_{|t=0} = u_0^\varepsilon, \end{cases}$$

where $V \in C^\infty(\mathbb{R}^n, \mathbb{R})$, $\varepsilon \in \,]0,1]$. In the semiclassical limit $\varepsilon \to 0$, the energy of the solution u^ε is carried by bicharacteristics, which are the integral curves associated with the classical Hamiltonian

$$p(t,x,\tau,\xi) = \tau + \frac{|\xi|^2}{2} + V(x)\,.$$

If the energy tends to concentrate in this case, one can expect that, for (1.1) with a focusing nonlinearity ($\lambda < 0$), blow-up in finite time, which corresponds to the concentration of the mass, may occur. Bicharacteristic curves solve

$$\begin{cases} \dot t = 1\,, \\ \dot x = \xi\,, \\ \dot\tau = 0\,, \\ \dot\xi = -\nabla V(x)\,. \end{cases}$$

Rays of geometric optics, which are the projection of bicharacteristic curves on (t,x) space, are of the form $x = x(t)$, with

$$\begin{cases} \ddot x + \nabla V(x) = 0 \ , \\ x(0) = x_0 \ , \ \dot x(0) = \xi_0\,. \end{cases} \tag{1.5}$$

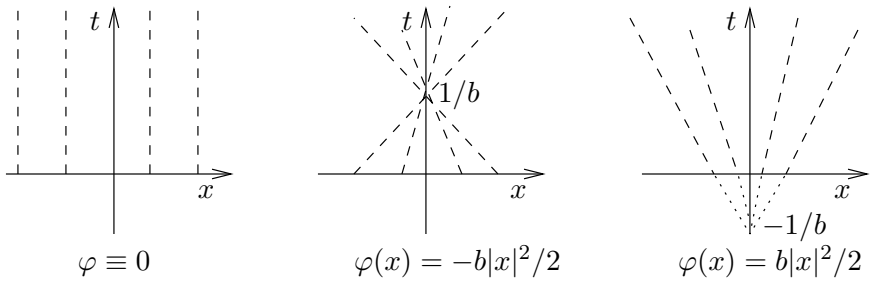


FIG. 1.1. *Geometry of rays: case* $V \equiv 0$.

If the initial datum is of the form $u_0^\varepsilon(x) = f(x)e^{i\varphi(x)/\varepsilon}$, then $\xi_0 = \nabla\varphi(x_0)$. We give three examples which are at the origin of this work and which correspond to cases where (1.5) can easily be solved.

Example 1. Suppose $V \equiv 0$. Then the solutions of (1.5) are

$$x(t) = x_0 + t\nabla\varphi(x_0)\,.$$

If no oscillation is present in the initial datum, the rays are parallel. More interesting is the case of quadratic oscillations (see also [5]). If $\varphi(x) = -b|x|^2/2$ with $b > 0$, then rays are given by $x(t) = x_0(1 - bt)$, and meet at the origin at time $t = 1/b$ (see Figure 1.1). There is focusing, which suggests that in a nonlinear setting such oscillations may cause wave collapse. If $\varphi(x) = b|x|^2/2$ with $b > 0$, then rays are given by $x(t) = x_0(1 + bt)$ and meet at the origin at time $t = -1/b$ (in the past). In particular, they are spread out for positive times, which suggests that in a nonlinear setting such oscillations may prevent wave collapse. This intuition described about the last two cases is confirmed by the results of Cazenave and Weissler [7].

Example 2. Suppose $V(x) = \omega^2|x|^2/2$, with $\omega > 0$. In the case $\varphi \equiv 0$, rays are given by $x(t) = x_0\cos(\omega t)$, and meet at the origin at time $t = \pi/(2\omega)$ (see Figure 1.2). The first example suggests that blow-up may occur more easily than when $V \equiv 0$. This phenomenon is reinforced by the case $\varphi(x) = -\omega\tan(\omega t_0)|x|^2/2$, $|t_0| < \pi/(2\omega)$, where

$$x(t) = \frac{x_0}{\cos(\omega t_0)}\cos\omega(t + t_0)\,.$$

Rays meet at the origin at time $t = \pi/(2\omega) - t_0$. If $t_0 > 0$, focusing is anticipated, while if $t_0 < 0$, it is delayed (but in no case prevented). This geometry can be compared with the second case of the first example.

Example 3. Suppose $V(x) = -\omega^2|x|^2/2$ with $\omega > 0$. In the case $\varphi \equiv 0$, rays are given by $x(t) = x_0\cosh(\omega t)$ and are strongly dispersed for positive times (see Figure 1.2). This geometry is to be compared with the third case of the first example; in that case the rays are scattered but go to infinity exponentially fast, instead of algebraically. If $\varphi(x) = -\omega b|x|^2/2$ with $b > 0$, then the rays are given by $x(t) = x_0(\cosh(\omega t) - b\sinh(\omega t))$, and their behavior is far less singular than in the case $V \equiv 0$. This is a first hint that such potentials may prevent blow-up.

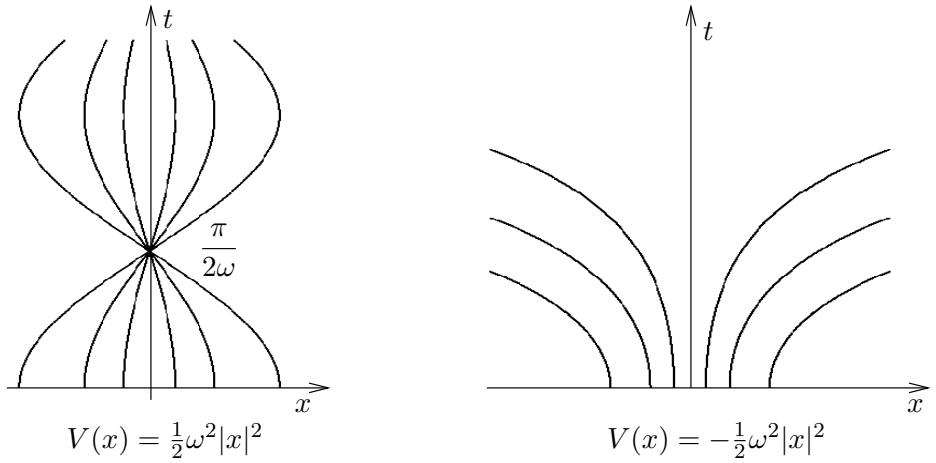


FIG. 1.2. *Geometry of rays, with* $\varphi \equiv 0$.

Inspired by the second example, we proved in [3] that if u solves

$$\begin{cases} i\partial_t u + \dfrac{1}{2}\Delta u = \omega^2 \dfrac{|x|^2}{2} u + \lambda |u|^{2\sigma} u, \quad (t,x) \in \mathbb{R} \times \mathbb{R}^n, \\ u_{|t=0} = u_0, \end{cases} \tag{1.6}$$

then under condition (1.4) (as a matter of fact, the value $E(u_0) = 0$ is allowed), u blows up at time $T \leq \pi/(2\omega)$; choosing ω large enough, the blow-up time is therefore anticipated by the action of a confining magnetic field. This is intimately related to the dynamics of the linear Schrödinger equation with a confining harmonic potential, whose solution is given by Mehler's formula (see, e.g., [10]), for $|t| < \pi/(2\omega)$,

$$u(t,x) = e^{-in\frac{\pi}{4}\operatorname{sgn} t} \left| \frac{\omega}{2\pi \sin \omega t} \right|^{n/2} \int_{\mathbb{R}^n} e^{\frac{i\omega}{\sin \omega t}\left(\frac{x^2+y^2}{2}\cos \omega t - x\cdot y\right)} u_0(y) dy\,. \tag{1.7}$$

At time $t = \pi/(2\omega)$, the fundamental solution is singular. In the nonlinear case, two phenomena cumulate: because of the linear dynamics, the solution tends to focus near the origin as time goes to $\pi/(2\omega)$; when the solution is sufficiently concentrated, the nonlinear term becomes important and causes wave collapse.

Setting $\lambda = 0$ in (1.2), we have the analogue of Mehler's formula,

$$u(t,x) = e^{-in\frac{\pi}{4}\operatorname{sgn} t} \left| \frac{\omega}{2\pi \sinh \omega t} \right|^{n/2} \int_{\mathbb{R}^n} e^{\frac{i\omega}{\sinh \omega t}\left(\frac{x^2+y^2}{2}\cosh \omega t - x\cdot y\right)} u_0(y) dy\,. \tag{1.8}$$

Not only does the kernel of $U_\omega(t) := \exp\{-it/2(-\Delta - \omega^2|x|^2)\}$ given by the above formula have no singularities for $t > 0$, but the dispersive effects are much stronger than in the case with no potential ($\omega = 0$). One might think that in the nonlinear case, the free dynamics can prevent the nonlinear mechanism of blow-up, at least for large ω. In section 4 we prove that this holds true. These results are summarized in the following theorem, whose first point is a consequence of [3].

THEOREM 1.1. *Let* $u_0 \in \Sigma$, $\lambda < 0$, $\sigma \geq 2/n$ *and* $\sigma < 2/(n-2)$ *if* $n \geq 3$.

1. *Assume that the solution* u *to* (1.3) *has negative energy; that is,* (1.4) *holds. Then* u *blows up in finite time* T. *Let* $\omega_0 = \pi/(2T)$; *for any* $\omega \geq \omega_0$, *the solution to* (1.6) *blows up before time* $\pi/(2\omega)$ *and, in particular, before time* T.

2. *If the initial datum* u_0 *satisfies*

$$\frac{1}{2}\|\nabla u_0\|_{L^2}^2+\frac{\lambda}{\sigma+1}\|u_0\|_{L^{2\sigma+2}}^{2\sigma+2}<-\frac{\omega^2}{2}\|xu_0\|_{L^2}^2\,,$$

then the solution to (1.2) *blows up in finite time, in the future or in the past.*

3. *If the initial datum* u_0 *satisfies*

$$\frac{1}{2}\|\nabla u_0\|_{L^2}^2+\frac{\lambda}{\sigma+1}\|u_0\|_{L^{2\sigma+2}}^{2\sigma+2}<-\frac{\omega^2}{2}\|xu_0\|_{L^2}^2-\omega\left|\mathrm{Im}\int\overline{u_0}x\cdot\nabla_x u_0\right|\,,$$

then the solution to (1.2) *blows up in finite time, in the future* and *in the past.*

4. *There exists* $\omega_1>0$ *such that for any* $\omega\geq\omega_1$, (1.2) *has a unique global solution* $u\in C(\mathbb{R},\Sigma)$.

In the particular case $\sigma=2/n$, a change of variables relating the solutions of (1.3) to those of (1.6) or (1.2) shows explicitly how the blow-up for solutions to (1.3) can be anticipated, delayed, or prevented by the introduction of quadratic potentials.

THEOREM 1.2. *Let* $u_0\in\Sigma$, $\lambda<0$, $\sigma=2/n$. *Assume that the solution to* (1.3) *blows up at time* $T>0$.

- *For any* $\omega>0$, *the solution to* (1.6) *blows up at time* $\arctan(\omega T)/\omega<T$.
- *If* $\omega<1/T$, *then the solution to* (1.2) *blows up at time* $\arg\tanh(\omega T)/\omega>T$.
- *If* $\omega\geq 1/T$, *then* (1.2) *has a unique global solution in* $C(\mathbb{R}_+,\Sigma)$.

REMARK 1.3. *As we recall in section* 2 *(see Lemma* 2.2*),* (1.8) *provides Strichartz inequalities that make it possible to study* (1.2) *with* $u_0\in L^2(\mathbb{R}^n)$ *if* $\sigma<2/n$ *or with* $u_0\in L^2(\mathbb{R}^n)$ *and* $\|u_0\|_{L^2}$ *small if* $\sigma=2/n$. *Our goal is precisely to understand the other cases; that is why we shall always assume* $u_0\in\Sigma$.

REMARK 1.4. *Replacing* ω *with* $\pm i\omega$ *formally turns* (1.6) *into* (1.2) *(and* vice versa*) and* (1.7) *into* (1.8). *All the algebraic manipulations we perform in section* 2 *can be retrieved using this argument; in particular,* (2.3), *Lemma* 2.3, *and the evolution law* (2.10) *can be deduced from the formulae established in* [3].

When $\lambda>0$, solutions to (1.3) are known to be global, and the classical issue is to understand their asymptotic behavior as $t\to\pm\infty$. For σ sufficiently large, the solutions are asymptotically free, while for $\sigma\leq 1/n$, a long range scattering theory is needed (see [1], [21], [23], [19], [13], [2]). Notice that, for the initial value problem (1.2), it is not obvious that the formal conservations of mass and energy imply global existence in Σ once local existence is known. These conservations read

$$\|u(t)\|_{L^2}\equiv\|u_0\|_{L^2};\quad\frac{1}{2}\|\nabla_x u(t)\|_{L^2}^2-\frac{\omega^2}{2}\|xu(t)\|_{L^2}^2+\frac{\lambda}{\sigma+1}\|u(t)\|_{L^{2\sigma+2}}^{2\sigma+2}=\mathrm{const}\,.$$

For (1.6), the analogue of these two conservation laws yields global existence in Σ when $\lambda>0$, because the energy is the sum of three positive terms. For (1.2), the energy functional is not always positive, even if $\lambda>0$. We prove in section 2 that a refined analysis of the conservation of energy, consisting in splitting the energy into two parts, yields global existence as soon as $\lambda>0$ (and in other cases). Moreover, the strong dispersive properties of U_ω lead to a different scattering theory for (1.2); the nonlinearity $u\mapsto|u|^{2\sigma}u$ is always short range.

THEOREM 1.5. *Let* $\lambda,\sigma>0$, *with* $\sigma<2/(n-2)$ *if* $n\geq 3$.

1. *For every* $u_-\in\Sigma$, *there exists a unique* $u_0\in\Sigma$ *such that the maximal solution* $u\in C(\mathbb{R},\Sigma)$ *to* (1.2) *satisfies*

$$\|U_\omega(-t)u(t)-u_-\|_\Sigma\underset{t\to-\infty}{\longrightarrow}0.$$

2. *For every $u_0 \in \Sigma$, there exists a unique $u_+ \in \Sigma$ such that the maximal solution $u \in C(\mathbb{R}, \Sigma)$ to (1.2) satisfies*

$$\|U_\omega(-t)u(t) - u_+\|_\Sigma \underset{t\to+\infty}{\longrightarrow} 0.$$

This paper is organized as follows. In section 2, we study the Cauchy problem for (1.2); we prove that it is locally well-posed in Σ and, in some cases, globally well-posed. This is the most important part of the paper. In section 3 we analyze the corresponding scattering theory, and prove Theorem 1.5. Theorem 1.1 is proven in section 4, and Theorem 1.2 is proven in section 5.

2. Solving the Cauchy problem. To solve the local Cauchy problem, we first introduce the classical notions of admissible pairs and Strichartz estimates.

DEFINITION 2.1. *A pair (q,r) is* admissible *if $2 \leq r < \frac{2n}{n-2}$ (resp., $2 \leq r \leq \infty$ if $n = 1$, $2 \leq r < \infty$ if $n = 2$) and*

$$\frac{2}{q} = \delta(r) := n\left(\frac{1}{2} - \frac{1}{r}\right).$$

Recall that $U_\omega(t)$ denotes the semigroup $\exp(it/2(\Delta + \omega^2|x|^2))$, which is given explicitly by (1.8).

LEMMA 2.2 (Strichartz estimates for U_ω). *Let $\omega > 0$.*

1. *For any admissible pair (q,r), there exists C_r independent of $\omega > 0$ such that*

$$\|U_\omega(.)\varphi\|_{L^q(\mathbb{R};L^r)} \leq C_r\|\varphi\|_{L^2} \tag{2.1}$$

for every $\varphi \in L^2(\mathbb{R}^n)$.

2. *For any admissible pairs (q_1,r_1) and (q_2,r_2) and any interval I, there exists C_{r_1,r_2} independent of $\omega > 0$ and I such that*

$$\left\|\int_{I\cap\{s\leq t\}} U_\omega(t-s)F(s)ds\right\|_{L^{q_1}(I;L^{r_1})} \leq C_{r_1,r_2}\left\|F\right\|_{L^{q_2'}(I;L^{r_2'})} \tag{2.2}$$

for every $F \in L^{q_2'}(I;L^{r_2'})$.

Proof. The semigroup U_ω is isometric on $L^2(\mathbb{R}^n)$, and from (1.8) it satisfies, for any $t \neq 0$ and $f \in L^1(\mathbb{R}^n)$,

$$\|U_\omega(t)f\|_{L^\infty} \leq \frac{1}{|2\pi t|^{n/2}}\|f\|_{L^1}.$$

It follows from the results of [15] that U_ω satisfies such Strichartz estimates as stated above. One can choose constants independent of $\omega > 0$, because the above dispersion estimate is independent of $\omega > 0$. One can actually take the same constants as in the case with no potential, $\omega = 0$. □

As mentioned in the introduction, this lemma makes it possible to study (1.2) if $u_0 \in L^2(\mathbb{R}^n)$ and $\sigma < 2/n$ by just mimicking the proof of the corresponding result for (1.3). Since our interest is to study (1.2) when $\sigma \geq 2/n$, in order to analyze finite-time blow-up, we assume $u_0 \in \Sigma$ and introduce two operators:

$$J(t) := \omega x\sinh(\omega t) + i\cosh(\omega t)\nabla_x, \quad H(t) := x\cosh(\omega t) + i\frac{\sinh(\omega t)}{\omega}\nabla_x. \tag{2.3}$$

For $f \in L^2(\mathbb{R}^n)$ and $t \in \mathbb{R}$, the property $J(t)f, H(t)f \in L^2(\mathbb{R}^n)$ implies $f \in \Sigma$, since

$$i\nabla_x = \cosh(\omega t)J(t) - \omega \sinh(\omega t)H(t), \quad x = \cosh(\omega t)H(t) - \frac{\sinh(\omega t)}{\omega}J(t)\,. \tag{2.4}$$

These two operators are the formal analogues of those we used in [3] to study (1.6), when ω is replaced by $\pm i\omega$. They have the remarkable property of being both Heisenberg observables *and* conjugate to ∇_x by a unitary factor.

LEMMA 2.3. *The operators J and H satisfy the following properties.*

1. *They are Heisenberg observables, which read as*

$$J(t) = U_\omega(t) i\nabla_x U_\omega(-t), \quad H(t) = U_\omega(t) x U_\omega(-t), \tag{2.5}$$

and consequently they commute with the linear part of (1.2)*:*

$$\left[i\partial_t + \frac{1}{2}\Delta + \omega^2\frac{|x|^2}{2}, J(t)\right] = \left[i\partial_t + \frac{1}{2}\Delta + \omega^2\frac{|x|^2}{2}, H(t)\right] = 0\,.$$

2. *They can be factored as follows: for $t \neq 0$,*

$$\begin{aligned} J(t) &= i\cosh(\omega t) e^{i\omega\frac{|x|^2}{2}\tanh(\omega t)} \nabla_x \left(e^{-i\omega\frac{|x|^2}{2}\tanh(\omega t)}\ \cdot \right), \\ H(t) &= i\frac{\sinh(\omega t)}{\omega} e^{i\omega\frac{|x|^2}{2}\coth(\omega t)} \nabla_x \left(e^{-i\omega\frac{|x|^2}{2}\coth(\omega t)}\ \cdot \right). \end{aligned} \tag{2.6}$$

3. *They yield modified Gagliardo–Nirenberg inequalities. Recall that if $r \geq 2$, with $r < 2n/(n-2)$ if $n \geq 3$, there exists c_r such that for any $f \in H^1(\mathbb{R}^n)$,*

$$\|f\|_{L^r} \leq c_r \|f\|_{L^2}^{1-\delta(r)} \|\nabla f\|_{L^2}^{\delta(r)}.$$

Then for every $f \in \Sigma$,

$$\begin{aligned} \big\|f\big\|_{L^r} &\leq \frac{c_r}{(\cosh(\omega t))^{\delta(r)}} \|f\|_{L^2}^{1-\delta(r)} \|J(t)f\|_{L^2}^{\delta(r)} \ \forall t \in \mathbb{R}, \\ \big\|f\big\|_{L^r} &\leq c_r \left(\frac{\omega}{\sinh(\omega t)}\right)^{\delta(r)} \|f\|_{L^2}^{1-\delta(r)} \|H(t)f\|_{L^2}^{\delta(r)} \ \forall t \neq 0. \end{aligned} \tag{2.7}$$

4. *They act like derivatives on the nonlinearities $F \in C^1(\mathbb{C},\mathbb{C})$ satisfying the gauge invariance property $F(z) = G(|z|^2)z$ $\forall z \in \mathbb{C}$; that is,*

$$\begin{aligned} J(t)F(u) &= \partial_z F(u) J(t)u - \partial_{\bar z} F(u)\overline{J(t)u}, \\ H(t)F(u) &= \partial_z F(u) H(t)v - \partial_{\bar z} F(u)\overline{H(t)u}. \end{aligned}$$

Proof. The first point is easily checked thanks to (1.8). The second assertion is obvious and implies the last two points. ∎

REMARK 2.4. *One could argue that we consider only isotropic potentials and not the general form*

$$V(x) = \frac{1}{2}\sum_{j=1}^n \delta_j \omega_j^2 x_j^2,$$

with $\delta_j \in \{-1,0,1\}$, $\omega_j > 0$, not necessarily equal. Strichartz estimates would still be available (locally in time only if some δ_j is positive), and one could construct operators

analogous to J and H that satisfy such properties as those stated in Lemma 2.3. *However, the evolution law* (2.10) *stated below (on which our study strongly relies) seems to be bound to isotropic potentials. Finally, the changes of variables we use in section* 5 *are also typical of isotropic potentials.*

Formally, solutions of (1.2) satisfy the following conservation laws:

$$\begin{aligned} &\text{Mass: } M = \|u(t)\|_{L^2} = \text{const} = \|u_0\|_{L^2}\,, \\ &\text{Energy: } E = \frac{1}{2}\|\nabla_x u(t)\|_{L^2}^2 - \frac{\omega^2}{2}\|xu(t)\|_{L^2}^2 + \frac{\lambda}{\sigma+1}\|u(t)\|_{L^{2\sigma+2}}^{2\sigma+2} = \text{const}\,. \end{aligned} \tag{2.8}$$

Notice that even if the nonlinearity is repulsive ($\lambda > 0$), one cannot deduce a priori estimates from the conservation of energy. One needs more precise information. Following [3], split the energy into two parts, which are not conserved in general:

$$\begin{aligned} E_1(t) &:= \frac{1}{2}\|J(t)u\|_{L^2}^2 + \frac{\lambda}{\sigma+1}\cosh^2(\omega t)\|u(t)\|_{L^{2\sigma+2}}^{2\sigma+2}\,, \\ E_2(t) &:= -\frac{\omega^2}{2}\|H(t)u\|_{L^2}^2 - \frac{\lambda}{\sigma+1}\sinh^2(\omega t)\|u(t)\|_{L^{2\sigma+2}}^{2\sigma+2}\,. \end{aligned}$$

Let us notice the identity $E_1(t) + E_2(t) \equiv E$. For (q,r) an admissible pair, define

$$\begin{aligned} Y_{r,\mathrm{loc}}(I) &:= \left\{u \in C(I;\Sigma);\ A(t)u \in L^q_{\mathrm{loc}}(I;L^r) \cap L^\infty_{loc}(I;L^2)\ \forall A \in \{Id,J,H\}\right\}, \\ Y_{\mathrm{loc}}(I) &:= \{u \in C(I;\Sigma);\ A(t)u \in L^q_{\mathrm{loc}}(I;L^r)\ \forall A \in \{Id,J,H\}, \forall (q,r)\ \text{admissible}\}\,. \end{aligned}$$

When (1.2) is globally well-posed, we will also use

$$\begin{aligned} Y_r(I) &:= \left\{u \in C(I;\Sigma);\ A(t)u \in L^q(I;L^r) \cap L^\infty(I;L^2)\ \forall A \in \{Id,J,H\}\right\}, \\ Y(I) &:= \{u \in C(I;\Sigma);\ A(t)u \in L^q(I;L^r)\ \forall A \in \{Id,J,H\}\ \forall (q,r)\ \text{admissible}\}\,. \end{aligned}$$

Proposition 2.5 (local well-posedness for (1.2)). *Let $\lambda \in \mathbb{R}$, $\sigma, \omega > 0$, with $\sigma < 2/(n-2)$ if $n \geq 3$.*

• *For every $u_0 \in \Sigma$, there exist $t_0 > 0$ independent of $\omega > 0$, and a unique solution $u \in Y_{2\sigma+2,loc}(]-2t_0, 2t_0[)$ to* (1.2). *Moreover, it belongs to $Y_{loc}(]-2t_0, 2t_0[)$ and there exists C_0 depending only on λ, n, σ, and $\|u_0\|_\Sigma$ such that*

$$\sup_{|t|\leq t_0}\|u(t)\|_{L^2} + \sup_{|t|\leq t_0}\|J(t)u\|_{L^2} + \sup_{|t|\leq t_0}\|H(t)u\|_{L^2} \leq C_0\ . \tag{2.9}$$

• *Mass and energy are conserved, that is,* (2.8) *holds. More precisely, E_1 and E_2 satisfy*

$$\frac{dE_1}{dt} = -\frac{dE_2}{dt} = \frac{\omega\lambda}{2\sigma+2}(2-n\sigma)\sinh(2\omega t)\|u(t)\|_{L^{2\sigma+2}}^{2\sigma+2}\,. \tag{2.10}$$

• *If $u_0^n \to u_0$ in Σ and $[t_1,t_2] \subset]-2t_0, 2t_0[$, then $u^n \to u$ in $Y([t_1,t_2])$, where u^n solves* (1.2) *with initial datum u_0^n.*

Proof. Notice that Duhamel's principle for (1.2) can be written as

$$u(t) = U_\omega(t)u_0 - i\lambda\int_0^t U_\omega(t-s)\left(|u|^{2\sigma}u\right)(s)ds\,. \tag{2.11}$$

The point is that we can reproduce the proof of local existence of solutions to (1.3) in Σ (see, e.g., [11]). Indeed, Duhamel's principle is similar: we have the same Strichartz

inequalities (from Lemma 2.2, with constants independent of $\omega > 0$), and operators J and H satisfy the same properties as those which are used in the proof of local existence of solutions to (1.3) in Σ. More precisely, the operators used in the case of (1.3) are ∇_x and $x + it\nabla_x$; they commute with the linear part of (1.3) and act on such nonlinearities as those we treat like derivatives, and ∇_x provides Gagliardo–Nirenberg inequalities (as does $x+it\nabla_x$, but this point is not used for local existence). Lemma 2.3 shows that J and H satisfy those properties, and in particular the first line of (2.7) provides the same Gagliardo–Nirenberg inequalities as for ∇_x; since $\cosh(x) \geq 1\ \forall x \in \mathbb{R}$,

$$\|f\|_{L^r} \leq \frac{c_r}{(\cosh(\omega t))^{\delta(r)}} \|f\|_{L^2}^{1-\delta(r)} \|J(t)f\|_{L^2}^{\delta(r)} \leq c_r \|f\|_{L^2}^{1-\delta(r)} \|J(t)f\|_{L^2}^{\delta(r)}. \tag{2.12}$$

Moreover, $J(0) = i\nabla_x$ and $H(0) = x$ are independent of $\omega > 0$, so the first point of the proposition follows.

One can check that the identities stated in the second point hold for smooth solutions. It follows that they hold for the solutions constructed in the first point, by the same argument as in the case of (1.3) (see, e.g., [6, Thm. 4.2.8 and Prop. 6.4.2]). Similarly, transposing the proof of [6, Thm. 4.2.8] yields the last point of the proposition. □

REMARK 2.6. *One might think that the above proposition would also hold for* (1.6)*. In that case, we would have some local existence on a time interval independent of* $\omega > 0$*, which contradicts the result of* [3]*, recalled in the first point of Theorem* 1.1*. The only difference which makes it impossible to conclude as above is that, in the analogue of* (2.7)*, hyperbolic functions are replaced by trigonometric functions; that is, the analogue of* (2.7) *is*

$$\|f\|_{L^r} \leq \frac{c_r}{(\cos(\omega t))^{\delta(r)}} \|f\|_{L^2}^{1-\delta(r)} \|J(t)f\|_{L^2}^{\delta(r)},$$
$$\|f\|_{L^r} \leq c_r \left(\frac{\omega}{\sin(\omega t)}\right)^{\delta(r)} \|f\|_{L^2}^{1-\delta(r)} \|H(t)f\|_{L^2}^{\delta(r)}.$$

We cannot eliminate the dependence upon ω *as we did in* (2.12)*; this prevents the existence of such a* t_0 *independent of* ω*.*

COROLLARY 2.7. *Let* $\lambda \in \mathbb{R}$*,* $\sigma, \omega > 0$*, with* $\sigma < 2/(n-2)$ *if* $n \geq 3$*.*

1. *Let* $u_0 \in \Sigma$ *and* $u \in Y_{loc}(I)$ *solve* (1.2) *for some time interval* I *containing* 0*. For any* $I \ni t > 0$*, the following properties are equivalent:*
   - $\nabla_x u(s)$ *is uniformly bounded in* $L^2(\mathbb{R}^n)$ *for* $s \in [0,t]$*;* $\nabla_x u \in L^\infty([0,t]; L^2)$*.*
   - $J(s)u$ *or* $H(s)u$ *is uniformly bounded in* $L^2(\mathbb{R}^n)$ *for* $s \in [0,t]$*.*
   - $J(s)u$ *and* $H(s)u$ *are uniformly bounded in* $L^2(\mathbb{R}^n)$ *for* $s \in [0,t]$*.*
   - $u(s,\cdot)$ *is uniformly bounded in* Σ *for* $s \in [0,t]$*;* $u \in L^\infty([0,t]; \Sigma)$*.*
2. *For every* $u_0 \in \Sigma$*, there exist* $T^*(u_0), T_*(u_0) > 0$*, and a unique maximal solution* $u \in Y_{2\sigma+2,loc}(]-T_*, T^*[)$ *to* (1.2)*, which actually belongs to* $Y_{loc}(]-T_*, T^*[)$*. It is maximal in the sense that if* $T^*(u_0) < \infty$*, then* $\|\nabla_x u(t)\|_{L^2} \to \infty$ *as* $t \uparrow T^*(u_0)$*, and if* $T_*(u_0) < \infty$*, then* $\|\nabla_x u(t)\|_{L^2} \to \infty$ *as* $t \downarrow -T_*(u_0)$*.*

Proof. First, notice that the equivalence of the last two properties of the first assertion is a consequence of formulae (2.3) and (2.4) and of the conservation of mass (2.8).

Assume that $\nabla_x u(s,\cdot)$ is uniformly bounded in $L^2(\mathbb{R}^n)$ for $s \in [0,t]$. Since $u \in Y_{loc}(I)$ solves (1.2), its L^2-norm is constant; thus $u(s,\cdot)$ is uniformly bounded in $H^1(\mathbb{R}^n)$ for $s \in [0,t]$. From the Sobolev embedding $H^1(\mathbb{R}^n) \subset L^{2\sigma+2}(\mathbb{R}^n)$, $u(s,\cdot)$ is

uniformly bounded in $L^{2\sigma+2}(\mathbb{R}^n)$ for $s \in [0,t]$, and from the conservation of energy (2.8), the first moment of u is uniformly bounded in $L^2(\mathbb{R}^n)$: $u \in L^\infty([0,t];\Sigma)$.

We now have only to prove that the second and third properties are equivalent, that is, that the second implies the third. Assume that $J(s)u$ is uniformly bounded in $L^2(\mathbb{R}^n)$ for $s \in [0,t]$. Then from (2.7), $u(s,\cdot)$ is uniformly bounded in $L^{2\sigma+2}(\mathbb{R}^n)$ for $s \in [0,t]$; $E_1(s)$ is uniformly bounded for $s \in [0,t]$. Since $E_1(s) + E_2(s) \equiv E$, $E_2(s)$ is uniformly bounded for $s \in [0,t]$, which proves that $H(s)u$ is uniformly bounded in $L^2(\mathbb{R}^n)$ for $s \in [0,t]$.

Assume that $H(s)u$ is uniformly bounded in $L^2(\mathbb{R}^n)$ for $s \in [0,t]$. From the first point of Proposition 2.5, $u \in L^\infty([0,t_0];\Sigma)$ for some $t_0 > 0$. We thus suppose that $t \geq t_0$. From (2.7), $u(s,\cdot)$ is uniformly bounded in $L^{2\sigma+2}(\mathbb{R}^n)$ for $s \in [t_0,t]$, and we can repeat the above argument.

The second assertion follows from the first and Proposition 2.5. □

We can now state sufficient conditions for the solution of (1.2) to be global. When $\lambda < 0$, let Q denote the unique spherically symmetric solution of (see [22], [16])

$$\begin{cases} -\frac{1}{2}\Delta Q + Q = -\lambda |Q|^{4/n} Q & \text{in } \mathbb{R}^n, \\ Q > 0 & \text{in } \mathbb{R}^n. \end{cases} \tag{2.13}$$

COROLLARY 2.8 (global existence). *Let $\lambda \in \mathbb{R}$, $\sigma, \omega > 0$ with $\sigma < 2/(n-2)$ if $n \geq 3$, $u_0 \in \Sigma$, and $u \in Y_{loc}(]-T_*, T^*[)$ be the maximal solution given by Corollary 2.7. We have $T_* = T^* = \infty$ in the following cases:*

- *the nonlinearity is repulsive, $\lambda \geq 0$;*
- *$\lambda < 0$ and $\sigma < 2/n$;*
- *$\lambda < 0$, $\sigma = 2/n$, and $\|u_0\|_{L^2} < \|Q\|_{L^2}$;*
- *$\lambda < 0$, $\sigma > 2/n$, and $\|u_0\|_{H^1}$ is sufficiently small.*

In addition, if $\lambda \geq 0$, we have $u \in Y(\mathbb{R})$, and (1.2) is globally well-posed.

Proof. We shall prove that under our assumptions, $T^* = \infty$; the proof that $T_* = \infty$ is similar. From Corollary 2.7, it suffices to proves that the L^2-norm of $J(t)u$ cannot blow up in finite time.

Assume $\lambda > 0$. If $\sigma \geq 2/n$, then (2.10) implies that for any $t \geq 0$, $E_1(t) \leq E_1(0)$, which yields an a priori bound for the L^2-norm of $J(t)u$. From Corollary 2.7, this yields $T^* = \infty$. If $\sigma < 2/n$, it follows from (2.10) that for $t \geq 0$,

$$\cosh^2(\omega t)\|u(t)\|_{L^{2\sigma+2}}^{2\sigma+2} \leq \frac{\sigma+1}{\lambda} E_1(0) + \omega\left(1 - \frac{n\sigma}{2}\right)\int_0^t \sinh(2\omega s)\|u(s)\|_{L^{2\sigma+2}}^{2\sigma+2} ds\,.$$

The Gronwall lemma applied to the function defined by the left-hand side of the above inequality yields

$$\|u(t)\|_{L^{2\sigma+2}}^{2\sigma+2} \leq \frac{\sigma+1}{\lambda} E_1(0)\big(\cosh(\omega t)\big)^{-n\sigma}.$$

Plugging this estimate into (2.10), we have

$$\|J(t)u\|_{L^2}^2 \lesssim e^{(2-n\sigma)\omega t}. \tag{2.14}$$

This yields $T^* = \infty$.

Now assume $\lambda < 0$. If $\sigma < 2/n$, it follows from (2.10) and (2.7) that, for $t \geq 0$,

$$\begin{aligned} \frac{1}{2}\|J(t)u\|_{L^2}^2 &\leq E_1(0) + \frac{|\lambda|}{\sigma+1}\cosh^2(\omega t)\|u(t)\|_{L^{2\sigma+2}}^{2\sigma+2} \\ &\leq E_1(0) + C\big(\cosh(\omega t)\big)^{2-n\sigma}\|u_0\|_{L^2}^{(2-n)\sigma+2}\|J(t)u\|_{L^2}^{n\sigma}\,. \end{aligned}$$

Since $\sigma < 2/n$, this means that

$$\sup_{t\geq 0} \left(\cosh(\omega t)\right)^{-2} \|J(t)u\|_{L^2}^2 < \infty,$$

and Corollary 2.7 yields global existence in the future.

If $\sigma = 2/n$, the same argument as above yields

$$\frac{1}{2}\|J(t)u\|_{L^2}^2 \leq E_1(0) + \frac{|\lambda|}{\sigma+1} c_{2+4/n}^{2+4/n} \|u_0\|_{L^2}^{4/n} \|J(t)u\|_{L^2}^2 \,,$$

where $c_{2+4/n}$ is the constant of Gagliardo–Nirenberg inequality mentioned in the third point of Lemma 2.3. M. Weinstein [24] proved that the best such constant satisfies

$$\frac{|\lambda|}{\sigma+1} c_{2+4/n}^{2+4/n} \|Q\|_{L^2}^{4/n} = \frac{1}{2}\,,$$

where Q is the radial solution of (2.13). Thus if $\|u_0\|_{L^2} < \|Q\|_{L^2}$, we obtain an a priori bound for $\|J(t)u\|_{L^2}$, which implies $T^* = \infty$.

Finally, if $\sigma > 2/n$, we have

$$\begin{aligned} E_1(t) &\leq E_1(0) + C\int_0^t \sinh(2\omega s)\|u(s)\|_{L^{2\sigma+2}}^{2\sigma+2} ds \\ &\leq C\big(\|u_0\|_{H^1}\big) + C\big(\|u_0\|_{L^2}\big) \sup_{0\leq s\leq t} \|J(s)u\|_{L^2}^{n\sigma} \int_0^t \frac{\sinh(\omega s)}{\cosh^{n\sigma-1}(\omega s)} ds\,. \end{aligned}$$

Therefore,

$$\sup_{0\leq s\leq t} \|J(s)u\|_{L^2}^2 \leq C\big(\|u_0\|_{H^1}\big) + C\big(\|u_0\|_{L^2}\big) \sup_{0\leq s\leq t} \|J(s)u\|_{L^2}^{n\sigma}$$

for $C\big(\|u_0\|_{H^1}\big)$ and $C\big(\|u_0\|_{L^2}\big)$ going to zero with their argument. Now we can use the following lemma, whose easy proof is omitted.

LEMMA 2.9 (bootstrap argument). *Let $M = M(t)$ be a nonnegative continuous function on $[0,T]$ such that, for every $t \in [0,T]$,*

$$M(t) \leq a + bM(t)^\theta,$$

where $a, b > 0$ and $\theta > 1$ are constants such that

$$a < \left(1 - \frac{1}{\theta}\right) \frac{1}{(\theta b)^{1/(\theta-1)}}\ , \quad M(0) \leq \frac{1}{(\theta b)^{1/(\theta-1)}}\ .$$

Then, for every $t \in [0,T]$, we have

$$M(t) \leq \frac{\theta}{\theta - 1}\ a\ .$$

Taking $\|u_0\|_{H^1}$ sufficiently small, we can apply the above lemma and obtain an a priori bound for $\|J(t)u\|_{L^2}$.

We now have to prove the last assertion of the corollary; that is, if $\lambda \geq 0$, then

$$A(t)u \in L^q(\mathbb{R}; L^r)\ \forall A \in \{Id, J, H\},\ \forall (q,r) \text{ admissible.} \tag{2.15}$$

Let $\lambda \geq 0$ and $A \in \{Id, J, H\}$. In the first part of the proof, we saw that according to the case considered ($\sigma \geq 2/n$, or $\sigma < 2/n$), either $J(t)u \in L^\infty(\mathbb{R}; L^2)$ or it satisfies estimates (2.14). It is easy to check that in either of these two cases, $H(t)u$ satisfies the same estimate as $J(t)u$. Since the second estimate is weaker than the first one, it suffices to prove that it yields (2.15). We will use the following algebraic lemma.

LEMMA 2.10. *Let $r = s = 2\sigma + 2$ and q be such that the pair (q, r) is admissible. Define k by*

$$k = \frac{2\sigma(2\sigma+2)}{2-(n-2)\sigma}\,.$$

Then k is finite, and the following algebraic identities hold:

$$\begin{cases} \dfrac{1}{r'} = \dfrac{1}{r} + \dfrac{2\sigma}{s}\ , \\ \dfrac{1}{q'} = \dfrac{1}{q} + \dfrac{2\sigma}{k}\ . \end{cases}$$

Let q, r, k, and s be as in the above lemma. From (2.14) and the conservation of mass, (2.7) yields

$$\|u\|_{L^k([T,\infty[;L^s)} \leq C \left\| e^{-n\sigma\delta(s)t/2} \right\|_{L^k([T,\infty[)} \leq Ce^{-n\sigma\delta(s)T/2}. \tag{2.16}$$

To prove that $A(t)u \in L^q([0,\infty[; L^r)$, write Duhamel's principle with the time origin equal to T, to be fixed later, as

$$u(t) = U_\omega(t-T)u(T) - i\lambda \int_T^t U_\omega(t-s)\left(|u|^{2\sigma}u\right)(s)ds\,.$$

Applying operator A yields

$$A(t)u = U_\omega(t-T)A(T)u - i\lambda \int_T^t U_\omega(t-s)A(s)\left(|u|^{2\sigma}u\right)ds\,,$$

and from Strichartz inequalities and Lemma 2.10, for any $S > T$,

$$\begin{aligned} \|A(t)u\|_{L^q([T,S];L^r)} &\leq C_r\|A(T)u\|_{L^2} + C_{r,r}\left\|A(t)\left(|u|^{2\sigma}u\right)\right\|_{L^{q'}([T,S];L^{r'})} \\ &\leq C_r\|A(T)u\|_{L^2} + \underline{C}\|u\|^{2\sigma}_{L^k([T,\infty[;L^s)}\|A(t)\|_{L^q([T,S];L^r)}\,, \end{aligned}$$

where $\underline{C}$ does not depend on T, S. From (2.16), choosing T sufficiently large, the second term of the right-hand side can be absorbed by the left-hand side, and

$$\|A(t)u\|_{L^q([T,S];L^r)} \leq 2C_r\|A(T)u\|_{L^2}\,.$$

Since $S > T$ is arbitrary, this implies $A(t)u \in L^q(\mathbb{R}_+; L^r)$. Similarly, $A(t)u \in L^q(\mathbb{R}; L^r)$; (2.15) is proven for the admissible pair (q, r) such that $r = 2\sigma + 2$. Strichartz inequality (2.2) then yields (2.15) for any admissible pair. Indeed, if (q_1, r_1) is admissible,

$$\begin{aligned} \|A(t)u\|_{L^{q_1}([0,S];L^{r_1})} &\leq C_{r_1}\|u_0\|_\Sigma + |\lambda|C_{r,1,r}\left\|A(t)\left(|u|^{2\sigma}u\right)\right\|_{L^{q'}([0,S];L^{r'})} \\ &\leq C + C\|u\|^{2\sigma}_{L^k(\mathbb{R};L^s)}\|A(t)\|_{L^q(\mathbb{R};L^r)}\ . \end{aligned}$$

This completes the proof of Corollary 2.8. □

3. Scattering theory. In this section, we prove that the influence of the nonlinear term in (1.2) is negligible as time becomes large (at least if $\lambda > 0$), without the usual restriction on the power of the nonlinearity encountered for scattering theory associated to (1.3). We first prove the existence of wave operators and then their asymptotic completeness.

PROPOSITION 3.1 (existence of wave operators). *In either of the cases considered in Corollary* 2.8*, the following holds.*

- *For every $u_- \in \Sigma$, there exists a unique $u_0 \in \Sigma$ such that the maximal solution $u \in C(\mathbb{R};\Sigma)$ of* (1.2) *satisfies*

$$\left\|U_\omega(-t)u(t) - u_-\right\|_{\Sigma} \underset{t\to-\infty}{\longrightarrow} 0\,.$$

- *For every $u_+ \in \Sigma$, there exists a unique $u_0 \in \Sigma$ such that the maximal solution $u \in C(\mathbb{R};\Sigma)$ of* (1.2) *satisfies*

$$\left\|U_\omega(-t)u(t) - u_+\right\|_{\Sigma} \underset{t\to+\infty}{\longrightarrow} 0\,.$$

Proof. We prove the first point; the proof of the second is similar. We solve the following equation by a fixed point argument:

$$u(t) = U_\omega(t)u_- - i\lambda\int_{-\infty}^{t} U_\omega(t-s)\big(|u|^{2\sigma}u\big)(s)ds\,. \tag{3.1}$$

Define $F(u)(t)$ as the right-hand side of (3.1), and let $R := \|u_-\|_\Sigma$. Recall that, as stated in Lemma 2.10, (q,r) is the admissible pair such that $r = 2\sigma+2$.

We first prove that there exists $T > 0$ such that the set

$$\begin{aligned} X_T := \big\{ & u \in Y_{2\sigma+2}(]-\infty,-T]);\ \|A(t)u\|_{L^2} \le 2R\ \forall t \le -T,\ A \in \{Id,J,H\},\\ & \|A(t)u\|_{L^q(]-\infty,-T];L^r)} \le 2C_{2\sigma}R\ \forall A \in \{Id,J,H\}\big\} \end{aligned}$$

is stable under the map F, where $C_{2\sigma}$ is the constant in Strichartz inequality (2.1). We then prove that choosing T even larger, F is a contraction on $L^q(]-\infty,-T];L^r)$.

For any pair (a,b), we use the notation

$$\|f\|_{L^a_T(L^b)} = \|f\|_{L^a(]-\infty,-T];L^b)}.$$

Let $u \in X_T$ and $A \in \{Id,J,H\}$. From Lemmas 2.2, 2.3, and 2.10,

$$\begin{aligned} \|A(t)F(u)\|_{L^\infty_T(L^2)} &\le \|u_-\|_\Sigma + C_{2,2\sigma+2}|\lambda|\left\|A(t)\left(|u|^{2\sigma}u\right)\right\|_{L^{q'}_T(L^{r'})}\\ &\le R + C\left\||u|^{2\sigma}A(t)u\right\|_{L^{q'}_T(L^{r'})}\\ &\le R + C\,\|u\|^{2\sigma}_{L^k_T(L^s)}\,\|A(t)u\|_{L^q_T(L^r)}\ . \end{aligned}$$

From (2.7) and Lemma 2.10,

$$\|u\|_{L^k_T(L^s)} \le C_k R\left\|\left(\frac{1}{\cosh(\omega t)}\right)^{\delta(s)}\right\|_{L^k(]-\infty,-T])} \le C(\omega,\sigma)Re^{-\omega\delta(s)T}.$$

It follows that

$$\|A(t)F(u)\|_{L^\infty_T(L^2)} \le R + CR^{2\sigma+1}e^{-2\sigma\omega\delta(s)T}\ . \tag{3.2}$$

Use Lemmas 2.2, 2.3, and 2.10 again to obtain

$$\begin{aligned}\|A(t)F(u)\|_{L^q_T(L^r)} &\leq C_{2\sigma}R + C\,\|u\|^{2\sigma}_{L^k_T(L^s)}\,\|A(t)u\|_{L^q_T(L^r)}\\ &\leq C_{2\sigma}R + CR^{2\sigma+1}e^{-2\sigma\omega\delta(s)T}\ .\end{aligned}$$

It is now clear that if T is sufficiently large, then X_T is stable under F.

To complete the proof of the proposition, following the argument used in [14], it is enough to prove contraction for large T in the weaker metric $L^q(]-\infty,-T];L^r)$. From Lemmas 2.2, 2.3, and 2.10, we have

$$\begin{aligned}\left\|F(u_2)-F(u_1)\right\|_{L^q_T(L^r)} &\leq C\left\|\left(|u_2|^{2\sigma}u_2-|u_1|^{2\sigma}u_1\right)\right\|_{L^{q'}_T(L^{r'})}\\ &\leq C\left(\|u_1\|^{2\sigma}_{L^k_T(L^s)}+\|u_2\|^{2\sigma}_{L^k_T(L^s)}\right)\|u_2-u_1\|_{L^q_T(L^r)}\,.\end{aligned}\tag{3.3}$$

As above, we have the estimate

$$\|u_1\|^{2\sigma}_{L^k_T(L^s)}+\|u_2\|^{2\sigma}_{L^k_T(L^s)}\leq CR^{2\sigma}e^{-2\sigma\omega\delta(s)T}.$$

Therefore, contraction follows for T sufficiently large.

From Corollary 2.8, the solution u we obtain by this fixed point argument is defined not only on $]-\infty,-T]$ for T large but also globally. Proposition 3.1 then follows from Corollaries 2.7 and 2.8. □

REMARK 3.2. *The fact that we limit ourselves to the cases considered in Corollary* 2.8 *in the above proposition is needed only to ensure that the solution u we construct is defined up to time $t=0$. To solve* (3.1) *in the neighborhood of $-\infty$, we used only the assumptions of Proposition* 2.5.

PROPOSITION 3.3 (asymptotic completeness). *Let $\lambda\geq 0$, $\sigma>0$, with $\sigma<2/(n-2)$ if $n\geq 3$.*

- *For every $u_0\in\Sigma$, there exists a unique $u_-\in\Sigma$ such that the maximal solution $u\in C(\mathbb{R};\Sigma)$ of* (1.2) *satisfies*

$$\left\|U_\omega(-t)u(t)-u_-\right\|_{\Sigma}\underset{t\to-\infty}{\longrightarrow}0\,.$$

- *For every $u_0\in\Sigma$, there exists a unique $u_+\in\Sigma$ such that the maximal solution $u\in C(\mathbb{R};\Sigma)$ of* (1.2) *satisfies*

$$\left\|U_\omega(-t)u(t)-u_+\right\|_{\Sigma}\underset{t\to+\infty}{\longrightarrow}0\,.$$

Proof. We prove the second point; the proof of the first is similar. Let $u_0\in\Sigma$.

Since Σ is a Hilbert space, it is enough to prove that the family $\left(U_\omega(-t)u(t)\right)_{t\geq 0}$ is a Cauchy sequence as t goes to $+\infty$. From Duhamel's principle (2.11), we have

$$U_\omega(-t)u(t)=u_0-i\lambda\int_0^t U_\omega(-s)\left(|u|^{2\sigma}u\right)(s)ds\,.$$

Let $B\in\{Id,\nabla_x,x\}$, and let $A\in\{Id,J,H\}$ be its counterpart given by the commutation property (2.5). We have

$$B\left(U_\omega(-t)u(t)\right)=Bu_0-i\lambda\int_0^t U_\omega(-s)A(s)\left(|u|^{2\sigma}u\right)ds\,.$$

Let $t_2 \geq t_1 > 0$. From Strichartz inequality (2.2),

$$\begin{aligned}\big\|B\big(U_\omega(-t_2)u(t_2) - U_\omega(-t_1)u(t_1)\big)\big\|_{L^2} &\\ \leq C\left\|\int_{t_1}^{t} U_\omega(-s)A(s)\left(|u|^{2\sigma}u\right)ds\right\|_{L^\infty([t_1,t_2];L^2)} &\\ \leq C\left\|A(t)\left(|u|^{2\sigma}u\right)\right\|_{L^{q'}([t_1,t_2];L^{r'})}.&\end{aligned}$$

From Lemmas 2.3 and 2.10, this yields

$$\begin{aligned}\big\|B\big(U_\omega(-t_2)u(t_2) - U_\omega(-t_1)u(t_1)\big)\big\|_{L^2} &\leq C\|u\|^{2\sigma}_{L^k([t_1,t_2];L^s)}\|A(t)u\|_{L^q([t_1,t_2];L^r)}\\ &\leq C\|u\|^{2\sigma}_{L^k([t_1,t_2];L^s)}\|A(t)u\|_{L^q([t_1,t_2];L^r)}.\end{aligned}$$

We saw in the proof of Corollary 2.8 that $u \in L^k(\mathbb{R};L^s)$ and $A(t)u \in L^q(\mathbb{R};L^r)$, which implies that $\big(B(U_\omega(-t)u(t))\big)_{t>0}$ is a Cauchy sequence in L^2, which completes the proof of the proposition. □

Propositions 3.1 and 3.3 imply Theorem 1.5, and even more, since we do not necessarily assume that the nonlinearity is defocusing.

4. Blow-up in finite time. In Corollary 2.8, we proved that if $\lambda < 0$ and $\sigma \geq 2/n$, the solution of (1.2) is global, provided that the initial datum u_0 is small. When u_0 is not small, we show that finite-time blow-up may occur, as in the case of (1.3). However, the sufficient condition we state below is stronger than its counterpart (1.4) for (1.3); in some sense, blow-up in finite time is less likely to occur for (1.2) than for (1.3).

PROPOSITION 4.1. *Let $u_0 \in \Sigma$, $\lambda < 0$, $\sigma \geq 2/n$, with $\sigma < 2/(n-2)$ if $n \geq 3$. If u_0 satisfies*

$$\frac{1}{2}\|\nabla u_0\|^2_{L^2} + \frac{\lambda}{\sigma+1}\|u_0\|^{2\sigma+2}_{L^{2\sigma+2}} < -\frac{\omega^2}{2}\|xu_0\|^2_{L^2},$$

then the solution u to (1.2) blows up in finite time, in the future or in the past. More precisely,

- *if $\operatorname{Im}\int \overline{u_0}x\cdot\nabla u_0 \leq 0$, then $T^* < \infty$—that is, u blows up in the future;*
- *if $\operatorname{Im}\int \overline{u_0}x\cdot\nabla u_0 \geq 0$, then $T_* < \infty$—that is, u blows up in the past.*

If, moreover,

$$\frac{1}{2}\|\nabla u_0\|^2_{L^2} + \frac{\lambda}{\sigma+1}\|u_0\|^{2\sigma+2}_{L^{2\sigma+2}} < -\frac{\omega^2}{2}\|xu_0\|^2_{L^2} - \omega\left|\operatorname{Im}\int \overline{u_0}x\cdot\nabla_x u_0\right|,$$

then u blows up in the past and *in the future.*

Proof. We follow the Zakharov–Glassey method. Denote $y(t) := \|xu(t)\|^2_{L^2}$. We show that $y(t)$ satisfies a second-order ordinary differential equation, from which the proposition follows.

Step 1. *Formal computations.* Differentiating $y(t)$ and using (1.2) yields

$$\dot{y}(t) = 2\operatorname{Im}\int \overline{u}(t,x)x\cdot\nabla_x u(t,x)dx.$$

Expanding $\|J(t)u\|^2_{L^2}$, we have

$$\begin{aligned}\|J(t)u\|^2_{L^2} = \omega^2\sinh^2(\omega t)y(t) + \cosh^2(\omega t)\|\nabla_x u(t)\|^2_{L^2}&\\ - \omega\sinh(2\omega t)\operatorname{Im}\int \overline{u}(t,x)x\cdot\nabla_x u(t,x)dx,&\end{aligned}$$

and from the conservation of energy (2.8),

$$E_1(t) = \frac{\omega^2}{2}\sinh^2(\omega t)y(t) + \cosh^2(\omega t)\left(E + \frac{\omega^2}{2}y(t)\right) - \frac{\omega}{2}\sinh(2\omega t)\,\mathrm{Im}\int \overline{u}(t,x)x\cdot\nabla_x u(t,x)dx\,.$$

Using the evolution law (2.10) and the above formula for $\dot y(t)$ yields

$$\frac{d}{dt}\,\mathrm{Im}\int \overline{u}(t,x)x\cdot\nabla_x u(t,x)dx = 2\omega^2 y(t) + 2E - \frac{\lambda}{\sigma+1}(2-n\sigma)\|u(t)\|^{2\sigma+2}_{L^{2\sigma+2}}\,.$$

It follows that

$$\ddot y(t) = 4\omega^2 y(t) + 4E - \frac{2\lambda}{\sigma+1}(2-n\sigma)\|u(t)\|^{2\sigma+2}_{L^{2\sigma+2}}\,. \tag{4.1}$$

Step 2. *Justification.* One has to know that $y\in C^1(]-T_*,T^*[)$; the rest of the computations follow from (2.10). The argument is classical (it consists of considering $y^\varepsilon(t) := \|e^{-\varepsilon|x|^2}xu(t)\|^2_{L^2}$ and eventually letting ε go to zero), and we refer to [6, sect. 6.4] for more details, as we did for the proof of (2.10).

Step 3. *Conclusion.* From classical ordinary differential equations methods, the solution of (4.1) is given by the formula

$$y(t) = y(0)\cosh(2\omega t) + \dot y(0)\frac{\sinh(2\omega t)}{2\omega} + \int_0^t \frac{\sinh(2\omega(t-s))}{2\omega}f(s)ds\,,$$

where $f(t) = 4E - \frac{2\lambda}{\sigma+1}(2-n\sigma)\|u(t)\|^{2\sigma+2}_{L^{2\sigma+2}}$. Since $\lambda<0$ and $\sigma\geq 2/n$,

$$\begin{aligned} y(t) &\leq y(0)\cosh(2\omega t) + \dot y(0)\frac{\sinh(2\omega t)}{2\omega} + \int_0^t \frac{\sinh(2\omega(t-s))}{2\omega}4Eds \\ &\leq y(0)\cosh^2(\omega t) + \frac{2\sinh^2(\omega t)}{\omega^2}\left(\frac12\|\nabla u_0\|^2_{L^2} + \frac{\lambda}{\sigma+1}\|u_0\|^{2\sigma+2}_{L^{2\sigma+2}}\right) \\ &\quad + \dot y(0)\frac{\sinh(2\omega t)}{2\omega}\,. \end{aligned}$$

Assume $\dot y(0)\leq 0$. Then for positive times, the above estimate implies

$$\begin{aligned} y(t) &\leq y(0)\cosh^2(\omega t) + \frac{2\sinh^2(\omega t)}{\omega^2}\left(\frac12\|\nabla u_0\|^2_{L^2} + \frac{\lambda}{\sigma+1}\|u_0\|^{2\sigma+2}_{L^{2\sigma+2}}\right) \\ &\leq \cosh^2(\omega t)\left(y(0) + \frac{2\tanh^2(\omega t)}{\omega^2}\left(\frac12\|\nabla u_0\|^2_{L^2} + \frac{\lambda}{\sigma+1}\|u_0\|^{2\sigma+2}_{L^{2\sigma+2}}\right)\right)\,. \end{aligned}$$

Since $\tanh(\mathbb{R}_+) = [0,1[$, it follows that if

$$\frac12\|\nabla u_0\|^2_{L^2} + \frac{\lambda}{\sigma+1}\|u_0\|^{2\sigma+2}_{L^{2\sigma+2}} < -\frac{\omega^2}{2}\|xu_0\|^2_{L^2}\,,$$

then if we suppose $T^*=\infty$, $y(t)$ becomes negative for possibly large t. This is absurd; therefore T^* is finite. Similarly, if $\dot y(0)\geq 0$, then T_* is finite.

Finally, if

$$\frac12\|\nabla u_0\|^2_{L^2} + \frac{\lambda}{\sigma+1}\|u_0\|^{2\sigma+2}_{L^{2\sigma+2}} < -\frac{\omega^2}{2}\|xu_0\|^2_{L^2} - \omega\left|\mathrm{Im}\int \overline{u_0}x\cdot\nabla_x u_0\right|\,,$$

then the same argument shows that T_* and T^* are finite. □

We now prove that, indeed, blow-up in finite time is less likely to occur for (1.2) than for (1.3). For a fixed $u_0 \in \Sigma$, the blow-up sufficient conditions stated in Proposition 4.1 become empty when ω is large. We prove that, for a fixed initial datum u_0, taking ω sufficiently large ensures the global existence of u.

PROPOSITION 4.2. *Let $u_0 \in \Sigma$, $\lambda < 0$, $\sigma > 2/n$, with $\sigma < 2/(n-2)$ if $n \geq 3$. There exists $\omega_1 > 0$ such that for any $\omega \geq \omega_1$, the solution u to* (1.2) *is global, and $u \in Y(\mathbb{R})$.*

Proof. From Proposition 2.5, there exist $t_0 > 0$ and C_0 independent of $\omega > 0$ such that (1.2) has a unique solution $u \in Y(]-2t_0, 2t_0[)$, which also satisfies estimate (2.9), that is,

$$\sup_{|t|\leq t_0} \|u(t)\|_{L^2} + \sup_{|t|\leq t_0} \|J(t)u\|_{L^2} + \sup_{|t|\leq t_0} \|H(t)u\|_{L^2} \leq C_0.$$

The idea is to mimic the proof of the fourth case in Corollary 2.8 by replacing the smallness assumption by the property $\omega \gg 1$.

Integrate the evolution law (2.10) between time t_0 and time $t > t_0$:

$$E_1(t) - E_1(t_0) = \frac{\omega\lambda}{2\sigma+2}(2-n\sigma)\int_{t_0}^{t} \sinh(2\omega s)\|u(s)\|_{L^{2\sigma+2}}^{2\sigma+2} ds\,.$$

From Proposition 2.5 and the fact that the nonlinearity we consider is focusing, we have

$$E_1(t_0) \leq \frac{1}{2}\|J(t_0)u\|_{L^2}^2 \leq \frac{1}{2}C_0^2\,,$$

where C_0 does not depend on ω.

Using modified Gagliardo–Nirenberg inequalities (2.7), we have

$$\begin{aligned} E_1(t) &\leq \frac{1}{2}C_0^2 + C(\lambda,\sigma)\omega\int_{t_0}^{t} \frac{\sinh(2\omega s)}{\big(\cosh(\omega s)\big)^{n\sigma}}\|u(s)\|_{L^2}^{2+(2-n)\sigma}\|J(s)u\|_{L^2}^{n\sigma} ds \\ &\leq \frac{1}{2}C_0^2 + C'(\lambda,\sigma)\|u_0\|_{L^2}^{2+(2-n)\sigma} \sup_{t_0\leq s\leq t} \|J(s)u\|_{L^2}^{n\sigma}\big(\cosh(\omega t_0)\big)^{2-n\sigma}. \end{aligned}$$

From the definition of E_1, this yields

$$\begin{aligned} \frac{1}{2}\|J(t)u\|_{L^2}^2 &\leq \frac{1}{2}C_0^2 + C\big(\cosh(\omega t)\big)^2\|u(t)\|_{L^{2\sigma+2}}^{2\sigma+2} \\ &\quad + C\sup_{t_0\leq s\leq t} \|J(s)u\|_{L^2}^{n\sigma}\big(\cosh(\omega t_0)\big)^{2-n\sigma} \\ &\leq \frac{1}{2}C_0^2 + C\big(\cosh(\omega t)\big)^{2-n\sigma}\|J(t)u\|_{L^2}^{n\sigma} \\ &\quad + C\sup_{t_0\leq s\leq t} \|J(s)u\|_{L^2}^{n\sigma}\big(\cosh(\omega t_0)\big)^{2-n\sigma} \\ &\leq \frac{1}{2}C_0^2 + C\big(\cosh(\omega t_0)\big)^{2-n\sigma}\|J(t)u\|_{L^2}^{n\sigma} \\ &\quad + C\sup_{t_0\leq s\leq t} \|J(s)u\|_{L^2}^{n\sigma}\big(\cosh(\omega t_0)\big)^{2-n\sigma}\,, \end{aligned}$$

where the above constants do not depend on ω. We finally obtain

$$\sup_{t_0\leq s\leq t} \|J(s)u\|_{L^2}^2 \leq\ C_0^2 + C\big(\cosh(\omega t_0)\big)^{2-n\sigma} \sup_{t_0\leq s\leq t} \|J(s)u\|_{L^2}^{n\sigma}\,,$$

which can also be written

$$\sup_{t_0\leq s\leq t}\|J(s)u\|_{L^2}^2\leq C_0^2+f(\omega)\left(\sup_{t_0\leq s\leq t}\|J(s)u\|_{L^2}^2\right)^{n\sigma/2}\,,$$

where C_0 does not depend on ω, and $f(\omega)\to 0$ as $\omega\to+\infty$, because $n\sigma>2$. Lemma 2.9 shows that for ω sufficiently large, $J(t)u$ is uniformly bounded in L^2 for $t\geq 0$. Corollary 2.7 then implies $u\in Y_{\rm loc}(\mathbb{R}_+)$, and one can repeat the end of the proof of Corollary 2.8 to deduce that $u\in Y(\mathbb{R}_+)$.

Proving $u\in Y(\mathbb{R}_-)$ is similar, so we omit that part. □

The proof of Theorem 1.1 is not complete yet, since in the above proposition, we assumed only $\sigma>2/n$, while in Theorem 1.1, we assumed $\sigma\geq 2/n$. The remaining case, $\sigma=2/n$, is treated in the next section.

5. The particular case $\sigma=2/n$. Let $\lambda\in\mathbb{R}$, $u_0\in\Sigma$. Let v solve (1.3) with a critical power, that is,

$$\left\{\begin{array}{l} i\partial_t v+\dfrac{1}{2}\Delta v=\lambda|v|^{4/n}v\,,\quad (t,x)\in\mathbb{R}\times\mathbb{R}^n\,,\\ v_{|t=0}=u_0\,.\end{array}\right. \tag{5.1}$$

Let $\omega>0$. In [4], we noticed that if u^+ is defined by

$$u^+(t,x)=\frac{1}{(\cos(\omega t))^{n/2}}e^{-i\frac{\omega}{2}|x|^2\tan(\omega t)}v\left(\frac{\tan(\omega t)}{\omega},\frac{x}{\cos(\omega t)}\right)\,, \tag{5.2}$$

then u^+ solves (1.6) with $\sigma=2/n$. We also proved that v blows up at time $T>0$ if and only if u^+ blows up at time $\arctan(\omega T)/\omega$. The first point of Theorem 1.2 is therefore a reminder of a result stated in [4].

As noticed in the introduction, replacing ω by $\pm i\omega$ formally turns (1.6) into (1.2). Following this idea again, define

$$u^-(t,x)=\frac{1}{(\cosh(\omega t))^{n/2}}e^{i\frac{\omega}{2}|x|^2\tanh(\omega t)}v\left(\frac{\tanh(\omega t)}{\omega},\frac{x}{\cosh(\omega t)}\right)\,. \tag{5.3}$$

Then from Proposition 2.5, u^- is the solution of

$$\left\{\begin{array}{l} i\partial_t u^-+\dfrac{1}{2}\Delta u^-=-\omega^2\dfrac{|x|^2}{2}u^-+\lambda|u^-|^{4/n}u^-\,,\quad (t,x)\in\mathbb{R}\times\mathbb{R}^n\,,\\ u^-_{|t=0}=u_0\,.\end{array}\right. \tag{5.4}$$

Now assume that $\lambda<0$ and that v blows up at some finite time $T>0$.

From the factorization (2.6), it is easy to see that

$$\|J(t)u^-\|_{L^2}=\left\|\nabla_x v\left(\frac{\tanh(\omega t)}{\omega}\right)\right\|_{L^2}\,. \tag{5.5}$$

Since $\tanh(\mathbb{R}_+)=[0,1[$, if $\omega\geq 1/T$, then the function of the right-hand side of (5.3) does not "see" the time T, and from Corollary 2.7, u^- does not blow up in finite time.

If $\omega<1/T$, then (5.5) and Corollary 2.7 show that u^- blows up at time

$$T_\omega=\frac{\arg\tanh(\omega T)}{\omega}\,,$$

which completes the proof of Theorem 1.2.

We can go further in the analysis of the influence of the parameter ω.

PROPOSITION 5.1. *Let $u_0 \in \Sigma$, $\lambda < 0$. For $\omega \geq 0$, denote u^ω the solution of the initial value problem*

$$\begin{cases} i\partial_t u^\omega + \dfrac{1}{2}\Delta u^\omega = -\omega^2 \dfrac{|x|^2}{2} u^\omega + \lambda |u^\omega|^{4/n} u^\omega \,, \quad (t,x) \in \mathbb{R} \times \mathbb{R}^n \,, \\ u^\omega_{|t=0} = u_0 \,. \end{cases} \tag{5.6}$$

Let $\omega_* \geq 0$.

• *If u^{ω_*} is defined globally, $u^{\omega_*} \in Y_{loc}(\mathbb{R})$, then for every $\omega \geq \omega_*$, u^ω is also defined globally, $u^\omega \in Y_{loc}(\mathbb{R})$.*

• *Suppose that there exists $T_* > 0$ such that u^{ω_*} blows up at time T_*. Then for every $0 \leq \omega < \omega_* / \tanh(\omega_* T_*)$, u^ω blows up in finite time, and for every $\omega \geq \omega_* / \tanh(\omega_* T_*)$, u^ω is defined globally, $u^\omega \in Y_{loc}(\mathbb{R})$.*

Proof. Let v be the solution of (5.1). If v is defined globally in Σ, then so is u^ω for any $\omega \geq 0$. If v blows up in finite time $T_0 > 0$ while u^{ω_*} is defined globally, then Theorem 1.2 implies $\omega_* \geq 1/T_0$. Using Theorem 1.2 again, u^ω is defined globally for any $\omega \geq \omega_* \geq 1/T_0$.

Now assume u^{ω_*} blows up at time $T_* > 0$. From Theorem 1.2, v blows up at time T_0, with

$$T_0 = \frac{\tanh(\omega_* T_*)}{\omega_*} \,.$$

The last point of Theorem 1.2 implies that if $\omega \geq 1/T_0$, then u^ω is defined globally. Similarly, if $0 \leq \omega < \omega_* / \tanh(\omega_* T_*)$, then u^ω blows up at time

$$T_\omega = \frac{1}{\omega} \operatorname{arg\,tanh}\left(\frac{\omega \tanh(\omega_* T_*)}{\omega_*} \right),$$

which completes the proof of the proposition. □

Assume that $\lambda < 0$ and that v blows up in finite time. Theorem 1.2 and Proposition 5.1 show that there is a critical value for the parameter ω, which is the inverse of the blow-up time for v. What happens for that critical value? We can answer this question in the case where the mass of the initial datum is critical. We saw in Corollary 2.8 that if $\|u_0\|_{L^2} < \|Q\|_{L^2}$, where Q is the spherically symmetric solution of (2.13), then the solution to (5.6) is global for any $\omega \geq 0$. If $\|u_0\|_{L^2} = \|Q\|_{L^2}$, then blow-up in finite time may occur. This phenomenon was studied very precisely by Merle in the case of (5.1).

THEOREM 5.2 (see [17, Thm. 1]). *Let $\lambda < 0$, $u_0 \in H^1(\mathbb{R}^n)$, and assume that the solution v of (5.1) blows up in finite time $T > 0$. Moreover, assume that $\|u_0\|_{L^2} = \|Q\|_{L^2}$, where Q is defined by (2.13). Then there exist $\theta \in \mathbb{R}$, $\delta > 0$, $x_0, x_1 \in \mathbb{R}^n$ such that*

$$u_0(x) = \left(\frac{\delta}{T}\right)^{n/2} e^{i\theta - i|x-x_1|^2/2T + i\delta^2/T} Q\left(\delta\left(\frac{x-x_1}{T} - x_0\right)\right),$$

and for $t < T$,

$$v(t,x) = \left(\frac{\delta}{T-t}\right)^{n/2} e^{i\theta - i|x-x_1|^2/2(T-t) + i\delta^2/(T-t)} Q\left(\delta\left(\frac{x-x_1}{T-t} - x_0\right)\right).$$

We use this result only to understand the role of ω in preventing blow-up when the mass is critical, but other applications are possible (see [4] for the case of a confining harmonic potential). With the above theorem and the change of variable (5.3), the following result is straightforward.

COROLLARY 5.3. *Let $\lambda < 0$ and $T > 0$. Assume that u_0 is given by*

$$u_0(x) = \frac{1}{T^{n/2}} e^{-i|x|^2/2T + i/T} Q\left(\frac{x}{T}\right).$$

For $\omega \geq 0$, denote by u^ω the solution of (5.6).

- *If $0 \leq \omega < 1/T$, then u^ω blows up at time* $\arg\tanh(\omega T)/\omega$, *with the profile Q.*
- *If $\omega > 1/T$, then u^ω is defined globally with exponential decay, $u^\omega \in Y(\mathbb{R})$.*
- *If $\omega = 1/T$, then u^ω is defined globally with only $u^\omega \in Y_{loc}(\mathbb{R})$. More precisely,*

$$\begin{aligned} u^{1/T}(t,x) &= \left(\omega e^{\omega t}\right)^{n/2} Q\left(\omega x e^{\omega t}\right) e^{-i\omega|x|^2/2 + i\omega(e^{2\omega t}+1)/2} \\ &= \left(\frac{e^{t/T}}{T}\right)^{n/2} Q\left(\frac{x e^{t/T}}{T}\right) e^{-i|x|^2/2T + i(e^{2t/T}+1)/2T}. \end{aligned}$$

The critical value $\omega = 1/T$ thus leads to a global solution (we already knew that from Theorem 1.2), which may have exponential growth (and does in the particular case $\|u_0\|_{L^2} = \|Q\|_{L^2}$).

REFERENCES

[1] J. E. BARAB, *Nonexistence of asymptotically free solutions for a nonlinear Schrödinger equation*, J. Math. Phys., 25 (1984), pp. 3270–3273.

[2] R. CARLES, *Geometric optics and long range scattering for one-dimensional nonlinear Schrödinger equations*, Comm. Math. Phys., 220 (2001), pp. 41–67.

[3] R. CARLES, *Remarks on nonlinear Schrödinger equations with harmonic potential*, Ann. Henri Poincaré, 3 (2002), pp. 757–772.

[4] R. CARLES, *Critical nonlinear Schrödinger equations with and without harmonic potential*, Math. Models Methods Appl. Sci., 12 (2002), pp. 1513–1523.

[5] R. CARLES, C. FERMANIAN, AND I. GALLAGHER, *On the role of quadratic oscillations in nonlinear Schrödinger equations*, J. Funct. Anal., to appear.

[6] T. CAZENAVE, *An Introduction to Nonlinear Schrödinger Equations*, Text. Met. Mat. 26, Univ. Fed. Rio de Jan., 1993.

[7] T. CAZENAVE AND F. WEISSLER, *Rapidly decaying solutions of the nonlinear Schrödinger equation*, Comm. Math. Phys., 147 (1992), pp. 75–100.

[8] A. DEBUSSCHE AND L. DI MENZA, *Numerical simulation of focusing stochastic nonlinear Schrödinger equations*, Phys. D, 162 (2002), pp. 131–154.

[9] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators. Part* II: *Spectral Theory. Self Adjoint Operators in Hilbert Space*, with the assistance of W. G. Bade and R. G. Bartle, Interscience Publishers/John Wiley & Sons, New York, London, 1963.

[10] R. P. FEYNMAN AND A. HIBBS, *Quantum Mechanics and Path Integrals*, International Series in Pure and Applied Physics, McGraw-Hill, Maidenhead, UK, 1965.

[11] J. GINIBRE, *An introduction to nonlinear Schrödinger equations*, in Nonlinear Waves (Sapporo, 1995), R. Agemi, Y. Giga, and T. Ozawa, eds., GAKUTO Internat. Ser. Math. Sci. Appl., Gakkōtosho, Tokyo, 1997, pp. 85–133.

[12] R. T. GLASSEY, *On the blowing up of solutions to the Cauchy problem for nonlinear Schrödinger equations*, J. Math. Phys., 18 (1977), pp. 1794–1797.

[13] N. HAYASHI AND P. NAUMKIN, *Asymptotics for large time of solutions to the nonlinear Schrödinger and Hartree equations*, Amer. J. Math., 120 (1998), pp. 369–389.

[14] T. KATO, *Nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré Phys. Théor., 46 (1987), pp. 113–129.

[15] M. KEEL AND T. TAO, *Endpoint Strichartz estimates*, Amer. J. Math., 120 (1998), pp. 955–980.

[16] M. K. KWONG, *Uniqueness of positive solutions of $\Delta u - u + u^p = 0$ in $\mathbb{R}^n$*, Arch. Ration. Mech. Anal., 105 (1989), pp. 243–266.

[17] F. Merle, *Determination of blow-up solutions with minimal mass for nonlinear Schrödinger equations with critical power*, Duke Math. J., 69 (1993), pp. 427–454.
[18] Y.-G. Oh, *Cauchy problem and Ehrenfest's law of nonlinear Schrödinger equations with potentials*, J. Differential Equations, 81 (1989), pp. 255–274.
[19] T. Ozawa, *Long range scattering for nonlinear Schrödinger equations in one space dimension*, Comm. Math. Phys., 139 (1991), pp. 479–493.
[20] M. Reed and B. Simon, *Methods of Modern Mathematical Physics.* II. *Fourier Analysis, Self-Adjointness*, Academic Press, New York, 1975.
[21] W. A. Strauss, *Nonlinear scattering theory*, in Scattering Theory in Mathematical Physics, J. Lavita and J. P. Marchand, eds., Reidel, Dordrecht, The Netherlands, 1974.
[22] W. A. Strauss, *Existence of solitary waves in higher dimensions*, Comm. Math. Phys., 55 (1977), pp. 149–162.
[23] W. A. Strauss, *Nonlinear scattering theory at low energy*, J. Funct. Anal., 41 (1981), pp. 110–133.
[24] M. I. Weinstein, *Nonlinear Schrödinger equations and sharp interpolation estimates*, Comm. Math. Phys., 87 (1982/83), pp. 567–576.
[25] K. Yajima, *Smoothness and non-smoothness of the fundamental solution of time dependent Schrödinger equations*, Comm. Math. Phys., 181 (1996), pp. 605–629.
[26] K. Yajima and G. Zhang, *Smoothing property for Schrödinger equations with potential superquadratic at infinity*, Comm. Math. Phys., 221 (2001), pp. 573–590.

SIAM J. MATH. ANAL.
Vol. 35, No. 4, pp. 844–867

TWIST CHARACTER OF THE LEAST AMPLITUDE PERIODIC SOLUTION OF THE FORCED PENDULUM*

JINZHI LEI†, XIONG LI†, PING YAN†, AND MEIRONG ZHANG†

Abstract. In this paper, we will derive some twist criteria for the periodic solution of a periodic scalar Newtonian equation using the third order approximation. As an application to the forced pendulum $\ddot{x} + \omega^2 \sin x = p(t)$, we will find an explicit bound $P(\omega)$ for the L^1 norm, $\|p\|_1$, of the periodic forcing $p(t)$ using the frequency ω as a parameter such that the least amplitude periodic solution of the forced pendulum is of twist type when $\|p\|_1 < P(\omega)$. The bound $P(\omega)$ has the order of $O(\omega^{1/2})$ when ω is bounded away from resonance of orders ≤ 4 and $\omega \to +\infty$.

Key words. forced pendulum, periodic solution, third order approximation, twist coefficient, twist character

AMS subject classifications. 37J25, 34D20, 34C25

DOI. 10.1137/S003614100241037X

1. Introduction.

This paper is motivated by studying the twist character of the least amplitude periodic solution $x_\omega(t)$ of the forced pendulum

$$\ddot{x} + \omega^2 \sin x = p(t), \tag{1.1}$$

where the frequency $\omega > 0$ and the forcing $p \in C(\mathbb{R}/2\pi\mathbb{Z})$.

Such a simple model presents very interesting dynamical phenomena and has been attracting much attention in the literature. See, e.g., the surveys [12, 13]. Before going to our topic, let us recall some interesting phenomena for (1.1).

The first one is from You [27]. The net flux (or Calabi invariant) of system (1.1) is given by the mean value of $p(t)$. When this is zero, it is shown in [27] that the Poincaré map of (1.1) satisfies the hypotheses of the Moser twist theorem [10, 14, 23] for large enough $\dot{x}$, and there are infinitely many invariant circles for $\dot{x}$ large. When the net flux is nonzero, there exist solutions such that $\dot{x}$ are unbounded. These give a portrait for solutions of (1.1) with very high energy.

The second one is an interesting result which is proved by Wiggins [25] and proved again by Hastings and McLeod [6] using a different approach. They show that there are many chaotic solutions of (1.1) in the following sense. For any sequence of positive integers $n_1, n_2, \ldots, n_{2k-1}, n_{2k}, \ldots$, (1.1) has a solution $x(t)$ such that it rotates n_1 times clockwise and then rotates n_2 times counterclockwise, and rotates n_3 times clockwise and then rotates n_4 times counterclockwise, etc. This phenomenon happens in the region of phase space with high, but not too high, energy.

The third one is the chaotic phenomenon obtained from the homoclinic orbit of unforced case. It can be analyzed using the Melnikov method. This deals with the solutions with a suitable energy.

As for the present paper, we are interested in the stability and twist character of the periodic solution of (1.1) which is near the stable equilibrium $x(t) \equiv 0$ of the

*Received by the editors August 8, 2002; accepted for publication (in revised form) February 21, 2003; published electronically October 14, 2003.

http://www.siam.org/journals/sima/35-4/41037.html

†Department of Mathematical Sciences, Tsinghua University, Beijing 100084, People's Republic of China (jzlei@math.tsinghua.edu.cn, xli@bnu.edu.cn, pyan@math.tsinghua.edu.cn, mzhang@math.tsinghua.edu.cn). The fourth author was supported by the National 973 Project of China (1999), NNSF of China (1998), and TRAPOYT-MOE of China (2001).

unforced system. Suppose that the forcing $p(t)$ ensures that (1.1) has 2π-periodic solutions. Then there exists one periodic solution $x_\omega(t)$ such that the L^∞ norm $\|x_\omega\|_\infty$ is smallest among all of the 2π-periodic solutions of (1.1). Such a periodic solution is called the *least amplitude periodic solution*. This corresponds to the stable equilibrium $x(t) \equiv 0$ for the unforced case. A basic problem concerning $x_\omega(t)$, namely, stability, is the main object of this paper.

More generally, let us consider the scalar Newtonian equation

$$\ddot{x} + f(t, x) = 0, \tag{1.2}$$

where $f(t, x)$ is 2π-periodic in t and is sufficiently smooth in (t, x), e.g., $f \in C^{0,4}(\mathbb{R}/2\pi\mathbb{Z} \times \mathbb{R})$. Suppose that $x = u(t)$ is a 2π-periodic solution of (1.2). A basic method to study the stability of $u(t)$ is to consider the third order approximation of (1.2) along the solution $u(t)$:

$$\ddot{x} + a(t)x + b(t)x^2 + c(t)x^3 + \cdots = 0, \tag{1.3}$$

where the coefficients $a(t),\ b(t),\ c(t) \in C(\mathbb{R}/2\pi\mathbb{Z})$ are

$$a(t) = \frac{\partial f}{\partial x}(t, u(t)), \quad b(t) = \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(t, u(t)), \quad c(t) = \frac{1}{6}\frac{\partial^3 f}{\partial x^3}(t, u(t)).$$

(We have transformed the solution $u(t)$ to $x \equiv 0$ in the above equation.) The linear part of (1.3) is the Hill equation:

$$\ddot{x} + a(t)x = 0. \tag{1.4}$$

The stability problem of $x \equiv 0$ (as a periodic solution of (1.3)) has the nonlocal character because (1.3) is a perturbation of (1.4) which cannot be integrated explicitly. Although there are some results for this problem in previous works such as [15, 23] which are based on the twist theorem [14], a breakthrough is Ortega's works [19, 20, 21]. In these papers, he has derived the (first) twist coefficient for the Birkhoff normal form of the Poincaré map of (1.3) when the linearization equation (1.4) is *R-elliptic* and is 4-*elementary* (for definitions, see section 3.2 or [21]). Under an assumption on (1.4) which implies that it is within the first stability zone [24], he obtained some interesting twist criteria for nonlinear equation (1.3). The results obtained there are based on the comparison between the coefficients $b(t)$ and $c(t)$. They have the characteristic that no small parameters are involved. An interesting application of his results is on the swing (or the pendulum of variable length)

$$\ddot{x} + \alpha(t)\sin x = 0, \tag{1.5}$$

where $\alpha(t)$ (> 0) is a periodic function. It was proved that the periodic solution $x(t) \equiv 0$ of (1.5) is of twist type (and consequently, is almost stable) and is "almost" equivalent to its linear stability; i.e., the corresponding linearization equation

$$\ddot{x} + \alpha(t)x = 0 \tag{1.6}$$

is elliptic. This result works when (1.6) is in higher order stability zones. Some further development in [17] shows that even when (1.6) is (unstable) parabolic, the nonlinear equation (1.5) may be stable in some cases. See also Liu [9] for a related problem. Compared to (1.1), problem (1.5) is relatively simple, because the periodic solution of

(1.5) is known, i.e., $x(t) \equiv 0$. Another advantage is that the second coefficient $b(t)$ for (1.5) vanishes everywhere. At this moment, it is worth mentioning a result of Núñez [16]. He obtained some twist results when $b(t)$ and $c(t)$ in (1.3) can change sign but with a more restricted assumption on the linear equation (1.4) than that obtained by Ortega. In particular, Núñez's results are applicable only to the case that (1.4) is in the first stability zone.

As for the forced pendulum (1.1), three factors need to be considered: (1) The least amplitude periodic solution $x_\omega(t)$ is not a priori known, although we can find in section 2 an upper estimate for $\|x_\omega\|_\infty$ when $\|p\|_1$ is not too large. (2) When we use the third order approximation of (1.1) along $x_\omega(t)$, the coefficients are

$$a(t) = \omega^2 \cos x_\omega(t), \quad b(t) = -\frac{\omega^2}{2} \sin x_\omega(t), \quad c(t) = -\frac{\omega^2}{6} \cos x_\omega(t). \tag{1.7}$$

So the second coefficient $b(t)$ changes sign. (3) A more serious disadvantage is that if ω is large, then $a(t)$ will be large. So the linearization equation (1.4) will be in any higher order stability zone in this case. Thus the results in [16, 19, 21] are not applicable to (1.1). Thus one needs to find new twist criteria in order to study the twist character of the least amplitude periodic solution $x_\omega(t)$ of (1.1).

The paper is organized as follows. In section 2, we will prove the existence of the least amplitude periodic solution $x_\omega(t)$ of (1.1) and give the upper bounds for $x_\omega(t)$ under the assumption on the L^1 norm of the forcing $p(t)$. See Theorem 2.1. In section 3, we will derive the formulas for the twist coefficient of (1.3) when the linearization equation (1.4) is elliptic and is 4-elementary. See (3.23) and (3.24). Then we will give some new twist criteria; cf. Theorem 3.1 and Theorem 3.2. In doing so, we find that it is crucial to find the estimates for the growth of the Floquet solutions of (1.4). This will be realized using several equations derived from the Hill equation (1.4), including the Ermakov–Pinney equation [22] and the Riccati equation. In section 4, we apply the results developed in sections 2 and 3 to obtain the twist character of $x_\omega(t)$ when ω is away from resonance of orders ≤ 4 and satisfies an explicit condition of the form $\|p\|_1 \le P(\omega)$. See Theorem 4.1. A remarkable conclusion is that $p(t)$ may be large in some sense, because $P(\omega)$ is of order $O(\omega^{1/2})$ when ω is bounded from resonance of orders ≤ 4 and $\omega \to \infty$. As a result of the Moser twist theorem, $x_w(t)$ is stable in the sense of Lyapunov. Furthermore, (1.1) has, in a neighborhood of $x_\omega(t)$, infinitely many subharmonics with periods tending to infinity, and infinitely many quasi-periodic solutions.

Throughout this paper the following notation will be used. Denote by $\mathbb{Z}^+ = \{0\} \cup \mathbb{N}$ the set of all nonnegative integers, where $\mathbb{N}$ is the set of positive integers. Let

$$\Omega_0 := \{\omega \in (0,\infty) : \omega \ne p/q \quad \text{for all } p,\ q \in \mathbb{N} \text{ with } 1 \le q \le 4\},$$
$$\Theta_0 := \{\theta \in (0,\infty) : \theta \ne 2n\pi/3 \quad \text{for all } n \in \mathbb{N}\}.$$

For $\ell \in [1,\infty]$ and a 2π-periodic function $r(t)$, we use $\|r\|_\ell$ to denote the L^ℓ norm of $r(t)$ over $[0, 2\pi]$. For two functions $f(t)$ and $g(t)$, $f \ll g$ means that $f(t) \le g(t)$ for all t and $f(t) < g(t)$ holds for t in a subset of positive measure.

2. The least amplitude periodic solution. In this section, we consider the periodic motion of the forced pendulum equation (1.1). When $\omega \notin \mathbb{N}$ and $\|p\|_1$ is not too large in some sense, we will prove that (1.1) has a unique 2π-periodic solution $x = x_\omega(t)$ such that it is near zero and will make the L^∞ norm $\|x_\omega\|_\infty$ be smallest among all of 2π-periodic solutions of (1.1). In this sense, $x_\omega(t)$ is called the *least amplitude periodic solution* of (1.1).

The proof of the following result is elementary.

LEMMA 2.1. *Let α and γ be positive parameters. Then the cubic equation*

$$\alpha X^3 + \gamma = X$$

has a positive root if and only if $27\alpha\gamma^2 \le 4$. In this case, the minimal positive root is given by

$$X = X^*(\alpha,\gamma) = 2(3\alpha)^{-1/2}\cos\frac{\vartheta+\pi}{3}, \quad \left(\vartheta = \arccos\left(\frac{3}{2}\gamma(3\alpha)^{1/2}\right) \in \left(0,\frac{\pi}{2}\right)\right), \tag{2.1}$$

which satisfies

$$X^*(\alpha,\gamma) \le \frac{3}{2}\gamma. \tag{2.2}$$

Now we give the existence of the least amplitude periodic solution.

THEOREM 2.1. *Consider the forced pendulum equation* (1.1). *Assume that $\omega \notin \mathbb{N}$. Let*

$$\alpha = \frac{\int_0^{\omega\pi} |\cos s| \mathrm{d}s}{6|\sin\omega\pi|}, \qquad \gamma = \frac{\|p\|_1}{2\omega|\sin\omega\pi|}. \tag{2.3}$$

If the condition

$$27\alpha\gamma^2 \le 4 \tag{2.4}$$

is satisfied, then equation (1.1) *has a unique 2π-periodic solution $x = x_\omega(t)$ such that $\|x_\omega\|_\infty$ is the smallest among all of 2π-periodic solutions of* (1.1). *Moreover, $x_\omega(t)$ satisfies*

$$\|x_\omega\|_\infty \le X^*(\alpha,\gamma) \le \frac{3\|p\|_1}{4\omega|\sin\omega\pi|}. \tag{2.5}$$

Proof. Let $G(t,s)$ be the Green's function associated with the problem

$$\ddot{x} + \omega^2 x = f(t), \qquad x(t) \text{ is } 2\pi\text{-periodic.}$$

Explicitly,

$$G(t,s) = \begin{cases} \frac{\cos\omega(\pi-t+s)}{2\omega\sin\omega\pi} & \text{if } 0 \le s \le t \le 2\pi, \\ \frac{\cos\omega(\pi-s+t)}{2\omega\sin\omega\pi} & \text{if } 0 \le t \le s \le 2\pi. \end{cases}$$

Now x is a 2π-periodic solution of (1.1) if and only if $x \in C(\mathbb{R}/2\pi\mathbb{Z})$ satisfies

$$x(t) = \int_0^{2\pi} G(t,s)(p(s) + \omega^2(x(s) - \sin x(s)))\mathrm{d}s =: (\mathcal{T}x)(t).$$

The operator $\mathcal{T}$ is a completely continuous operator from $C(\mathbb{R}/2\pi\mathbb{Z})$ (with the uniform norm $\|\cdot\|_\infty$) to itself. It follows from the basic estimate $|y - \sin y| \le \frac{1}{6}|y|^3$ that we have, for any $x \in C(\mathbb{R}/2\pi\mathbb{Z})$,

$$|(\mathcal{T}x)(t)| \le \max_{(t,s)} |G(t,s)|\,\|p\|_1 + \frac{\omega^2}{6}\left(\max_t \int_0^{2\pi} |G(t,s)|\mathrm{d}s\right)\|x\|_\infty^3 = \gamma + \alpha\|x\|_\infty^3,$$

where α and γ are as in (2.3). This yields

$$\|\mathcal{T}x\|_\infty \le \gamma + \alpha\|x\|_\infty^3$$

for all $x \in C(\mathbb{R}/2\pi\mathbb{Z})$.

If α and γ satisfy (2.4), then $\mathcal{T}$ maps the closed ball $\mathcal{B} = \{x \in C(\mathbb{R}/2\pi\mathbb{Z}) : \|x\|_\infty \le X^*(\alpha,\gamma)\}$ into itself. Thus it follows immediately from the Schauder fixed point theorem that $\mathcal{T}$ has a fixed point x_ω in $\mathcal{B}$, namely, x_ω is a 2π-periodic solution of (1.1).

Now we prove the uniqueness. Let $x,\ y \in \mathcal{B}$. Then, using the estimate (2.2), we have

$$|(x(s) - \sin x(s)) - (y(s) - \sin y(s))| \le \frac{1}{2}(X^*(\alpha,\gamma))^2|x(s) - y(s)| \le \frac{9}{8}\gamma^2|x(s) - y(s)|$$

and

$$\begin{aligned}|(\mathcal{T}x)(t) - (\mathcal{T}y)(t)| &= \left|\int_0^{2\pi} G(t,s)\omega^2((x(s) - \sin x(s)) - (y(s) - \sin y(s)))\mathrm{d}s\right| \\ &\le \frac{9}{8}\omega^2\gamma^2\int_0^{2\pi} |G(t,s)|\,|x(s) - y(s)|\mathrm{d}s.\end{aligned}$$

Hence

$$\|\mathcal{T}x - \mathcal{T}y\|_\infty \le \frac{9}{8}\omega^2\gamma^2\left(\max_t \int_0^{2\pi} |G(t,s)|\mathrm{d}s\right)\|x - y\|_\infty = \frac{27}{4}\alpha\gamma^2\|x - y\|_\infty$$

for all $x,\ y \in \mathcal{B}$. Thus, if the strict inequality in condition (2.4) is satisfied, we know that $\mathcal{T} : \mathcal{B} \to \mathcal{B}$ is actually a strict contraction. So $\mathcal{T}$ has a unique fixed point x_ω in $\mathcal{B}$.

Note that if $27\alpha\gamma^2 = 4$, one can also obtain the uniqueness from the proof above, although $\mathcal{T}$ may not be a strict contraction.

By the uniqueness of the 2π-periodic solution of (1.1) in $\mathcal{B}$, we know that $\|x_\omega\|_\infty$ is smaller than other possible 2π-periodic solutions of (1.1). □

Remark 2.1. (1) The existence condition (2.4) can be expressed as

$$\|p\|_1 \le \frac{4\sqrt{2}}{3}\frac{\omega|\sin\omega\pi|^{3/2}}{\left(\int_0^{\pi\omega}|\cos s|\mathrm{d}s\right)^{1/2}} =: P_1(\omega). \tag{2.6}$$

Note that when ω is bounded away from resonance, i.e., when $\mathrm{dist}\,(w,\mathbb{Z}^+) \ge \varepsilon_0 > 0$, then

$$P_1(\omega) = O(\omega^{1/2}) \qquad \text{as } w \to +\infty.$$

It follows now from (2.2), (2.3), (2.5), and (2.6) that

$$\|x_\omega\|_\infty \le Q(\omega) := \frac{(2|\sin\omega\pi|)^{1/2}}{\left(\int_0^{\pi\omega}|\cos s|\mathrm{d}s\right)^{1/2}} = O(\omega^{-1/2}) \quad \text{as } w \to +\infty. \tag{2.7}$$

A more precise upper bound for $x_\omega(t)$ can be derived from (2.1) and (2.5).

(2) The existence of periodic solutions of (1.1) is a central problem in nonlinear analysis; see [11, 13]. However, when we study the twist character, it is necessary to give a quantitative estimate to the periodic solution. In previous works such as [16, 18], this is done using the method of upper and lower solutions [2]. However, this method is applicable to (1.1) when the frequency ω is small. Although the estimate in Theorem 2.1 is not optimal, it will yield a satisfactory result in section 4 when we study the twist character of $x_\omega(t)$.

3. Twist results basing on the third order approximation. For the forced pendulum equation (1.1), we will consider the case that ω is bounded away from the resonance and ω is large. Recall from (1.7) and (2.7) that $a(t) = a_\omega(t) = \omega^2 \cos x_\omega(t) > 0$, $c(t) = c_\omega(t) = -(\omega^2/6)\cos x_\omega(t) < 0$, and $b(t) = b_\omega(t) = -(\omega^2/2)\sin x_w(t)$ changes sign, and all of them will be large in general when ω is so. In particular, $\lambda = 0$ is not within the first stability zone (defined at the end of the next subsection) of the linearization equation

$$\ddot{x} + (\lambda + a_\omega(t))x = 0.$$

We will follow [16, 19, 20, 21] to derive some new twist results for (1.3) which are applicable to the forced pendulum equation. The results obtained in this section are of independent interest, because we are mainly concerned with the case of higher order stability zones for the linearization equations. In doing so, we mostly concentrate on linearization equation (1.4). Since (1.4) cannot be integrated explicitly, a lot of theories for the Hill equations and their variants will be engaged in the discussion below.

3.1. Rotation numbers and Floquet multipliers. We consider the Hill equation (1.4). Let $x = r\cos\psi$ and $\dot{x} = -r\sin\psi$ in (1.4). Then the equation for $\psi(t)$ is

$$\dot{\psi} = \sin^2\psi + a(t)\cos^2\psi. \tag{3.1}$$

Since the right-hand side of (3.1) is periodic in both t and ψ, it is well known that the *rotation number* of (1.4),

$$\rho = \rho(a) = \lim_{t\to\infty} \psi(t)/t,$$

does exist and is independent of the choice of the solution $\psi(t)$ of (3.1) in defining the rotation number. See Hartman [5].

Some well-known properties on rotation numbers are listed in the following lemma.

LEMMA 3.1.

(1) $0 \le \rho(a) < \infty$.

(2) *$\rho(a)$ is continuous in $a(\cdot)$ with respect to the L^1 norm of a's.*

(3) *$\rho(a)$ is monotone with respect to $a(t)$. More precisely, if $a_1 \ll a_2$, then $\rho(a_1) < \rho(a_2)$.*

(4) *When $a(t) \equiv \omega^2$ is a constant, then $\rho(a) = \omega$.*

Some further properties on rotation numbers and their applications can be found in [28].

Rewrite (1.4) as an equivalent planar, linear system:

$$\dot{x} = y, \qquad \dot{y} = -a(t)x. \tag{3.2}$$

Let M be the Poincaré matrix associated with (3.2). The eigenvalues $\lambda_{1,2}$ of M are called the *Floquet multipliers* of (1.4). Since $\det M = 1$, $\lambda_1 \cdot \lambda_2 = 1$. As usual, we say that (1.4) is *elliptic*, *parabolic*, and *hyperbolic* if $\lambda_{1,2} \in S^1\backslash\{\pm 1\}$, $\lambda_{1,2} = \pm 1$, and $|\lambda_{1,2}| \ne 1$, respectively.

In the following we are interested only in the elliptic case, which can also be described using rotation numbers.

LEMMA 3.2. *Equation* (1.4) *is elliptic if and only if $\rho = \rho(a) \notin \frac{1}{2}\mathbb{Z}^+$. In this case, the Floquet multipliers of* (1.4) *are given by $\lambda_{1,2} = e^{\pm i\theta}$, where*

$$\theta = 2\pi\rho. \tag{3.3}$$

Proof. An elementary proof for this fact is given in [4]. □

Note that the θ in the expression of the Floquet multipliers is only defined by modulo 2π. However, we will always take θ as in (3.3) when (1.4) is elliptic.

Let $n \in \mathbb{N}$. If θ is contained in the interval $((n-1)\pi, n\pi)$, we say that 0 is in the nth *stability zone* of (1.4) (see [24]), or simply that $a(t)$ is in the nth stability zone. This is equivalent to the fact that $\lambda = 0$ is in the nth spectrum interval of the parameterized Hill equation

$$\ddot{x} + (\lambda + a(t))x = 0.$$

3.2. Ellipticity and twist coefficients. Let $\Psi(t) = \phi_1(t) + i\phi_2(t)$ be the (complex) solution of (1.4) with the initial data $\Psi(0) = 1$ and $\dot{\Psi}(0) = i$, where ϕ_1 and ϕ_2 are, respectively, the real and imaginary parts of Ψ. Now the Poincaré matrix of (3.2) is

$$M = \begin{pmatrix} \phi_1(2\pi) & \phi_2(2\pi) \\ \dot{\phi}_1(2\pi) & \dot{\phi}_2(2\pi) \end{pmatrix}.$$

When (1.4) is elliptic, it is easy to see that $\Psi(t) \neq 0$ for all t. Thus it can be written in the form $\Psi(t) = r(t)e^{i\varphi(t)}$, where $r,\ \varphi \in C^2(\mathbb{R})$, $r(t) > 0$, and they have initial data

$$r(0) = 1, \quad \dot{r}(0) = 0, \quad \varphi(0) = 0, \quad \dot{\varphi}(0) = 1. \tag{3.4}$$

We say that an elliptic equation (1.4) is 4-*elementary* if its multipliers $\lambda = e^{\pm i\theta}$ satisfy $\lambda^q \neq 1$ for $1 \leq q \leq 4$. This is simply equivalent to

$$\rho = \theta/(2\pi) \in \Omega_0, \tag{3.5}$$

where Ω_0 is as in the end of section 1.

We say that (1.4) is R-*elliptic* (with respect to $e^{i\theta}$) if (1.4) is elliptic and

$$\Psi(t + 2\pi) \equiv e^{i\theta}\Psi(t). \tag{3.6}$$

In this case, the Poincaré matrix M is simply a rigid rotation with the angle θ. Furthermore, $r(t)$ is 2π-periodic and $\varphi(t)$ is strictly increasing (see (3.20) below) and satisfies

$$\varphi(t + 2\pi) \equiv \varphi(t) + \theta. \tag{3.7}$$

This gives an expression for θ in the Floquet multipliers using the function $\varphi(t)$. In particular, $\varphi(0) = 0$ and $\varphi(2\pi) = \theta$. Condition (3.6) means also that $\Psi(t)$ is a Floquet solution with the multiplier $e^{i\theta}$. For another expression of θ, see (3.22) below.

From now on we consider the nonlinear equation (1.3), where $a,\ b,\ c \in C(\mathbb{R}/2\pi\mathbb{Z})$. At the moment, we assume that $a \in C(\mathbb{R}/2\pi\mathbb{Z})$ is such that (1.4) is R-elliptic. However, we will not confine ourself to the case that $a(\cdot)$ is in the first stability zone.

Let

$$\hat{F}(x_0, y_0) = (\hat{F}_1(x_0, y_0), \hat{F}_2(x_0, y_0))$$

be the Poincaré map of (1.3). Write $\hat{F}$ in the complex form, with $z = x_0 + iy_0$,

$$F(z, \bar{z}) = \hat{F}_1\left((z + \bar{z})/2, (z - \bar{z})/(2i)\right) + i\hat{F}_2\left((z + \bar{z})/2, (z - \bar{z})/(2i)\right).$$

When $\lambda = e^{i\theta}$ is 4-elementary, it is well known that $F(z, \bar{z})$ is C^∞ conjugate, in the group of area-preserving diffeomorphisms, to

$$N(z, \bar{z}) = \lambda(z + i\beta|z|^2 z + \cdots),$$

where $\beta \in \mathbb{R}$. Such a form of $N(z, \bar{z})$ is called the *Birkhoff normal form* of F. The coefficient β, which depends only on a, b, c and is invariant under conjugacies of area-preserving diffeomorphisms, is called the *(first) twist coefficient* of (1.3). When $\beta \neq 0$, we say that the solution $x = 0$ of (1.3) (as a 2π-periodic solution) is of *twist type*. In this case, the Moser twist theorem is applicable and will yield the typical dynamical behavior near 0, as mentioned in the introduction.

Under the assumption that (1.4) is 4-elementary and is R-elliptic (cf. (3.5) and (3.6)), Ortega [19, 21] uses the expansion of $F(z, \bar{z})$ at $z = 0$ to have derived the formula of the twist coefficient β. See formula (2.6) and Proposition 4.4 of [21]. If one exploits the notation above β can be written as

$$\begin{aligned}\beta = &-\frac{3}{8}\int_{[0,2\pi]} c(t)r^4(t)\mathrm{d}t + \iint_{[0,2\pi]^2} b(t)b(s)r^3(t)r^3(s)\chi_1(|\varphi(t) - \varphi(s)|)\mathrm{d}t\mathrm{d}s \\ &+\frac{3}{16}\cot\frac{\theta}{2}\left|\int_{[0,2\pi]} b(t)r^3(t)e^{-i\varphi(t)}\mathrm{d}t\right|^2 + \frac{1}{16}\cot\frac{3\theta}{2}\left|\int_{[0,2\pi]} b(t)r^3(t)e^{3i\varphi(t)}\mathrm{d}t\right|^2,\end{aligned} \tag{3.8}$$

where

$$\chi_1(x) = \frac{1}{8}(2 + \cos 2x)\sin x = \frac{3\sin x - 2\sin^3 x}{8}, \qquad x \in [0, \theta]. \tag{3.9}$$

Formula (3.8) can be written in a more compact form [29]:

$$\beta = -\frac{3}{8}\int_{[0,2\pi]} c(t)r^4(t)\mathrm{d}t + \iint_{[0,2\pi]^2} b(t)b(s)r^3(t)r^3(s)\chi_2(|\varphi(t) - \varphi(s)|)\mathrm{d}t\mathrm{d}s, \tag{3.10}$$

where the kernel $\chi_2(\cdot)$ is

$$\chi_2(x) = \frac{3}{16}\frac{\cos(x - \theta/2)}{\sin(\theta/2)} + \frac{1}{16}\frac{\cos 3(x - \theta/2)}{\sin(3\theta/2)}, \qquad x \in [0, \theta]. \tag{3.11}$$

Roughly speaking, the twist coefficient β is the sum of a linear functional of $c(\cdot)$ and a quadratic form of $b(\cdot)$. However, the kernels in the functionals are dependent upon the solutions $r(t)$ and $\varphi(t)$ of the Hill equation (1.4) in a complicated way. The properties of β are far from being understood completely. For discussions on some hidden mystery of it, see the recent work [29]. Some applications of Ortega's works can be found in [8, 18].

Suppose now that (1.4) is elliptic (not necessarily R-elliptic) and 4-elementary. Ortega has shown in [19, Proposition 7] that there exist some $t_0 \in \mathbb{R}$ and $\sigma > 0$ such that the change of variables

$$\xi = x, \qquad \tau = \sigma(t - t_0) \tag{3.12}$$

will transform (1.4) into an R-elliptic equation,

$$\frac{d^2\xi}{\mathrm{d}\tau^2} + a^*(\tau)\xi = 0. \tag{3.13}$$

Correspondingly, (3.12) transforms (1.3) into

$$\frac{\mathrm{d}^2\xi}{\mathrm{d}\tau^2} + a^*(\tau)\xi + b^*(\tau)\xi^2 + c^*(\tau)\xi^3 + \cdots = 0. \tag{3.14}$$

Here

$$a^*(\tau) = \sigma^{-2}a(t_0 + \sigma^{-1}\tau), \quad b^*(\tau) = \sigma^{-2}b(t_0 + \sigma^{-1}\tau), \quad c^*(\tau) = \sigma^{-2}c(t_0 + \sigma^{-1}\tau),$$

and the new period is $T^* = 2\pi\sigma$.

Note that the R-ellipticity condition for (3.13) may be with respect to $e^{-i\theta}$. However, this can be transformed into the R-ellipticity defined as in (3.6) by reversing time. Thus we always assume that (3.13) is R-elliptic as in (3.6) (with 2π replaced trivially by the new period T^*).

If we introduce $\Psi^*(\tau) = r^*(\tau)e^{i\varphi^*(\tau)}$ for the R-elliptic and 4-elementary equation (3.13) as before, then the first twist coefficient of (3.14) is given by (cf. (3.8))

$$\begin{aligned}\beta^* = &-\frac{3}{8}\int_0^{T^*} c^*(\tau){r^*}^4(\tau)\mathrm{d}\tau + \iint_{[0,T^*]^2} b^*(\tau)b^*(\zeta){r^*}^3(\tau){r^*}^3(\zeta)\chi_1(|\varphi^*(\tau) - \varphi^*(\zeta)|)\mathrm{d}\tau\mathrm{d}\zeta \\ &+ \frac{3}{16}\cot\frac{\theta}{2}\left|\int_0^{T^*} b^*(\tau){r^*}^3(\tau)e^{-i\varphi^*(\tau)}\mathrm{d}\tau\right|^2 + \frac{1}{16}\cot\frac{3\theta}{2}\left|\int_0^{T^*} b^*(\tau){r^*}^3(\tau)e^{3i\varphi^*(\tau)}\mathrm{d}\tau\right|^2.\end{aligned} \tag{3.15}$$

Note that (3.13) has the same θ as (1.4). A basic relationship between β for (1.3) and β^* for (3.14) is

$$\operatorname{sign}\beta = \operatorname{sign}\beta^*.$$

Thus we are mainly concerned with the estimates of β^* in the following.

Let us use the solutions of (1.4), not that of the transformed equation (3.13), to express the coefficient β^*. Set

$$r(t) = \sigma^{-1/2}r^*(\sigma(t - t_0)), \qquad \varphi(t) = \varphi^*(\sigma(t - t_0)). \tag{3.16}$$

Using initial conditions (3.4) for $r^*(\tau)$ and $\varphi^*(\tau)$, we see that $r(t)$ and $\varphi(t)$ satisfy

$$r(t_0) = \sigma^{-1/2}, \quad \dot r(t_0) = 0, \quad \varphi(t_0) = 0, \quad \dot\varphi(t_0) = \sigma. \tag{3.17}$$

Since $r^*(\tau)$ is T^*-periodic, $r(t)$ is 2π-periodic. Another fact is that $\Psi(t) := r(t)e^{i\varphi(t)} = \sigma^{-1/2}\Psi^*(\sigma(t - t_0))$ satisfies (1.4). Substituting this into (1.4), we have

$$0 = \ddot\Psi(t) + a(t)\Psi(t) = e^{i\varphi}\left[(\ddot r - r\dot\varphi^2 + a(t)r) + i(2\dot r\dot\varphi + r\ddot\varphi)\right].$$

Thus

$$2\dot r\dot\varphi + r\ddot\varphi = 0, \qquad \ddot r - r\dot\varphi^2 + a(t)r = 0. \tag{3.18}$$

From the first equation, we have

$$\dot\varphi = c/r^2$$

for some constant c. Using the initial data (3.17), one sees that $c = 1$. By the second equation of (3.18), $r(t)$ satisfies the so-called Ermakov–Pinney equation [22]

$$\ddot{r} + a(t)r = \frac{1}{r^3}, \tag{3.19}$$

while $\varphi(t)$ satisfies

$$\dot{\varphi} = \frac{1}{r^2}. \tag{3.20}$$

In conclusion, the function $r(t)$ in (3.16) is a positive 2π-periodic solution of (3.19). In Lemma 3.3 below, we will prove that the Ermakov–Pinney equation (3.19) has a unique positive 2π-periodic solution $r(t)$ when (1.4) is elliptic. As $\varphi(t)$ satisfies $\varphi(t_0) = 0$ and $\varphi(t_0 + 2\pi) = \theta$ and $r(t)$ is 2π-periodic, we obtain from (3.20) that

$$\varphi(t) = \int_{t_0}^{t} \frac{\mathrm{d}t}{r^2(t)} \qquad \text{for all } t, \tag{3.21}$$

and

$$\int_{t_0}^{t_0+2\pi} \frac{\mathrm{d}t}{r^2(t)} = \int_{0}^{2\pi} \frac{\mathrm{d}t}{r^2(t)} = \theta. \tag{3.22}$$

The latter implies that $\varphi(t)$ also satisfies (3.7) for all t.

Exploiting these $r(t)$ and $\varphi(t)$, we make use of the change of variables $\tau = \sigma(t - t_0)$ in (3.15) and obtain the following "explicit" formula for β^*.

PROPOSITION 3.1. *The twist coefficient β^* can be rewritten as*

$$\begin{aligned}\beta^* = \sigma\Bigg[&-\frac{3}{8}\int_{t_0}^{t_0+2\pi} c(t)r^4(t)\mathrm{d}t + \iint_{[t_0,t_0+2\pi]^2} b(t)b(s)r^3(t)r^3(s)\chi_1(|\varphi(t)-\varphi(s)|)\mathrm{d}t\mathrm{d}s \\ &+\frac{3}{16}\cot\frac{\theta}{2}\left|\int_{t_0}^{t_0+2\pi} b(t)r^3(t)e^{-i\varphi(t)}\mathrm{d}t\right|^2 + \frac{1}{16}\cot\frac{3\theta}{2}\left|\int_{t_0}^{t_0+2\pi} b(t)r^3(t)e^{3i\varphi(t)}\mathrm{d}t\right|^2\Bigg],\end{aligned} \tag{3.23}$$

where $r(t)$ and $\varphi(t)$ are as above, while the constant σ is related with the critical value $r(t_0)$ (see [19, Proposition 7]*) and is not of importance in the estimates below.*

Analogously, we obtain from (3.10) another "explicit" formula for β^*.

PROPOSITION 3.2.

$$\begin{aligned}\beta^* = \sigma\Bigg[&-\frac{3}{8}\int_{t_0}^{t_0+2\pi} c(t)r^4(t)\mathrm{d}t \\ &+ \iint_{[t_0,t_0+2\pi]^2} b(t)b(s)r^3(t)r^3(s)\chi_2(|\varphi(t)-\varphi(s)|)\mathrm{d}t\mathrm{d}s\Bigg],\end{aligned} \tag{3.24}$$

where $r(t)$, $\varphi(t)$, and σ are as in Proposition 3.1*, and $\chi_2(\cdot)$ is given by* (3.11)*.*

Note from (3.23) and (3.24) that it is important to estimate the growth of $r(t)$, the (unique) positive 2π-periodic solution of the Ermakov–Pinney equation (3.19), in estimating β^*. This will be done in subsection 3.5.

3.3. Discussion on the kernels. In this subsection, we estimate the kernels $\chi_i(\cdot)$, $i = 1, 2$, in (3.23) and (3.24).

The estimate for $\chi_1(x)$ is simple:

$$\max_{x\in[0,\theta]} |\chi_1(x)| \le \sqrt{2}/8.$$

Combining this with the third and fourth terms in formula (3.23), we introduce the following function of θ:

$$K_1(\theta) := \frac{\sqrt{2}}{8} + \max\left\{-\frac{3}{16}\cot\frac{\theta}{2},\, 0\right\} + \max\left\{-\frac{1}{16}\cot\frac{3\theta}{2},\, 0\right\}. \tag{3.25}$$

Note that $K_1(\theta)$ is well defined in $\theta \in \Theta_0$ and is 2π-periodic in θ.

Sometimes, we will use (3.24) to estimate the twist coefficient β^*. We can rewrite the kernel $\chi_2(x)$ in another form:

$$\chi_2(x) = \frac{2\cos^3(x-\theta/2) + 3\cos\theta\cos(x-\theta/2)}{8\sin(3\theta/2)}, \qquad x \in [0, \theta]. \tag{3.26}$$

Let

$$K_2(\theta) := \max_{x\in[0,\theta]} |\chi_2(x)|.$$

Then $K_2(\theta)$ is defined in $\theta \in \Theta_0$ and is 2π-periodic in θ. Using the expression (3.26), we see that

$$K_2(\theta) = \begin{cases} |2 + 3\cos\theta|/(8|\sin(3\theta/2)|) & \text{if } \theta \in (0, 2\pi/3) \cup (4\pi/3, 2\pi), \\ |\cos\theta|\sqrt{-2\cos\theta}/(8|\sin(3\theta/2)|) & \text{if } \theta \in (2\pi/3, 4\pi/3). \end{cases}$$

For most of θ, $K_1(\theta) < K_2(\theta)$. However, $K_1(\theta) > K_2(\theta)$ when θ tends from left to $2n\pi/3$, $n \in \mathbb{N}$. Define

$$K(\theta) = \min\{K_1(\theta),\, K_2(\theta)\}, \qquad \theta \in \Theta_0. \tag{3.27}$$

By (3.26), we have

$$K(\theta) \le \frac{5}{8|\sin(3\theta/2)|}, \qquad \theta \in \Theta_0.$$

Both of the functions $K_1(\theta)$ and $K(\theta)$ are increasing for θ in any interval from Θ_0. In particular, we have

$$\max_{\theta\in[\theta_1,\theta_2]} K_1(\theta) = K_1(\theta_2) \le \frac{5}{8|\sin(3\theta_2/2)|} \tag{3.28}$$

when $\theta_1,\ \theta_2$, with $\theta_1 \le \theta_2$, are from the same interval of Θ_0. The graph of $K(\theta)$ is as in Figure 1.

3.4. Estimating periodic solutions of the Ermakov–Pinney equation. In this subsection, we concentrate on estimating the growth of the positive 2π-periodic solution $r(t)$ of (3.19). This is the crucial estimate to be used in the next subsection where we estimate the twist coefficients.

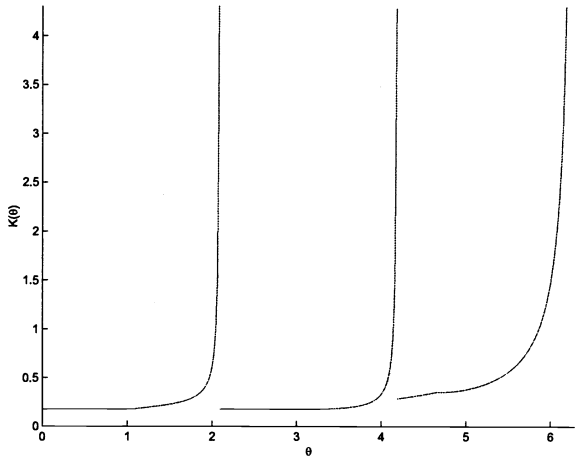


FIG. 1. *The graph of* $K(\theta)$.

LEMMA 3.3. *Assume that* $a \in C(\mathbb{R}/2\pi\mathbb{Z})$ *such that* (1.4) *is elliptic with the Floquet multipliers* $e^{\pm i\theta}$. *Then the Ermakov–Pinney equation* (3.19) *has a unique positive* 2π*-periodic solution, denoted by* $r(t)$. *Moreover,* $r(t)$ *satisfies* (3.22). *(This gives another expression for* θ *of* (1.4) *using the function* $r(t)$ *associated with* (1.4).*)*

Proof. The existence result of a positive periodic solution $r(t)$ of (3.19) has been explained in subsection 3.2 using Floquet solutions, where the connection between the Hill equation (1.4) and the Ermakov–Pinney equation (3.19) is used.

Now we prove the uniqueness result. Let $r_1(t)$ be another positive 2π-periodic solution of (3.19). Take t_1 as a critical point of $r_1(t)$, i.e., $\dot{r}_1(t_1) = 0$. Define $\varphi_1(t)$ by

$$\varphi_1(t) = \int_{t_1}^{t} \frac{\mathrm{d}s}{r_1^2(s)};$$

cf. (3.21). Then $(r_1(t), \varphi_1(t))$ satisfies the system (3.19)–(3.20). So $\Psi_1(t) = r_1(t)e^{i\varphi_1(t)}$ is a solution of (1.4). Moreover, as $r_1(t)$ is 2π-periodic, we obtain from the definition of $\varphi_1(t)$ that $\varphi_1(t+2\pi) - \varphi_1(t)$ is independent of t and equal to

$$\theta_1 = \int_0^{2\pi} \frac{\mathrm{d}s}{r_1^2(s)}.$$

Thus $\Psi_1(t)$ satisfies $\Psi_1(t+2\pi) \equiv e^{i\theta_1}\Psi_1(t)$ and is also a Floquet solution of (1.4) with the multiplier $e^{i\theta_1}$. By the uniqueness result for Floquet solutions, we have

$$\theta_1 = \theta + 2m\pi, \qquad r_1(t) = c\, r(t)$$

for some $m \in \mathbb{Z}$ and some $c > 0$. Since both $r(t)$ and $r_1(t)$ satisfy (3.19), we have necessarily that $c = 1$. Thus $r_1(t) \equiv r(t)$. This proves the uniqueness result and (3.22) is satisfied. □

Since the positive 2π-periodic solution $r(t)$ of (3.19) is uniquely determined by $a(t)$ when (1.4) is elliptic, we know that the minimum and the maximum of $r(t)$ are also uniquely determined by $a(t)$. These facts have been generalized in [1, 3, 26] to Ermakov–Pinney-type equations when they study the nonresonance problem of equations with singularities.

Now we give the estimates of the L^4 norm $\|r\|_4$ of $r(t)$. The estimate for lower bounds of $\|r\|_4$ is made simple by using the constraint (3.22).

LEMMA 3.4. *Assume that $r(t)$ is a positive 2π-periodic function satisfying* (3.22). *Then, for any $\ell \geq 2$,*

$$\|r\|_\ell \geq (2\pi)^{1/\ell}(2\pi/\theta)^{1/2}.$$

Proof. Let $\ell \geq 2$. Set the exponents $p = (2+\ell)/2$, $q = (2+\ell)/\ell$, and $\alpha = 2\ell/(2+\ell)$. Using the Hölder inequality, we have

$$\begin{aligned} 2\pi &= \int_0^{2\pi} 1\mathrm{d}t = \int_0^{2\pi} r^\alpha \cdot r^{-\alpha}\mathrm{d}t \\ &\leq \left(\int r^{\alpha p}\mathrm{d}t\right)^{1/p}\left(\int r^{-\alpha q}\mathrm{d}t\right)^{1/q} \\ &= \left(\int r^{\ell}\mathrm{d}t\right)^{1/p}\left(\int r^{-2}\mathrm{d}t\right)^{1/q} \\ &= \theta^{1/q}\left(\int r^{\ell}\mathrm{d}t\right)^{1/p}. \end{aligned}$$

Thus

$$\|r\|_\ell \geq (2\pi)^{p/\ell}/\theta^{p/(q\ell)},$$

which is just the inequality described in the lemma. □

In order to estimate the upper bounds of $\|r\|_4$, we need the following comparison result for Riccati equations.

LEMMA 3.5. *Assume that $a_j \in C(\mathbb{R})$. Let $\xi_j(t; z_j)$ be (real) solutions of equations*

$$\dot{x} = x^2 + a_j(t), \qquad j = 1, 2,$$

satisfying $\xi_j(0) = z_j$. If $a_1(t) \geq a_2(t)$ and $z_1 \geq z_2$, then

$$\xi_1(t, z_1) \geq \xi_2(t, z_2) \qquad \text{for all } t \in [0, t^*),$$

where t^* is such that $\xi_j(t, z_1) < +\infty$ for $t \in [0, t^*)$, $j = 1, 2$.

In the next lemma, we use $\langle a, b\rangle$ to denote the interval $[a, b]$ for $a \leq b$ or the interval $[b, a]$ for $a \geq b$.

LEMMA 3.6. *Let $M_0 = (0, \frac{1}{4})$, $M_n^+ = (\frac{n}{2}, \frac{n}{2} + \frac{1}{4})$, and $M_n^- = (\frac{n}{2} - \frac{1}{4}, \frac{n}{2})$ for $n \in \mathbb{N}$. Assume that $a \in C(\mathbb{R}/2\pi\mathbb{Z})$ satisfies*

$$\sigma_1^2 \leq a(t) \leq \sigma_2^2 \qquad \text{for all } t, \tag{3.29}$$

where σ_1 and σ_2 satisfy one of the following conditions:

$$\sigma_1,\ \sigma_2 \in M_0, \tag{3.30}$$

$$\sigma_1,\ \sigma_2 \in M_n^+, \qquad n \in \mathbb{N}, \tag{3.31}$$

$$\sigma_1,\ \sigma_2 \in M_n^-, \qquad n \in \mathbb{N}. \tag{3.32}$$

Then we have the following estimates.

(1) *Equation* (1.4) *is elliptic with the Floquet multipliers* $e^{\pm i\theta}$, *where* θ *satisfies*

$$2\pi\sigma_1 \le \theta \le 2\pi\sigma_2. \tag{3.33}$$

(2) *The positive* 2π*-periodic solution* $r(t)$ *of* (3.19) *satisfies*

$$r(t) \in \left\langle (\sigma_1\sigma_2 \tan 2\pi\sigma_1 \cot 2\pi\sigma_2)^{-1/4},\ (\sigma_1\sigma_2 \cot 2\pi\sigma_1 \tan 2\pi\sigma_2)^{-1/4} \right\rangle \quad \textit{for all } t, \tag{3.34}$$

and

$$\int_0^{2\pi} r^4(t)\mathrm{d}t \in \left\langle \frac{2\pi}{\sigma_1\sigma_2}\frac{\tan 2\pi\sigma_2}{\tan 2\pi\sigma_1},\ \frac{2\pi}{\sigma_1\sigma_2}\frac{\tan 2\pi\sigma_1}{\tan 2\pi\sigma_2} \right\rangle. \tag{3.35}$$

Proof. Conclusion (1) follows immediately from Lemma 3.1. Conclusion (2) will be established using the connection between the Hill equation and the Riccati equation [7]. Let $r(t)$ be as in the lemma. Suppose that t_0 is a critical point of $r(t)$. As in the proof of Lemma 3.3, let $\varphi(t)$ be defined by

$$\varphi(t) = \int_{t_0}^{t} \frac{\mathrm{d}s}{r^2(s)}.$$

Then $\Psi(t) = r(t)e^{i\varphi(t)}$ is a solution of (1.4). Define

$$w(t) = -\frac{\dot\Psi(t)}{\Psi(t)} = -\frac{\dot r}{r} - \frac{i}{r^2},$$

which is 2π-periodic. It is well known that $w(t)$ is a (complex) solution of the Riccati equation

$$\dot w = w^2 + a(t). \tag{3.36}$$

Now the estimates (3.34) are reduced to estimate the critical values $r(t_0) =: r_0 > 0$ of $r(t)$. Without loss of generality, we assume here that $t_0 = 0$. Let $w(t; z)$ be the solution of (3.36) satisfying $w(0; z) = z$. When the values are considered on the Riemannian sphere, $w(t; z)$ is well defined for all $t \in \mathbb{R}$. See [7, Chapter 4]. Since the coefficient $a(t)$ is real, it is well known that the Poincaré map of (3.36) is a Möbius transformation

$$T(z) = w(2\pi; z) = \frac{az + b}{cz + d},$$

where $a,\ b,\ c,\ d$ are real. The fixed points z_0 of T correspond to initial values of 2π-periodic solutions of (3.36). In our situation, $z_0 = -i/r_0^2$ is a fixed point of T. Since z_0 is purely imaginary, we know that $a = d$ and $b/c < 0$. Note that if $a = d = 0$, then $w(2\pi; 0) = T(0) = \infty$, which is impossible in our situation (see (3.37) below). Let us assume that $a = d = 1$ for simplicity. In this case, we know that r_0 is given by $r_0 = (-c/b)^{1/4}$. So the estimate for r_0 follows from estimating the coefficients b and c in the Poincaré map T of (3.36).

We will first estimate $b = T(0)$. Then $c = -1/T^{-1}(\infty)$ can be obtained in a similar way. The estimates will be done under the following assumption:

$$(0 <)\ \frac{n}{2} - \frac{1}{4} < \sigma_1 \le \sigma_2 < \frac{n}{2} + \frac{1}{4}.$$

It is easy to see that the condition above implies that

$$\frac{(2k-1)\pi}{2\sigma_2} \le \frac{(2k-1)\pi}{2\sigma_1} < \frac{(2k+1)\pi}{2\sigma_2} \qquad \text{for all } 1 \le k \le n.$$

Consider the equations

$$\dot{w}_1 = w_1^2 + \sigma_1^2,$$

$$\dot{w}_2 = w_2^2 + \sigma_2^2.$$

Let $w_1(t) = \sigma_1 \tan \sigma_1 t$ and $w_2(t) = \sigma_2 \tan \sigma_2 t$ be solutions of above equations with initial data $w_j(0) = 0$, respectively.

We will construct intervals of I_k of $[0, 2\pi]$ such that Lemma 3.5 is applicable on each I_k and $2\pi \in I_n$. Thus $b = w(2\pi; 0) \in [\sigma_1 \tan 2\pi\sigma_1,\, \sigma_2 \tan 2\pi\sigma_2]$ by Lemma 3.5. Denote $w(t) = w(t; 0)$. Then $w(t)$ is real because the initial data $w(0; 0) = 0$ and the coefficient $a(t)$ are real. Set

$$I_0 = [0, \pi/(2\sigma_2)),$$

and for $k = 1, \dots, n$,

$$I_k = \left(\frac{(2k-1)\pi}{2\sigma_1},\, \frac{(2k+1)\pi}{2\sigma_2}\right) \cap [0, 2\pi], \quad J_k = \left(\frac{(2k-1)\pi}{2\sigma_2},\, \frac{(2k-1)\pi}{2\sigma_1}\right) \cap [0, 2\pi].$$

We claim that there exist $t_k^* \in \bar{J}_k$ (the closure of J_k) such that

$$\lim_{t \to t_k^* \mp 0} w(t_k^*) = \pm\infty.$$

For example, the existence of t_1^* can be explained as below. From Lemma 3.5, it is easy to see that $w_1(t) \le w(t) \le w_2(t)$ for any $t \in I_0$. Thus $w(t) < +\infty$, $t \in I_0$. If $w(t) < +\infty$ for any $t \in J_1$, then for all $t \in I_0 \cup J_1$, we have $w(t) > w_1(t)$. On the other hand, since $\lim_{t \to (\pi/2\sigma_1) - 0} w_1(t) = +\infty$, we have $\lim_{t \to (\pi/2\sigma_1) - 0} w(t) = +\infty$. Thus we can always choose a $t_1^* \in \bar{J}_1$ such that $\lim_{t \to t_1^* - 0} w(t_1^*) = +\infty$. Consequently, $\lim_{t \to t_1^* + 0} w(t_1^*) = -\infty$. Let $L_1 := (\frac{\pi}{2\sigma_1}, t_1^*)$ and $L_2 := (t_1^*, \frac{\pi}{2\sigma_2})$. Then $w(t) \ge w_1(t)$ for all $t \in L_1$. Since

$$\lim_{t \to t_1^* + 0} w(t) = -\infty \le \lim_{t \to t_1^* + 0} w_2(t),$$

we have $w(t) \le w_2(t)$ for $t \in L_2$. Next, by

$$\lim_{t \to \pi/2\sigma_1 + 0} w_1(t) = -\infty \le w(\pi/2\sigma_1) \le w_2(\pi/2\sigma_2),$$

we have $w_1(t) \le w(t) \le w_2(t)$ for $t \in I_2$. The existence of t_k^* is similar using this argument step by step. Thus we have $t_k^* \in \bar{J}_k$ such that $\lim_{t \to t_k^* \mp 0} w(t_k^*) = \pm\infty$, and $w(t)$ is finite for $t \in [0, 2\pi] \backslash \{t_k\}$. Moreover, let

$$J_k' = \left(\frac{(2k-1)\pi}{2\sigma_2}, t_k^*\right), \quad J_k'' = \left(t_k^*, \frac{(2k-1)\pi}{2\sigma_1}\right), \quad k = 1, 2, \dots, n.$$

Then $[0, 2\pi]$ is divided into intervals $I_0,\ J_1',\ J_1'',\ I_1,\ J_2',\ J_2'', \dots, I_n$ by the points

$$0 < \frac{\pi}{2\sigma_2} < t_1^* < \frac{\pi}{2\sigma_1} < \frac{3\pi}{2\sigma_2} < t_2^* < \frac{3\pi}{2\sigma_1} < \cdots < \frac{(2n-1)\pi}{2\sigma_2} < t_n^* < \frac{(2n-1)\pi}{2\sigma_1} < 2\pi.$$

From the same arguments as above, we have

$$\begin{cases} w_1(t) \le w(t) \le w_2(t), & t \in I_k,\ k = 0, 1, \dots, n, \\ w_1(t) \le w(t), & t \in J'_k,\ k = 1, 2, \dots, n, \\ w(t) \le w_2(t), & t \in J''_k,\ k = 1, 2, \dots, n. \end{cases}$$

Since $2\pi \in I_n$, we have

$$-\infty < \sigma_1 \tan 2\pi\sigma_1 = w_1(2\pi) \le T(0) = w(2\pi; 0) \le w_2(2\pi) = \sigma_2 \tan 2\pi\sigma_2 < +\infty, \tag{3.37}$$

which implies that

$$b = T(0) \in [\sigma_1 \tan 2\pi\sigma_1,\ \sigma_2 \tan 2\pi\sigma_2].$$

Now we consider the estimates of c. Let $\tau = -t,\ u = 1/w$. Then $u(\tau)$ satisfies

$$\dot{u} = a(-\tau)u^2 + 1. \tag{3.38}$$

Denote the Poincaré map of (3.38) by $T^*(z)$. Then $T^{-1}(\infty) = 1/T^*(0)$. Similar to the arguments as above, we have

$$\sigma_1^{-1} \tan 2\pi\sigma_1 < T^*(0) < \sigma_2^{-1} \tan 2\pi\sigma_2.$$

Hence

$$-c = T^*(0) \in [\sigma_1^{-1} \tan 2\pi\sigma_1\ \sigma_2^{-1} \tan 2\pi\sigma_2].$$

Suppose now that σ_1, σ_2 are in I_n^+ or in I_0. Then $0 < \tan 2\pi\sigma_1 \le \tan 2\pi\sigma_2$. Thus

$$0 < \sigma_1 \tan 2\pi\sigma_1 \le b \le \sigma_2 \tan 2\pi\sigma_2, \qquad 0 < \sigma_2 \cot 2\pi\sigma_2 \le -1/c \le \sigma_1 \cot 2\pi\sigma_1,$$

and

$$-b/c \in [\sigma_1\sigma_2 \tan 2\pi\sigma_1 \cot 2\pi\sigma_2,\ \sigma_1\sigma_2 \cot 2\pi\sigma_1 \tan 2\pi\sigma_2].$$

If σ_1, $\sigma_2 \in I_n^-$, then $\tan 2\pi\sigma_1 \le \tan 2\pi\sigma_2 < 0$. So we have

$$0 < -\sigma_2 \tan 2\pi\sigma_2 \le -b \le -\sigma_1 \tan 2\pi\sigma_1, \qquad 0 < -\sigma_1 \cot 2\pi\sigma_1 \le 1/c \le -\sigma_2 \cot 2\pi\sigma_2,$$

and

$$-b/c \in [\sigma_1\sigma_2 \cot 2\pi\sigma_1 \tan 2\pi\sigma_2,\ \sigma_1\sigma_2 \tan 2\pi\sigma_1 \cot 2\pi\sigma_2].$$

In both cases, we have

$$r_0 = (-b/c)^{-1/4} \in \left\langle (\sigma_1\sigma_2 \tan 2\pi\sigma_1 \cot 2\pi\sigma_2)^{-1/4},\ (\sigma_1\sigma_2 \cot 2\pi\sigma_1 \tan 2\pi\sigma_2)^{-1/4} \right\rangle.$$

The statement (3.35) follows from (3.34) directly. □

Remark 3.1. The lower bound in (3.35) can be improved as follows. By (3.22) and (3.33), we obtain from Lemma 3.4 that

$$\|r\|_4 \ge (2\pi)^{1/4}(2\pi/\theta)^{1/2} \ge (2\pi)^{1/4}\sigma_2^{-1/2}.$$

When $\sigma_1,\ \sigma_2 \in M_0$, (1.4) is in the first stability zone. In this case, a comparison result for the Hill equations holds within one period. Núñez proved in [16, Lemma 4.2] that

$$\sigma_2^{-1/2} \le r(t) \le \sigma_1^{-1/2} \qquad \text{for all } t. \tag{3.39}$$

This improves (3.34) in this case. It seems to us that the estimates (3.39) do not hold for higher order stability zones. Thus we will use the upper bound in (3.35) for general cases. Denote

$$N(\sigma_1,\sigma_2) = \max\left\{ \left(\frac{2\pi}{\sigma_1\sigma_2}\frac{\tan 2\pi\sigma_2}{\tan 2\pi\sigma_1}\right)^{1/2},\ \left(\frac{2\pi}{\sigma_1\sigma_2}\frac{\tan 2\pi\sigma_1}{\tan 2\pi\sigma_2}\right)^{1/2} \right\}. \tag{3.40}$$

So we have $\|r\|_4^2 \le N(\sigma_1,\sigma_2)$.

3.5. Estimating twist coefficients. The following theorem gives a sufficient condition for the zero solution $x = 0$ of (1.3) to be of twist type.

THEOREM 3.1. *Assume $a(t) \in C(\mathbb{R}/2\pi\mathbb{Z})$ satisfies (3.29) for some $\sigma_1,\ \sigma_2$ in an interval from Ω_0. Then (1.4) is 4-elementary and there exists a constant $\mu = \mu(\sigma_1,\sigma_2) > 0$ such that $x = 0$ (as a periodic solution of (1.3)) is of twist type provided that $b(t)$ and $c(t)$ satisfy*

$$\max_{t\in\mathbb{R}} c(t) < -\mu\|b\|_4^2. \tag{3.41}$$

Proof. Let $\sigma_1,\ \sigma_2$ be in an interval from Ω_0. Thus one of the conditions (3.30)–(3.32) is satisfied for some $n \in \mathbb{N}$. So the estimates in Lemma 3.6 hold in this case. By Lemma 3.2, (1.4) is 4-elementary.

We will prove that β^* given by (3.23) is positive under (3.41). Note that $(r,\ \varphi)$ in (3.23) is a solution of (3.19)+(3.20) and $r(t) > 0$ is 2π-periodic.

Let $C_- := \min_t(-c(t)) > 0$. Then

$$-\frac{3}{8}\int_{t_0}^{t_0+2\pi} c(t)r^4(t)\mathrm{d}t \ge \frac{3}{8}C_-\int_{t_0}^{t_0+2\pi} r^4(t)\mathrm{d}t = \frac{3}{8}C_-\|r\|_4^4,$$

where the last equality is due to the 2π-periodicity of $r(t)$.

For the terms in (3.23) containing $b(\cdot)$, we use (3.25) to obtain

$$\begin{aligned}
&\iint_{[t_0,t_0+2\pi]^2} b(t)b(s)r^3(t)r^3(s)\chi_1(|\varphi(t)-\varphi(s)|)\mathrm{d}s\mathrm{d}t\\
&+\frac{3}{16}\cot\frac{\theta}{2}\left|\int_{t_0}^{t_0+2\pi} b(t)r^3(t)e^{-i\varphi(t)}\mathrm{d}t\right|^2 + \frac{1}{16}\cot\frac{3\theta}{2}\left|\int_{t_0}^{t_0+2\pi} b(t)r^3(t)e^{3i\varphi(t)}\mathrm{d}t\right|^2\\
&\ge -K_1(\theta)\left(\int_{t_0}^{t_0+2\pi}|b(t)|r^3(t)dt\right)^2 \ge -K_1(\theta)\|b\|_4^2\|r\|_4^6,
\end{aligned}$$

where the Hölder inequality is used.

Combining these estimates with Lemma 3.6, we have

$$\beta^* \ge \left(\frac{3}{8}C_- - K_1(\theta)\|r\|_4^2\|b\|_4^2\right)\sigma\|r\|_4^4 \ge \left(\frac{3}{8}C_- - K_1(\theta)N(\sigma_1,\sigma_2)\|b\|_4^2\right)\sigma\|r\|_4^4,$$

where $N(\sigma_1, \sigma_2)$ is defined by (3.40). This implies that $\beta^* > 0$ if

$$C_- > \frac{8}{3} K_1(\theta) N(\sigma_1, \sigma_2) \|b\|_4^2.$$

By Lemma 3.6, we get from (3.28) that the constant μ in (3.41) can take

$$\mu = \mu_1(\sigma_1, \sigma_2) := \frac{8}{3} K_1(2\pi\sigma_2) N(\sigma_1, \sigma_2). \qquad \square \tag{3.42}$$

Remark 3.2. If we use (3.24) to estimate β^*, a similar argument shows that μ in (3.42) can be replaced by

$$\mu = \frac{8}{3} K_2(2\pi\sigma_2) N(\sigma_1, \sigma_2).$$

Consequently, using the function $K(\cdot)$ defined by (3.27), we know that the constant μ in (3.41) can take

$$\mu = \mu_2(\sigma_1, \sigma_2) := \frac{8}{3} K(2\pi\sigma_2) N(\sigma_1, \sigma_2). \tag{3.43}$$

In the above proof, the most important factor is just the upper bound of $\|r\|_4$ for the positive 2π-periodic solution $r(t)$ of (3.19). In fact, if some upper bound for $\|r\|_\ell$ for certain $\ell \geq 4$ can be found, one can then obtain a twist condition similar to (3.41). As for our Theorem 3.1, Lemma 3.6 actually gives an L^∞ estimate for $r(t)$, although it may not be optimal. As mentioned in Lemma 3.1, this can be improved especially when (1.4) is in the first stability zone. This will done in the next subsection.

3.6. An improvement for the first stability zone. Assume that $a(t) \in C(\mathbb{R}/2\pi\mathbb{Z})$ satisfies (3.29) for some $\sigma_1,\ \sigma_2 \in M_0 = (0, 1/4)$. In this case $\theta \in (0, \pi/2)$ and $a(t)$ is in the first stability zone.

For a function $f(t)$, let

$$f_+(t) = \max\{f(t), 0\}, \qquad f_-(t) = \max\{-f(t), 0\}$$

be the positive and the negative parts of $f(t)$. Note that $f = f_+ - f_-$.

Let $r(t)$ be the unique positive 2π-periodic solution of (3.19). Denote

$$r_0 = \min\{r(t) : t \in [0, 2\pi]\}, \qquad r_\infty = \max\{r(t) : t \in [0, 2\pi]\}.$$

We estimate the twist coefficient as follows. The term containing $c(t)$ is

$$\begin{aligned} -\frac{3}{8}\int_{t_0}^{t_0+2\pi} c(t) r^4(t) \mathrm{d}t &= \frac{3}{8}\int_0^{2\pi} c_-(t) r^4(t) \mathrm{d}t - \frac{3}{8}\int_0^{2\pi} c_+(t) r^4(t) \mathrm{d}t \\ &\geq \frac{3}{8} r_0^4 \|c_-\|_1 - \frac{3}{8} r_\infty^4 \|c_+\|_1. \end{aligned} \tag{3.44}$$

Now we use formula (3.24). Note that when $0 < \theta < \pi/2$, the kernel $\chi_2(x) > 0$ for all $x \in [0, \theta]$. Let

$$\chi_{20}(\theta) := \min_{x\in[0,\theta]} \chi_2(x) = \chi_2(0) = \frac{3\cos(\theta/2) + 2\cos(3\theta/2)}{8\sin(3\theta/2)},$$

$$\chi_{2\infty}(\theta) := \max_{x\in[0,\theta]} \chi_2(x) = \chi_2(\theta/2) = \frac{3\cos\theta + 2}{8\sin(3\theta/2)}.$$

Thus the term containing $b(\cdot)$ is

$$\begin{aligned}
&\iint_{[t_0,t_0+2\pi]^2} b(t)b(s)r^3(t)r^3(s)\chi_2(|\varphi(t)-\varphi(s)|)\mathrm{d}t\mathrm{d}s\\
&= \iint_{[t_0,t_0+2\pi]^2} (b_+(t)b_+(s)+b_-(t)b_-(s))r^3(t)r^3(s)\chi_2(|\varphi(t)-\varphi(s)|)\mathrm{d}t\mathrm{d}s\\
&\quad - \iint_{[t_0,t_0+2\pi]^2} (b_+(t)b_-(s)+b_-(t)b_+(s))r^3(t)r^3(s)\chi_2(|\varphi(t)-\varphi(s)|)\mathrm{d}t\mathrm{d}s\\
&\geq \chi_{20}(\theta)r_0^6 \iint_{[t_0,t_0+2\pi]^2} (b_+(t)b_+(s)+b_-(t)b_-(s))\mathrm{d}t\mathrm{d}s\\
&\quad -\chi_{2\infty}(\theta)r_\infty^6 \iint_{[t_0,t_0+2\pi]^2} (b_+(t)b_-(s)+b_-(t)b_+(s))\mathrm{d}t\mathrm{d}s\\
&= \chi_{20}(\theta)r_0^6(\|b_+\|_1^2+\|b_-\|_1^2) - 2\chi_{2\infty}(\theta)r_\infty^6\|b_+\|_1\|b_-\|_1. \qquad (3.45)
\end{aligned}$$

A very rough result from (3.44) and (3.45) is

$$\beta^* \geq \sigma\left[\frac{3}{8}r_0^4\|c_-\|_1 - \frac{3}{8}r_\infty^4\|c_+\|_1 - 2\chi_{2\infty}(\theta)r_\infty^6\|b_+\|_1\|b_-\|_1\right], \quad \theta\in(0,\pi/2). \qquad (3.46)$$

When $0<\theta\leq\pi/3$, which is just the case studied by Núñez [16], we can also use (3.23) to estimate β^* as follows. Note that

$$\chi_{10}(\theta) := \min_{x\in[0,\theta]}\chi_1(x) = \chi_1(0) = 0,$$

$$\chi_{1\infty}(\theta) := \max_{x\in[0,\theta]}\chi_1(x) = \begin{cases} (3\sin\theta - 2\sin^3\theta)/8, & 0<\theta\leq\pi/4,\\ \sqrt{2}/8, & \pi/4\leq\theta\leq\pi/3.\end{cases}$$

Thus

$$\begin{aligned}
&\iint_{[t_0,t_0+2\pi]^2} b(t)b(s)r^3(t)r^3(s)\chi_1(|\varphi(t)-\varphi(s)|)\mathrm{d}t\mathrm{d}s\\
&= \iint_{[t_0,t_0+2\pi]^2} (b_+(t)b_+(s)+b_-(t)b_-(s))r^3(t)r^3(s)\chi_1(|\varphi(t)-\varphi(s)|)\mathrm{d}t\mathrm{d}s\\
&\quad - \iint_{[t_0,t_0+2\pi]^2} (b_+(t)b_-(s)+b_-(t)b_+(s))r^3(t)r^3(s)\chi_1(|\varphi(t)-\varphi(s)|)\mathrm{d}t\mathrm{d}s\\
&\geq -\chi_{1\infty}(\theta)r_\infty^6 \iint_{[t_0,t_0+2\pi]^2} (b_+(t)b_-(s)+b_-(t)b_+(s))\mathrm{d}t\mathrm{d}s\\
&= -2\chi_{1\infty}(\theta)r_\infty^6\|b_+\|_1\|b_-\|_1. \qquad (3.47)
\end{aligned}$$

Since $\cot(\theta/2)\geq 0$ and $\cot(3\theta/2)\geq 0$ for $\theta\in(0,\pi/3]$, the other two terms in (3.23) containing $b(t)$ are nonnegative. Thus we get from (3.44) and (3.47) that

$$\beta^* \geq \sigma\left[\frac{3}{8}r_0^4\|c_-\|_1 - \frac{3}{8}r_\infty^4\|c_+\|_1 - 2\chi_{1\infty}(\theta)r_\infty^6\|b_+\|_1\|b_-\|_1\right], \quad \theta\in(0,\pi/3]. \qquad (3.48)$$

Note that $\chi_{1\infty}(\theta) < \chi_{2\infty}(\theta)$ for all $\theta\in(0,\pi/3)$. Thus (3.48) improves (3.46) when $0<\theta\leq\pi/3$. We simply use the following estimates:

$$\chi_{1\infty}(\theta)\leq\sqrt{2}/8, \qquad \theta\in(0,\pi/3],$$

and

$$\chi_{2\infty}(\theta) \le 7/16, \qquad \theta \in (\pi/3, \pi/2).$$

Recalling the estimates (3.39) for r_0 and r_∞, we conclude from (3.46) and (3.48) the following result.

THEOREM 3.2. *Suppose, in Theorem* 3.1, *that* $\sigma_1,\ \sigma_2 \in M_0 = (0,\ 1/4)$. *Then for any* $b,\ c \in C(\mathbb{R}/2\pi\mathbb{Z})$ *(which may change sign) satisfying*

$$\sigma_1^3\|c_-\|_1 - \sigma_1\sigma_2^2\|c_+\|_1 > \frac{7}{3}\,\sigma_2^2\|b_+\|_1\|b_-\|_1, \tag{3.49}$$

then the zero solution $x = 0$ *of* (1.3) *is of twist type. When* $\sigma_2 \le 1/6$, *which implies that* $\theta \in (0, \pi/3]$, (3.49) *can be improved as*

$$\sigma_1^3\|c_-\|_1 - \sigma_1\sigma_2^2\|c_+\|_1 > \frac{2\sqrt{2}}{3}\sigma_2^2\|b_+\|_1\|b_-\|_1. \tag{3.50}$$

Note that (3.50) improves the main of result of [16]. Moreover, Theorem 3.2 shows that the assumption that $0 < \theta \le \pi/3$ in [16] can be relaxed as $0 < \theta < \pi/2$, which is natural from the 4-elementary condition. See the remark following [16, Theorem 2.2]. As a result, his application to (1.1), which is based on the antimaximum principle [2], can be improved accordingly.

The proof above shows that, for any $0 < \sigma_1 \le \sigma_2 < 1/4$, there always exists some constant $\nu = \nu(\sigma_1, \sigma_2) > 0$ such that

$$\sigma_2^3\|c_-\|_1 - \sigma_1^2\sigma_2\|c_+\|_1 > \nu(\sigma_1, \sigma_2)\|b_+\|_1\|b_-\|_1 \tag{3.51}$$

ensures the twist character of $x = 0$ of (1.3). An explicit formula for the constant $\nu(\sigma_1, \sigma_2)$ can be obtained by carefully examining the functions $\chi_{1\infty}(\theta)$ and $\chi_{2\infty}(\theta)$ in (3.46) and (3.48). A twist condition similar to (3.51) can be worked out when the negative part $c_-(t)$ of $c(t)$ is dominated by the positive part $c_+(t)$.

As a final remark, we note that

$$\|b_+\|_1\|b_-\|_1 \le (\|b\|_1)^2 \le (2\pi)^{3/2}\|b\|_4^2.$$

Thus conditions (3.49)–(3.51) improve (3.41) because we can deal with the case where $b(t)$ and $c(t)$ may change sign.

4. Applications to the forced pendulum. In this section we apply the results in section 3 to study the twist character of the least amplitude periodic solution $x_\omega(t)$ of (1.1), where $\omega > 0$ and $p(t) \in C(\mathbb{R}/2\pi\mathbb{Z})$ satisfy (2.6). We use the notation from section 2. By Theorem 2.1, $\|x_\omega\|_\infty \le X^*(\alpha, \gamma) \le 3\gamma/2$. We always assume that

$$X^*(\alpha, \gamma) \le 3\gamma/2 < \pi/2. \tag{4.1}$$

Denote

$$\eta = \cos^{1/2}(3\gamma/2) \in (0, 1]. \tag{4.2}$$

Recall the formulas (1.7) of $a_\omega(t),\ b_\omega(t),\ c_\omega(t)$. Then

$$(\omega\eta)^2 \le a_\omega(t) = \omega^2 \cos x_w(t) \le \omega^2.$$

So we can take $\sigma_1 = \omega\eta$ and $\sigma_2 = \omega$. Since $b_\omega(t) = -(\omega^2/2)\sin x_w(t)$, we have $\|b_\omega\|_4^2 \le (2\pi)^{1/2}(\omega^4/4)(1-\eta^4)$. Using $c_\omega(t) = -(\omega^2/6)\cos x_\omega(t)$, one can take $C_- = \omega^2\eta^2/6$.

Let $I_n = (a_n, b_n)$ be an interval from Ω_0, i.e., I_n is one of the following intervals for some $n \in \mathbb{N}$:

$$I_n^1 = \Big(n-1,\, n-\frac{3}{4}\Big), \quad I_n^2 = \Big(n-\frac{3}{4},\, n-\frac{2}{3}\Big), \quad I_n^3 = \Big(n-\frac{2}{3},\, n-\frac{1}{2}\Big),$$

$$I_n^4 = \Big(n-\frac{1}{2},\, n-\frac{1}{3}\Big), \quad I_n^5 = \Big(n-\frac{1}{3},\, n-\frac{1}{4}\Big), \quad I_n^6 = \Big(n-\frac{1}{4},\, n\Big).$$

In the following, we restrict our discussion to $\omega \in I_n$. If

$$\eta > Q_1(\omega) := a_n/\omega, \qquad \omega \in I_n = (a_n, b_n), \tag{4.3}$$

then $\sigma_1 = \omega\eta > a_n$ and $\sigma_1,\ \sigma_2 \in I_n$. So Theorem 3.1 is applicable to this case.

By Theorem 3.1 and (3.43), $x_\omega(t)$ is of twist type when η satisfies

$$\frac{\omega^2\eta^2}{6} > \frac{8}{3}K(2\pi\omega)N(\omega\eta,\omega)(2\pi)^{1/2}\frac{\omega^4}{4}(1-\eta^4). \tag{4.4}$$

Let

$$S(\omega,\eta) = \max\left\{\left(\frac{\tan(2\pi\omega\eta)}{\tan(2\pi\omega)}\right)^{1/2}, \left(\frac{\tan(2\pi\omega)}{\tan(2\pi\omega\eta)}\right)^{1/2}\right\}.$$

Then

$$N(\omega\eta,\omega) = \frac{(2\pi)^{1/2}}{\omega\eta^{1/2}}S(\omega,\eta).$$

So (4.4) can be rewritten as

$$\eta^{5/2} > 8\pi\omega K(2\pi\omega)S(\omega,\eta)(1-\eta^4). \tag{4.5}$$

Note that both (4.3) and (4.5) are satisfied for $\eta = 1$. Thus conditions (4.3) and (4.5) can be rewritten as a single one like

$$\eta > Q_2(\omega), \qquad \omega \in I_n. \tag{4.6}$$

Here the function $Q_2(\omega)$ can be found numerically and estimated using the facts that $\eta^{5/2} > \eta^4$ for all $\eta \in (0,1)$ and $S(\omega,\eta) \to 1$ when $\eta \to 1$.

Recall (4.1) and (4.2). Let us introduce a function

$$P_2(\omega) = \min\left\{P_1(\omega),\, (4\omega|\sin\omega\pi|/3)\arccos Q_2^2(\omega)\right\}, \qquad \omega \in I_n. \tag{4.7}$$

If $\omega \in I_n$ and $p(t) \in C(\mathbb{R}/2\pi\mathbb{Z})$ satisfies

$$\|p\|_1 < P_2(\omega), \qquad \omega \in I_n,$$

then all conditions (2.6), (4.1), (4.3), and (4.4) are satisfied and $x_\omega(t)$ is thus of twist type. It is not difficult to check that $P_2(\omega)$ has the order $O(\omega^{1/2})$ when ω is bounded away from resonance of orders ≤ 4 and tends to ∞.

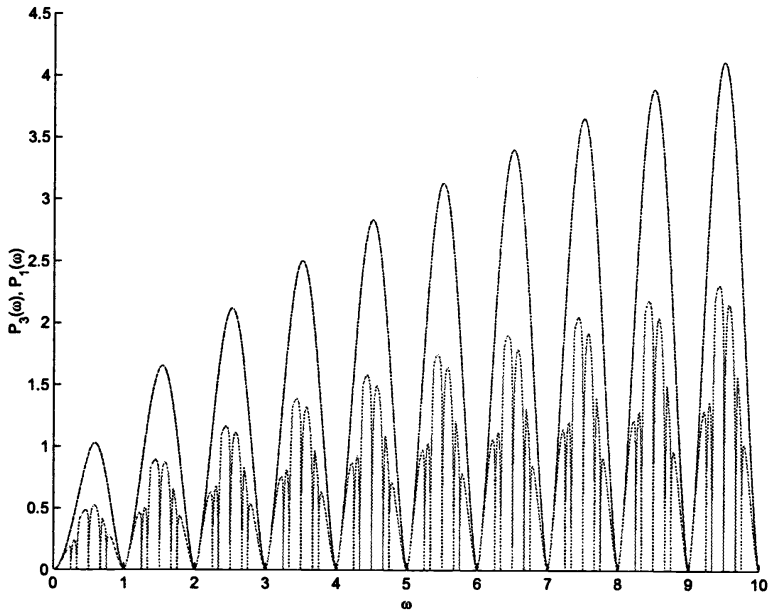


FIG. 2. *The graphs of* $P_1(\omega)$ *and* $P_3(\omega)$.

THEOREM 4.1. *There exists a nonnegative function* $P(\omega)$ *defined for all* $\omega > 0$ *such that if* $p(t) \in C(\mathbb{R}/2\pi\mathbb{Z})$ *satisfies*

$$\|p\|_1 < P(\omega),$$

then the least amplitude 2π*-periodic solution* $x_\omega(t)$ *of* (1.1) *is of twist type. Moreover,* $P(\omega) > 0$ *for all* $\omega \in \Omega_0$ *and* $P(\omega)$ *is of order* $O(\omega^{1/2})$ *when* ω *is bounded away from the resonance of orders* ≤ 4 *and tends to* ∞.

Remark 4.1. One can take the function $P(\omega)$ in Theorem 4.1 as $P_2(\omega)$ given by (4.7). If (4.2) is replaced by a more precise estimate

$$\eta = \cos^{1/2} X^*(\alpha, \gamma),$$

where $X^*(\alpha, \gamma)$ is given by (2.1), we find that the upper bounds $P(\omega)$ can be improved as $\|p\|_1 < P_3(\omega)$, where

$$\begin{aligned} P_3(\omega) &= \frac{4\omega|\sin\omega\pi|}{3(3\alpha)^{1/2}} \cos\left[3\arccos\left(\frac{1}{2}(3\alpha)^{1/2}\arccos Q_2^2(\omega)\right) - \pi\right] \\ &= \frac{4\sqrt{2}}{3} \frac{\omega|\sin\omega\pi|^{3/2}}{\left(\int_0^{\omega\pi}|\cos s|\mathrm{d}s\right)^{1/2}} \cos\left[3\arccos\left(\left(\frac{\int_0^{\omega\pi}|\cos s|\mathrm{d}s}{8|\sin\omega\pi|}\right)^{1/2}\arccos Q_2^2(\omega)\right) - \pi\right]. \end{aligned} \tag{4.8}$$

A comparison between $P_1(\omega)$ and $P_3(\omega)$, which are given by (2.6) and (4.8), respectively, is plotted in Figure 2.

Acknowledgment. The authors are grateful to the referees for their suggestions and their careful reading of the first version of the present work, especially the proof of Lemma 3.6.

REFERENCES

[1] D. Bonheure and C. De Coster, *Forced Singular Oscillators and the Method of Lower and Upper Solutions*, Preprint, 2001.
[2] C. De Coster and P. Habets, *Upper and lower solutions in the theory of ODE boundary value problems: Classical and recent results*, in Nonlinear Analysis and Boundary Value Problems for Ordinary Differential Equations, F. Zanolin, ed., CISM Courses and Lectures 371, Springer-Verlag, New York, 1996, pp. 1–78.
[3] M. A. del Pino, R. F. Manásevich, and A. Montero, *T-periodic solutions for some second order differential equations with singularities*, Proc. Roy. Soc. Edinburgh Sect. A, 120 (1992), pp. 231–243.
[4] S. Gan and M. Zhang, *Resonance pockets of Hill's equations with two-step potentials*, SIAM J. Math. Anal., 32 (2000), pp. 651–664.
[5] P. Hartman, *Ordinary Differential Equations*, 2nd ed., Birkhäuser, Boston, Basel, Stuttgart, 1982.
[6] S. P. Hastings and J. B. McLeod, *Chaotic motion of a pendulum with oscillatory forcing*, Amer. Math. Monthly, 100 (1993), pp. 563–572.
[7] E. Hille, *Ordinary Differential Equations in the Complex Domain*, John Wiley & Sons, New York, 1976.
[8] J. Lei and M. Zhang, *Twist property of periodic motion of an atom near a charged wire*, Lett. Math. Phys., 60 (2002), pp. 9–17.
[9] B. Liu, *The stability of the equilibrium of a conservative system*, J. Math. Anal. Appl., 202 (1996), pp. 133–149.
[10] J. N. Mather, *Existence of quasi-periodic orbits for twist homeomorphisms of the annulus*, Topology, 21 (1982), pp. 457–467.
[11] J. Mawhin, *Recent results on periodic solutions of the forced pendulum equation*, Rend. Istit. Mat. Univ. Trieste, 19 (1987), pp. 119–129.
[12] J. Mawhin, *The forced pendulum: A paradigm for nonlinear analysis and dynamical systems*, Expo. Math., 6 (1988), pp. 271–287.
[13] J. Mawhin, *Seventy-five years of global analysis around the forced pendulum equation*, in Equadiff 9 Proceedings, R. P. Agarwal, F. Neuman, and J. Vosmanský, eds., Electronic Publishing House, Stony Brook, NY, 1998, pp. 115–145.
[14] J. Moser, *On invariant curves of area preserving mappings of an annulus*, Nachr. Akad. Wiss. Göttingen Math.-Phys. Kl. II, (1962), pp. 1–20.
[15] J. Moser, *Recent developments in the theory of Hamiltonian systems*, SIAM Rev., 28 (1986), pp. 459–485.
[16] D. Núñez, *The method of lower and upper solutions and the stability of periodic oscillations*, Nonlinear Anal., 51 (2002), pp. 1207–1222.
[17] D. Núñez and R. Ortega, *Parabolic fixed points and stability criteria for nonlinear Hill's equations*, Z. Angew. Math. Phys., 51 (2000), pp. 890–911.
[18] D. Núñez and P. J. Torres, *Periodic solutions of twist type of an earth satellite equation*, Discrete Cont. Dynam. Systems, 7 (2001), pp. 303–306.
[19] R. Ortega, *The twist coefficient of periodic solutions of a time-dependent Newton's equation*, J. Dynam. Differential Equations, 4 (1992), pp. 651–665.
[20] R. Ortega, *The stability of the equilibrium of a nonlinear Hill's equation*, SIAM J. Math. Anal., 25 (1994), pp. 1393–1401.
[21] R. Ortega, *Periodic solution of a Newtonian equation: Stability by the third approximation*, J. Differential Equations, 128 (1996), pp. 491–518.
[22] E. Pinney, *The nonlinear differential equation $y''(x) + p(x)y + cy^{-3} = 0$*, Proc. Amer. Math. Soc., 1 (1950), p. 681.
[23] C. L. Siegel and J. Moser, *Lecture on Celestial Mechanics*, Springer-Verlag, Berlin, 1971.
[24] V. M. Staržinskiĭ, *A survey of works on the conditions of stability of the trivial solution of a system of linear differential equations with periodic coefficients*, Amer. Math. Soc. Transl. Ser. 2, Vol. 1, AMS, Providence, RI, 1955, pp. 189–237.
[25] S. Wiggins, *On the detection and dynamical consequences of orbits homoclinic to hyperbolic periodic orbits and normally hyperbolic invariant tori in a class of ordinary differential equations*, SIAM J. Appl. Math., 48 (1988), pp. 262–285.

[26] P. Yan and M. Zhang, *Higher order nonresonance for differential equations with singularities*, Math. Methods Appl. Sci., 26 (2003), pp. 1067–1074.
[27] J. You, *Invariant tori and Lagrange stability of pendulum-type equations*, J. Differential Equations, 85 (1990), pp. 54–65.
[28] M. Zhang, *The rotation number approach to eigenvalues of the one-dimensional p-Laplacian with periodic potentials*, J. London Math. Soc. (2), 64 (2001), pp. 125–143.
[29] M. Zhang, *The best bound on the rotations in the stability of periodic solutions of a Newtonian equation*, J. London Math. Soc. (2), 67 (2003), pp. 137–148.

SIAM J. MATH. ANAL.
Vol. 35, No. 4, pp. 868–883

DISPERSION AND STRICHARTZ INEQUALITIES FOR SCHRÖDINGER EQUATIONS WITH SINGULAR COEFFICIENTS*

VALERIA BANICA†

Abstract. In this paper we prove the global dispersion and the Strichartz inequalities for a class of one-dimensional Schrödinger equations with step-function coefficients having a finite number of discontinuities. The local and global dispersion and Strichartz inequalities are discussed for certain Schrödinger equations with low regularity coefficients oscillating at infinity.

Key words. Schrödinger equation, nonsmooth coefficients, dispersion and Strichartz inequalities, Bloch waves

AMS subject classifications. 35J10, 35R05, 35B45, 35C

DOI. 10.1137/S0036141002415025

1. Introduction. Strichartz estimates [7], [11] are an important tool for the understanding of nonlinear evolution equations. In the study of the dispersive properties of the Schrödinger equation with variable coefficients, the absence of the property of finite speed of propagation raises more difficulties than in the case of the wave equation. A way to "replace" this property is to impose a nontrapping condition on the trajectories. There are many results of wellposedness and smoothing effect for Schrödinger operators with smooth coefficients which are asymptotically flat and satisfy a nontrapping condition [4], [5], [8]. Staffilani and Tataru [10] proved the Strichartz estimates under the same conditions, but for lower regularity coefficients, only of $\mathcal{C}^2$-class. However, in order to have wellposedness for nonlinear Schrödinger equations (NLS), the nontrapping condition can be dropped. In their recent paper [2], Burq, Gérard, and Tzvetkov have obtained Strichartz estimates with fractional loss of derivative for metrics on $\mathbb{R}^d$ with uniformity assumptions at infinity, without geometric conditions. These new dispersive estimates imply local and global existence results for the Cauchy problem.

In this paper we study the dispersion property and the Strichartz inequalities for the one-dimensional Schrödinger equation

$$\text{(S)} \qquad \begin{cases} (i\,\partial_t + \partial_x a(x)\partial_x)\,u(t,x) = 0 \text{ for } (t,x) \in (0,\infty)\times\mathbb{R}, \\ u(0,x) = u_0(x) \in \mathbb{L}^2(\mathbb{R}) \end{cases}$$

for certain rough coefficients $a(x)$ without any geometric nontrapping condition.

In section 2 we prove global dispersion in the case of positive lamina coefficients, i.e., step functions with a finite number of singularities. Let us note in this situation the existence of trapped trajectories.

THEOREM 1.1. *Consider a partition of the real axis*

$$-\infty = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = \infty$$

and a step function

$$a(x) = b_i^{-2} \text{ for } x \in (x_{i-1}, x_i),$$

*Received by the editors September 24, 2002; accepted for publication (in revised form) February 7, 2003; published electronically October 14, 2003.
http://www.siam.org/journals/sima/35-4/41502.html
†Université de Paris Sud, Mathématiques, Bât. 425, 91405 Orsay Cedex, France (Valeria.Banica@math.u-psud.fr).

where b_i are positive numbers.

The solution of the Schrödinger equation (S) *satisfies the dispersion inequality*

$$\|u(t,\cdot)\|_{\mathbb{L}^\infty(\mathbb{R})} \leq \frac{C_n}{\sqrt{t}}\|u_0\|_{\mathbb{L}^1(\mathbb{R})}$$

and the Strichartz inequalities

$$\|u\|_{\mathbb{L}^p(\mathbb{R},\mathbb{L}^q(\mathbb{R}))} \leq C_n\|u_0\|_{\mathbb{L}^2(\mathbb{R})}$$

for every pair (p,q) verifying

$$\frac{2}{p}+\frac{1}{q}=\frac{1}{2}\,.$$

The proof consists of writing the solution by using the resolvent of the operator $-\partial_x a(x)\partial_x$. The resolvent is calculated and expressed in terms of series of exponentials. In order to get global dispersion, we discuss these series within the framework of the theory of Wiener's almost periodic functions.

We can also prove a similar result for the operator

$$i\,\partial_t + \frac{1}{\rho(x)}\partial_x a(x)\partial_x\,,$$

where $\rho(x)$ is a step function of the same type as $a(x)$.

Moreover, if $v(t,x)$ is the solution of the associated wave system

$$\text{(O)}\qquad \begin{cases} \left(\partial_t^2 - \partial_x a(x)\partial_x\right) v(t,x) = 0 \text{ for } x\in\mathbb{R}, \\ v(0,x) = u_0(x) \in \mathbb{L}^2(\mathbb{R}), \\ \partial_t v(0,x) = 0, \end{cases}$$

the same method gives us the following estimate:

$$\sup_{x\in\mathbb{R}}\int_{-\infty}^{\infty}|v(t,x)|dt \leq C_n\|u_0\|_{\mathbb{L}^1(\mathbb{R})}.$$

Dispersion is not satisfied if the step function coefficients are periodic. In section 3, by using the Krönig–Penney model, we show that the local dispersion fails in the case of 2-valued periodic step function coefficients.

THEOREM 1.2. *Let $x_0 \in (0,1)$ and let b_0, b_1 be positive numbers satisfying $b_0x_0 = b_1(1-x_0)$. Consider the 1-periodic function*

$$a(x) = \begin{cases} b_0^{-2} \text{ for } x\in[0,x_0), \\ b_1^{-2} \text{ for } x\in[x_0,1). \end{cases}$$

The local dispersion estimate fails for the Schrödinger equation (S).

The proof is based on the representation of the solution by its Floquet decomposition.

The fact that the coefficient a is not very oscillating at infinity seems to be essential for having dispersion. Applying the method used by Avellaneda, Bardos, and Rauch in [1], we can construct counterexamples for global dispersion and Strichartz's inequalities in the case of certain continuous coefficients oscillating at infinity.

Also, as Castro and Zuazua have recently shown in [3], even if the coefficients are flat at infinity, but rough ($C^{0,\alpha}$) and locally very oscillating, the local Strichartz inequalities fail.

All these results suggest the conjecture that the one-dimensional Schrödinger equations with strictly positive BV (bounded variation) coefficients satisfy the dispersion property.

2. Laminar media.

2.1. Representation of the resolvent of $-\partial_x a(x)\partial_x$. The operator $-\partial_x a(x)\partial_x$, defined from

$$\{h \in \mathbb{H}^1(\mathbb{R}), a\,\partial_x\, h \in \mathbb{H}^1(\mathbb{R})\}$$

to $\mathbb{L}^2(\mathbb{R})$, is self-adjoint. For $\omega \geq 0$ let R_ω be its resolvent

$$R_\omega g = (-\partial_x a(x)\partial_x + \omega^2 I)^{-1} g.$$

In order to obtain the expression of the resolvent on the intervals where a is constant, the second-order equations

$$\frac{1}{b_i^2}(R_\omega g)'' = \omega^2 R_\omega g - g$$

must be solved. Then, for $x \in (x_{i-1}, x_i)$, we have

$$R_\omega g(x) = c_{2i-1}e^{\omega b_i x} + c_{2i}e^{-\omega b_i x} + \int_{-\infty}^{\infty} \frac{g(y)}{2\omega} b_i e^{-\omega b_i |x-y|} dy.$$

Since $R_\omega g$ belongs to $\mathbb{L}^2(\mathbb{R})$ the coefficients c_2 and c_{2n-1} are zero. The conditions of continuity of $R_\omega g$ and of $a\,\partial_x R_\omega g$ at the points x_i give a system of $2n-2$ equations on the c_i's. The matrix D_n of this system is

$$\begin{pmatrix}
e^{\omega b_1 x_1} & -e^{\omega b_2 x_1} & -e^{-\omega b_2 x_1} & 0 & 0 & 0 & 0 & 0 \\
b_2 e^{\omega b_1 x_1} & -b_1 e^{\omega b_2 x_1} & b_1 e^{-\omega b_2 x_1} & 0 & 0 & 0 & 0 & 0 \\
0 & e^{\omega b_2 x_2} & e^{-\omega b_2 x_2} & -e^{\omega b_3 x_2} & -e^{-\omega b_3 x_2} & 0 & 0 & 0 \\
0 & b_3 e^{\omega b_2 x_2} & -b_3 e^{-\omega b_2 x_2} & -b_2 e^{\omega b_3 x_2} & b_2 e^{-\omega b_3 x_2} & 0 & 0 & 0 \\
: & : & : & : & : & : & : & : \\
0 & 0 & 0 & 0 & 0 & e^{\omega b_{n-1} x_{n-1}} & e^{-\omega b_{n-1} x_{n-1}} & -e^{-\omega b_n x_{n-1}} \\
0 & 0 & 0 & 0 & 0 & b_n e^{\omega b_{n-1} x_{n-1}} & -b_n e^{-\omega b_{n-1} x_{n-1}} & b_{n-1} e^{-\omega b_n x_{n-1}}
\end{pmatrix}.$$

The right-hand side of the system is

$$T_n = \begin{pmatrix} t_1 \\ : \\ t_{n-1} \end{pmatrix},$$

with

$$t_i = \begin{pmatrix} \int_{-\infty}^{\infty} \frac{g(y)}{2\omega}(-b_i e^{-\omega b_i |xi-y|} + b_{i+1} e^{-\omega b_{i+1}|x_i - y|})dy \\ \int_{-\infty}^{\infty} \frac{g(y)}{2\omega} b_{i+1} b_i (-e^{-\omega b_i |x_i - y|} + e^{-\omega b_{i+1}|x_i - y|})\mathrm{sign}(x_i - y)dy \end{pmatrix}.$$

Therefore the resolvent on each interval (x_i, x_{i+1}) is a finite sum of terms

$$R_\omega g(x) = \sum_{finite} C e^{\omega\beta(x)} \int_{I(x_i)} \frac{g(y)}{2\omega} \frac{e^{\pm\omega b_i y}}{\det D_n(\omega)} dy + \int_{-\infty}^{\infty} \frac{g(y)}{2\omega} b_i e^{-\omega b_i |x-y|} dy, \tag{1}$$

where $\beta(x)$ are real functions depending on $\{x, x_i, b_i\}$, C is a constant depending of $\{b_i\}$ and bounded by $(\max b_i^{-2})^n$, and $I(x_i)$ is either $(-\infty, x_i)$ or (x_i, ∞). Let $\widetilde{D}_n$ be the same matrix as D_n, with the last two terms of the last column replaced by

$$\begin{pmatrix} -e^{\omega b_n x_{n-1}} \\ -b_{n-1} e^{\omega b_n x_{n-1}} \end{pmatrix}.$$

The development of the determinants of D_n and $\widetilde{D}_n$ with respect to the last column gives the following induction relations:

$$\begin{cases} \det D_n = e^{-\omega b_n x_{n-1}}\Big[(b_{n-1}-b_n)e^{-\omega b_{n-1}x_{n-1}}\det\widetilde{D_{n-1}} - \\ \qquad\qquad -(b_{n-1}+b_n)e^{\omega b_{n-1}x_{n-1}}\det D_{n-1}\Big], \\ \det\widetilde{D_n} = e^{\omega b_n x_{n-1}}\Big[(b_{n-1}-b_n)e^{\omega b_{n-1}x_{n-1}}\det D_{n-1} - \\ \qquad\qquad -(b_{n-1}+b_n)e^{-\omega b_{n-1}x_{n-1}}\det\widetilde{D_{n-1}}\Big]. \end{cases}$$

Let us define for $n \geq m \geq 2$

$$Q_m(\omega) = e^{-2\omega b_m x_m}\frac{\det\widetilde{D_m}}{\det D_m}.$$

By denoting

$$d_{m-1} = \frac{b_{m-1}-b_m}{b_{m-1}+b_m},$$

we have for $n \geq 3$

$$\det D_n(\omega) = (b_1+b_2)e^{-\omega(b_2-b_1)x_1}\prod_{i=2\ldots n-1}(b_i+b_{i+1})e^{\omega(b_i-b_{i+1})x_i}(1-d_iQ_i(\omega)), \tag{2}$$

and for $n = 2$

$$\det D_2(\omega) = (b_1+b_2)e^{-\omega(b_2-b_1)x_1}. \tag{3}$$

Also, we obtain an induction formula on the Q_m's:

$$Q_m(\omega) = e^{-2\omega b_m(x_m-x_{m-1})}\frac{-d_{m-1}+Q_{m-1}(\omega)}{1-d_{m-1}Q_{m-1}(\omega)}. \tag{4}$$

Note that a Möbius transform on the unit disc occurs in this expression.

Let $\epsilon_n > 0$ be such that for every complex ω with

$$\Re\omega > -\epsilon_n,$$

the estimate

$$|Q_2(\omega)| = |d_1e^{-2\omega b_2(x_2-x_1)}| < 1$$

holds and gives by induction

$$|Q_m(\omega)| < 1.$$

Hence $(\det D_n(\omega))^{-1}$ is uniformly bounded and well defined in this region, which contains the imaginary axis. Therefore $\omega R_\omega u_0(x)$ can be analytically continued, and we can use the following spectral theory lemma.

LEMMA 2.1. *The solution of the Schrödinger equation* (S) *verifies*

$$u(t,x) = \int_{-\infty}^{\infty} e^{it\tau^2}\tau R_{i\tau}u_0(x)\frac{d\tau}{\pi}. \tag{5}$$

2.2. The algebra of Wiener's almost-periodic functions. Let us recall the structure of the Banach algebra of Wiener's almost-periodic functions:

$$B = \left\{ h : \mathbb{R} \mapsto \mathbb{C}\,,\ h(t) = \sum_{\lambda \in \mathbb{R}} c(\lambda) e^{i\lambda t} \text{ with } \|h\|_B = \sum_{\lambda \in \mathbb{R}} |c(\lambda)| < \infty \right\}.$$

We define for $h \in B$

$$\|h\|_\infty = \sup_{t \in \mathbb{R}} |h(t)|$$

and

$$\rho(h) = \inf\{r > 0 \ \mid\ \exists\, C_r > 0 \text{ for all } k \in \mathbb{N}\,,\ \|h^k\|_B \leq C_r r^k\}.$$

The following classical result, which is a consequence of Theorems 6§4 and 2§29 of [6], will be used.

THEOREM 2.2. *For all $h \in B$ we have*

$$\rho(h) = \|h\|_\infty.$$

COROLLARY 2.3. *Let $h \in B$ with $\|h\|_\infty < 1$ and let α be a complex number on the open unit disc. Then*

$$g = \frac{h - \alpha}{1 - \overline{\alpha} h}$$

also belongs to B and

$$\rho(g) < 1.$$

Proof. The function $\overline{\alpha} h$ belongs to B and

$$\|\overline{\alpha} h\|_\infty < |\alpha| < 1.$$

By using Theorem 2.2 we have

$$\|(\overline{\alpha} h)^k\|_B \leq C |\alpha|^k.$$

Since

$$\frac{h - \alpha}{1 - \overline{\alpha} h} = (h - \alpha) \sum_{k=0}^{\infty} (\overline{\alpha} h)^k,$$

it follows that g belongs to B. Moreover, by the maximum principle,

$$\|g\|_\infty < 1.$$

By again applying Theorem 2.2, the corollary is proved. ∎

2.3. The dispersion inequality. The $Q_m(i\tau)$'s are series of complex exponentials. In this subsection we will show that they belong to B with respect to the real variable τ. The estimates of their norm in this algebra will imply the dispersion for the Schrödinger equation (S).

Let us define

$$r_2 = |d_1|, \;\; r_m = \frac{|d_{m-1}| + r_{m-1}}{1 - |d_{m-1}| r_{m-1}}.$$

Obviously $Q_2 \in B$ and

$$\|Q_2\|_\infty = r_2.$$

Therefore Theorem 2.2 gives us

$$\rho(Q_2) = r_2 < 1.$$

By using Corollary 2.3 and the Möbius transform which occurs in formula (4), one can show by induction that $Q_m \in B$ and

$$\rho(Q_m) \leq r_m < 1.$$

Then formulae (2) and (3) lead us to the estimate

$$\|(\det D_n(i\tau))^{-1}\|_B < K_n, \tag{6}$$

where K_n is a constant depending on b_i.

In order to prove dispersion, it is sufficient, using (1) and (5), to estimate terms of the following type:

$$J_i(t,x) = \int_{-\infty}^{\infty} e^{it\tau^2} C e^{i\tau\beta(x)} \int_{I(x_i)} \frac{u_0(y)}{2i\tau} \frac{e^{\pm i\tau b_i y}}{\det D_n(i\tau)} dy \, \tau \frac{d\tau}{2\pi}.$$

By performing a change of variable in τ,

$$|J_i(t,x)| \leq C \int_{I(x_i)} \frac{|u_0(y)|}{4\pi\sqrt{t}} \left| \int_{-\infty}^{\infty} e^{is^2} \frac{e^{i\frac{s}{\sqrt{t}}(\beta(x) \pm b_i y)}}{\det D_n(i\frac{s}{\sqrt{t}})} ds \right| dy$$

$$\leq C \frac{\|u_o\|_{\mathbb{L}^1(\mathbb{R})}}{\sqrt{t}} \|(\det D_n(i\xi))^{-1}\|_B.$$

Then (6) implies that

$$\sup_x |J_i(t,x)| \leq K_n \frac{\|u_0\|_{\mathbb{L}^1(\mathbb{R})}}{\sqrt{t}},$$

so the dispersion inequality for the Schrödinger equation (S) is satisfied.

Remark 2.4. The finite sum in (1) contains $n2^n$ terms. Therefore, by estimating the solution as above, term by term, we cannot obtain the dispersion for equation (S) if $a(x)$ has an infinite number of steps. Therefore the method is too rough to prove dispersion for an arbitrary strictly positive BV coefficient $a(x)$.

Strichartz inequalities follow from the dispersion inequality by the classical duality argument TT^* [12], so the proof of Theorem 1.1 is complete.

Since we can express the solution of the wave equation (O) as

$$v(t,x)=\int_{-\infty}^{\infty}e^{it\tau}R_{i\tau}u_0(x)i\tau\frac{d\tau}{2\pi},$$

the property

$$\sup_{x\in\mathbb{R}}\int_{-\infty}^{\infty}|v(t,x)|dt\leq C\|u_0\|_{\mathbb{L}^1(\mathbb{R})}$$

follows similarly to the dispersion inequality for the solution of (S).

3. Periodic laminar media.

3.1. General theory of periodic-coefficient equations. Let θ be a number in $[0,2\pi]$ and consider the operator on $\mathbb{L}^2(\mathbb{S}^1)$

$$A_\theta=-(i\theta+\partial_x)a(x)(i\theta+\partial_x).$$

This operator is self-adjoint with a compact resolvent, hence the eigenvalues form a sequence of strictly positive numbers $\{\omega_{\theta,n}^2\}_{n\in\mathbb{N}}$. Moreover, the set of the corresponding eigenfunctions $p_n(\theta,x)$ is an orthonormal basis of $\mathbb{L}^2(\mathbb{S}^1)$.

Let us provide a way to construct the elements of this basis. Finding the eigenfunction $p_n(\theta,x)$ is equivalent to finding the function

$$\Psi_n(\theta,x)=e^{i\theta x}p_n(\theta,x)$$

that satisfies

$$-\partial_x a(x)\partial_x\Psi_n(\theta,x)=\omega_{\theta,n}^2\Psi_n(\theta,x). \tag{$\mathrm{H}_{\theta,n}$}$$

Note that this new function has the quasi-periodic property

$$\Psi_n(\theta,x+1)=e^{i\theta}\Psi_n(\theta,x).$$

Equation $(\mathrm{H}_{\theta,n})$ is of the type

$$-\partial_x a(x)\partial_x\Psi(x)=\lambda^2\Psi(x) \tag{H}$$

on

$$\{\Psi\in\mathbb{H}^1_{loc}(\mathbb{R}),\ a\,\partial_x\Psi\in\mathbb{H}^1_{loc}(\mathbb{R})\}.$$

This equation can be treated similarly to Hill's equation [9]. Let T be an operator acting on the solution space as follows:

$$T(\Psi)(x)=\Psi(x+1).$$

On the one hand, the eigenvalues of T verify

$$x^2-x\mathrm{Tr}(T)+\det T=0.$$

On the other hand, the generalized Wronskian

$$W = \Psi_1 a\partial_x \Psi_2 - \Psi_2 a\partial_x \Psi_1$$

associated with (Ψ_1, Ψ_2), a normalized basis of solutions of (H), i.e.,

$$\Psi_1(0) = (a\,\partial_x \Psi_2)(0) = 1, \qquad (a\,\partial_x \Psi_1)(0) = \Psi_2(0) = 0,$$

is constant. Therefore

$$\det T = W(1) = W(0) = 1,$$

and the eigenvalues are $e^{i\xi}$ and $e^{-i\xi}$ for some complex ξ. If $|\mathrm{Tr}(T)|$ is larger than 2, then ξ is purely imaginary and there exists a basis of solutions of exponential growth. In this case λ^2 belongs to an instability interval of the equation. Otherwise, if $|\mathrm{Tr}(T)|$ is less than or equal to 2, ξ is real and λ^2 belongs to a stability interval. Moreover, if $\xi \in \pi\mathbb{Z}$, periodic solutions exist. If $\xi \in \mathbb{R}\setminus\pi\mathbb{Z}$, the existence of a basis of quasi-periodic solutions is assured.

So, the eigenvalues of A_θ are exactly the values λ^2 for which the operator T associated with (H) admits $e^{i\theta}$ and $e^{-i\theta}$ as eigenvalues. If $\theta \in (0,\pi)\cup(\pi,2\pi)$, then these eigenvalues are simple. Therefore, in order to construct the $\mathbb{L}^2(\mathbb{S}^1)$ basis made of the eigenfunctions of A_θ, one has to find all λ for which the operator T associated with (H) verifies

$$\mathrm{Tr}T = 2\cos\theta.$$

For such a λ, we consider (Ψ_1, Ψ_2) a normalized basis of solutions of (H). If $\Psi_2(1) \neq 0$, then

$$\Psi(x) = \Psi_1(x) - \frac{\Psi_1(1) - e^{i\theta}}{\Psi_2(1)}\Psi_2(x) \tag{7}$$

is a solution of (H) and an eigenfunction of T for the eigenvalue $e^{i\theta}$. Finally,

$$p(x) = \Psi(x)e^{-i\theta x}$$

is an eigenfunction of the operator A_θ, associated with the eigenvalue λ^2.

3.2. Representation of solutions. In order to find the representation of the solution of (S), we decompose the initial data as follows:

$$u_0(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ix\xi}\,\widehat{u_0}(\xi)d\xi = \frac{1}{2\pi}\sum_{k\in\mathbb{Z}}\int_{2k\pi}^{2(k+1)\pi} e^{ix\xi}\,\widehat{u_0}(\xi)d\xi$$

$$= \frac{1}{2\pi}\sum_{k\in\mathbb{Z}}\int_0^{2\pi} e^{i(2k\pi+\theta)x}\,\widehat{u_0}(2k\pi+\theta)d\theta.$$

Thus u_0 can be written

$$u_0(x) = \frac{1}{2\pi}\int_0^{2\pi} v(\theta, x)\,d\theta,$$

with

$$v(\theta,x)=\sum_{k\in\mathbb{Z}}e^{i(2k\pi+\theta)x}\,\widehat{u_0}(2k\pi+\theta).\tag{8}$$

Moreover,

$$\|u_0\|^2_{\mathbb{L}^2(\mathbb{R})}=\frac{1}{2\pi}\|\widehat{u_0}\|^2_{\mathbb{L}^2(\mathbb{R})}=\frac{1}{2\pi}\sum_{k\in\mathbb{Z}}\int_{2k\pi}^{2(k+1)\pi}|\widehat{u_0}(x)|^2dx=\sum_{k\in\mathbb{Z}}\int_0^{2\pi}|\widehat{u_0}(2k\pi+\theta)|^2d\theta$$

$$=\int_0^{2\pi}\int_0^1|e^{-i\theta x}v(\theta,x)|^2dxd\theta=\int_0^{2\pi}\int_0^1|v(\theta,x)|^2dxd\theta.$$

Since v satisfies the quasi-periodicity property

$$v(\theta,x+1)=e^{i\theta}v(\theta,x),$$

then $v(\theta,x)e^{-i\theta x}$ is 1-periodic. Therefore we can decompose it with respect to the $\mathbb{L}^2(\mathbb{S}^1)$ basis of eigenfunctions of the operator A_θ introduced in section 3.1. If $\theta\in(0,\pi)\cup(\pi,2\pi)$, the eigenvalues of A_θ are simple and we can write

$$v(\theta,x)e^{-i\theta x}=\sum_{n\in\mathbb{N}}c_n(\theta)p_n(\theta,x);$$

that is,

$$v(\theta,x)=\sum_{n\in\mathbb{N}}c_n(\theta)\Psi_n(\theta,x).\tag{9}$$

Finally,

$$u(t,x)=\frac{1}{2\pi}\int_0^{2\pi}\sum_{n\in\mathbb{N}}e^{it\omega^2_{\theta,n}}c_n(\theta)\Psi_n(\theta,x)d\theta\tag{10}$$

is the solution of the Schrödinger equation (S). Moreover, using the above link between the $\mathbb{L}^2$ norms of the initial datum u_0 and of v,

$$\|u_0\|^2_{\mathbb{L}^2(\mathbb{R})}=\sum_{n\in\mathbb{N}}\|c_n\|^2_{\mathbb{L}^2(0,2\pi)}.$$

Let us now express the solution u in terms of the initial datum u_0. By using the definitions (8) and (9),

$$c_n(\theta)=\langle v(\theta,\cdot),\Psi_n(\theta,\cdot)\rangle=\sum_{k\in\mathbb{Z}}\widehat{u_0}(2k\pi+\theta)\langle e^{i(2k\pi+\theta)\,\cdot},\Psi_n(\theta,\cdot)\rangle.$$

Since $e^{-i\theta x}\Psi_n(\theta,x)$ is 1-periodic, its Fourier decomposition contains only even exponentials:

$$e^{-i\theta x}\Psi_n(\theta,x)=\sum_{k\in\mathbb{Z}}d_{n,k}(\theta)e^{i2\pi kx}.$$

Therefore

$$c_n(\theta) = \sum_{k\in\mathbb{Z}} \widehat{u_0}(2k\pi+\theta)\overline{d}_{n,k}(\theta) = \int_{-\infty}^{\infty} u_0(y)e^{-iy\theta}\sum_{k\in\mathbb{Z}} e^{-i2k\pi y}\overline{d}_{n,k}(\theta)dy$$

$$= \int_{-\infty}^{\infty} u_0(y)\overline{\Psi}_n(\theta,y)dy.$$

In conclusion, for any initial datum u_0, the solution of the Schrödinger equation (S) is

$$u(t,x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} u_0(y)\int_0^{2\pi}\sum_{n\in\mathbb{N}} e^{it\omega_{\theta,n}^2}\Psi_n(\theta,x)\overline{\Psi}_n(\theta,y)d\theta dy.$$

3.3. Explicit solutions for the Krönig–Penney model. Let

$$a(x) = \begin{cases} b_0^{-2} \text{ for } x\in[0,x_0), \\ b_1^{-2} \text{ for } x\in[x_0,1) \end{cases}$$

as defined in the statement of Theorem 1.2. Fix $\theta\in(0,\pi)\cup(\pi,2\pi)$. Following the approach presented in section 3.1, in this subsection we will explicitly find the functions $\Psi_n(\theta,x)$.

The basis of normalized solutions associated with the (H) is

$$\begin{cases} \Psi_1(x) = \begin{cases} \frac{1}{2}e^{i\lambda b_0 x} + \frac{1}{2}e^{-i\lambda b_0 x} \text{ for } x\in(0,x_0), \\ a_j^1 e^{i\lambda b_1 x} + b_j^1 e^{-i\lambda b_1 x} \text{ for } x\in(x_0,1), \end{cases} \\ \Psi_2(x) = \begin{cases} -\frac{ib_0}{2\lambda}e^{i\lambda b_0 x} + \frac{ib_0}{2\lambda}e^{-i\lambda b_0 x} \text{ for } x\in(0,x_0), \\ a_j^2 e^{i\lambda b_1 x} + b_j^2 e^{-i\lambda b_1 x} \text{ for } x\in(x_0,1) \end{cases} \end{cases}$$

with

$$\begin{cases} a_j^1 = \frac{1}{4b_0}[(b_0+b_1)e^{i\lambda x_0(b_0-b_1)} + (b_0-b_1)e^{-i\lambda x_0(b_0+b_1)}], \\ b_j^1 = \frac{1}{4b_1}[(b_0+b_1)e^{-i\lambda x_0(b_0-b_1)} + (b_0-b_1)e^{i\lambda x_0(b_0+b_1)}], \\ a_j^2 = \frac{i}{4\lambda}[-(b_0+b_1)e^{i\lambda x_0(b_0-b_1)} + (b_0-b_1)e^{-i\lambda x_0(b_0+b_1)}], \\ b_j^2 = \frac{i}{4\lambda}[(b_0+b_1)e^{-i\lambda x_0(b_0-b_1)} - (b_0-b_1)e^{i\lambda x_0(b_0+b_1)}]. \end{cases}$$

The trace of the shift operator T is

$$\mathrm{Tr}T = \Psi_1(1) + \frac{1}{b_1^2}\,\partial_x\Psi_2(1).$$

One can calculate

$$\mathrm{Tr}(T) = (r+1)\cos[\lambda(x_0b_0+(1-x_0)b_1)] - (r-1)\cos[\lambda(x_0b_0-(1-x_0)b_1)],$$

where

$$r = \frac{b_0^2+b_1^2}{2b_0b_1}.$$

By setting the conditions

$$\mathrm{Tr}(T) = 2\cos\theta, \;\; x_0 b_0 = (1-x_0)b_1,$$

it follows that

$$2\cos\theta = (r+1)\cos(\lambda 2x_0 b_0) - (r-1).$$

Hence we have

$$\lambda \in \left\{ \frac{2\pi j + f(\theta)}{2x_0 b_0}\,,\, j \in \mathbb{Z} \right\},$$

where $f(\theta)$ is the analytic function

$$f(\theta) = \arccos \frac{r - 1 + 2\cos\theta}{r+1}.$$

As the solutions Ψ_1 and Ψ_2 are the same for λ and for $-\lambda$, we have to check if there exist different integers j and k such that

$$2\pi j + f(\theta) = \pm(2\pi k + f(\theta)).$$

If this is true, it follows that

$$j + k = \frac{f(\theta)}{\pi}.$$

Since $r > 1$ gives $f(\theta) < \pi$ and $\theta \neq 0$ gives $f(\theta) \neq 0$, then j and k must satisfy

$$0 < |j + k| < 1.$$

In conclusion, the values

$$\left| \frac{2\pi j + f(\theta)}{2x_0 b_0} \right|$$

are different, so we can consider the eigenvalues of the operator A_θ indexed by $j \in \mathbb{Z}$ as follows:

$$\omega_{\theta,j} = \frac{2\pi j + f(\theta)}{2x_0 b_0}. \tag{11}$$

Note that since θ has been fixed in $(0,\pi) \cup (\pi, 2\pi)$,

$$\omega_{\theta,j} \neq 0 \text{ for all } j \in \mathbb{Z}.$$

By using (7), we obtain a quasi-periodic solution for equation $(\mathrm{H}_{\theta,j})$:

$$\widetilde{\Psi}_j(\theta, x) = \left(\frac{1}{2} + h_j(\theta)\right) e^{i\omega_{\theta,j} b_0 x} + \left(\frac{1}{2} - h_j(\theta)\right) e^{-i\omega_{\theta,j} b_0 x} \text{ for } x \in (0, x_0) \tag{12}$$

with

$$h_j(\theta) = i\frac{(b_0+b_1)\cos(2\omega_{\theta,j} b_0 x_0) + (b_0 - b_1) - e^{i\theta}}{(b_0+b_1)\sin(2\omega_{\theta,j} b_0 x_0)}.$$

The definition (11) of $\omega_{\theta,j}$ gives

$$h_j(\theta) = h(\theta) = i\frac{(b_0+b_1)\cos f(\theta) + (b_0-b_1) - e^{i\theta}}{(b_0+b_1)\sin f(\theta)}.$$

Then we can calculate for $x \in (0, x_0)$

$$\widetilde{\Psi}_j(\theta, x) = \cos(\omega_{\theta,j}\, b_0 x) + 2h(\theta)\sin(\omega_{\theta,j}\, b_0 x),$$

and for $x \in (x_0, 1)$

$$\widetilde{\Psi}_j(\theta, x) = \left(a_j^1 - a_j^2\, h(\theta)\frac{2\omega_{\theta,j}}{ib_0}\right) e^{i\omega_{\theta,j} b_1 x} + \left(b_j^1 - b_j^2\, h(\theta)\frac{2\omega_{\theta,j}}{ib_0}\right) e^{-i\omega_{\theta,j} b_1 x}$$

$$= \frac{b_0+b_1}{4b_0}(1+2h(\theta))e^{i\omega_{\theta,j}(x_0(b_0-b_1)+b_1 x)} + \frac{b_0-b_1}{4b_0}(1-2h(\theta))e^{-i\omega_{\theta,j}(x_0(b_0+b_1)-b_1 x)}$$

$$+\frac{b_0+b_1}{4b_0}(1-2h(\theta))e^{-i\omega_{\theta,j}(x_0(b_0-b_1)+b_1 x)} + \frac{b_0-b_1}{4b_0}(1+2h(\theta))e^{i\omega_{\theta,j}(x_0(b_0+b_1)-b_1 x)}.$$

It follows that

$$\int_0^1 |\widetilde{\Psi}_j(\theta, x)|^2 dx = \alpha_j(\theta) = \beta(\theta) + \frac{\gamma(\theta)}{2\pi j + f(\theta)},$$

with $\beta(\theta)$ strictly positive. Let $\Psi_j(\theta, x)$ be the $\mathbb{L}^2$ normalization of $\widetilde{\Psi}_j(\theta, x)$:

$$\Psi_j(\theta, x) = \frac{\widetilde{\Psi}_j(\theta, x)}{\sqrt{\alpha_j(\theta)}}.$$

We are now in the context described in section 3.2.

3.4. The failure of local dispersion. Let $\mathcal{X}$ be a 2π-periodic function whose restriction to $(0, 2\pi)$ is $\mathcal{C}_0^\infty$. One can write

$$\mathcal{X}(\xi) = \sum_{k\in\mathbb{Z}} s_k e^{ik\xi}.$$

Let v_0 be the Fourier localization outside $2\pi\mathbb{Z}$ points of the initial data u_0

$$\widehat{v_0}(\xi) = \widehat{u_0}(\xi)\mathcal{X}(\xi).$$

By applying Plancherel's theorem one has

$$v_0(x) = \int_{-\infty}^{\infty} e^{ix\xi}\widehat{u_0}(\xi)\mathcal{X}(\xi)\frac{d\xi}{2\pi} = \sum_{k\in\mathbb{Z}} u_0(x+k)s_k.$$

Since $\mathcal{X}|_{(0,2\pi)}$ is in $\mathcal{C}_0^\infty$,

$$\sum_{k\in\mathbb{Z}} |s_k| = S < \infty,$$

so the localization preserves the regularity $\mathbb{L}^1(\mathbb{R})\cap\mathbb{L}^2(\mathbb{R})$ with

$$\begin{cases}\|v_0\|_{\mathbb{L}^1(\mathbb{R})}\leq C\|u_0\|_{\mathbb{L}^1(\mathbb{R})},\\ \|v_0\|_{\mathbb{L}^2(\mathbb{R})}\leq C\|u_0\|_{\mathbb{L}^2(\mathbb{R})}.\end{cases}$$

For such an initial datum v_0, the coefficients $c_j(\theta)$ defined in section 3.2 are

$$c_j(\theta)=\sum_{k\in\mathbb{Z}}\widehat{u_0}(2k\pi+\theta)\mathcal{X}(2k\pi+\theta)\overline{d}_{j,k}(\theta)$$

$$=\mathcal{X}(\theta)\int_{-\infty}^{\infty}u_0(y)e^{-iy\theta}\sum_{k\in\mathbb{Z}}e^{-i2k\pi y}\overline{d}_{j,k}(\theta)dy=\mathcal{X}(\theta)\int_{-\infty}^{\infty}u_0(y)\overline{\Psi}_{\theta,j}(y)dy.$$

Then, by the representation formula (10), the solution $v(t,x)$ of the equation (S) with initial datum v_0 can be written as

$$v(t,x)=\int_{-\infty}^{\infty}u_0(y)K_t(x,y)dy,$$

where

$$K_t(x,y)=\frac{1}{2\pi}\int_0^{2\pi}\sum_{j\in\mathbb{Z}}e^{it\omega_{\theta,j}^2}\Psi_{\theta,j}(x)\overline{\Psi}_{\theta,j}(y)\mathcal{X}(\theta)d\theta.$$

Since

$$\|v_0\|_{\mathbb{L}^1(\mathbb{R})}\leq C\|u_0\|_{\mathbb{L}^1(\mathbb{R})},$$

in order to have the dispersion inequality

$$\|v(t,\cdot)\|_{\mathbb{L}^\infty(\mathbb{R})}\leq\frac{C}{\sqrt{t}}\|v_0\|_{\mathbb{L}^1(\mathbb{R})},$$

the dispersion kernel must satisfy

$$\|K_t\|_{\mathbb{L}^\infty(x,y)}\leq\frac{C}{\sqrt{t}}.$$

We will show that there exist times t, arbitrarily small, for which K_t is not an $\mathbb{L}^\infty(x,y)$ function.

Let us change t in $\frac{t}{4b_o^2x_0^2}$ and x in $\frac{x}{2x_0}$. By using definition (11) of $\omega_{\theta,j}$ and formula (12) for $\widetilde{\Psi}_j(\theta,x)$, we have that $K_t(x,y)$ is, for $x<x_0$, equal to

$$\frac{1}{4\pi}\sum_{j\in\mathbb{Z}}\int_0^{2\pi}e^{it(2\pi j+f(\theta))^2}\left(e^{ix(2\pi j+f(\theta))}(1+2h(\theta))+e^{-ix(2\pi j+f(\theta))}(1-2h(\theta))\right)$$

$$\times\left(e^{-iy(2\pi j+f(\theta))}(1+2\overline{h}(\theta))+e^{iy(2\pi j+f(\theta))}(1-2\overline{h}(\theta))\right)\frac{\mathcal{X}(\theta)}{\alpha_j(\theta)}d\theta.$$

It follows that the kernel is the sum of four terms of the following type:

$$J_t(x,y)=\frac{1}{4\pi}\sum_{j\in\mathbb{Z}}\int_0^{2\pi}e^{it(2\pi j+f(\theta))^2}e^{i(x-y)(2\pi j+f(\theta))}(1+2h(\theta))(1+2\overline{h}(\theta))\frac{\mathcal{X}(\theta)}{\alpha_j(\theta)}d\theta.$$

In view of the forthcoming applications of the stationary phase formula, we can consider that $J_t(x,y)$ is, modulo an $\mathbb{L}^\infty$ function, the same sum as above, with α_0 replaced by α_1. Since $|f(\theta)| < \pi$, one can choose a function $\alpha_\xi(\theta)$ which is strictly positive, bounded, and $\mathcal{C}^\infty$ with respect to the variable ξ such that

$$\alpha_\xi(\theta) = \beta(\theta) + \frac{\gamma(\theta)}{\xi + f(\theta)} \text{ for } |\xi| > \pi.$$

This allows us to apply the Poisson formula, so $J_t(x,y)$ can be written as

$$\frac{1}{2}\sum_{l\in\mathbb{Z}} e^{i\xi l}\int_{-\infty}^{\infty}\int_0^{2\pi} e^{it(\xi+f(\theta))^2}e^{i(x-y)(\xi+f(\theta))}(1+2h(\theta))(1+2\overline{h}(\theta))\frac{\mathcal{X}(\theta)}{\alpha_\xi(\theta)}d\theta d\xi.$$

By changing $\xi + f(\theta)$ into ζ,

$$J_t(x,y) = \frac{1}{2}\sum_{l\in\mathbb{Z}}\int_0^{2\pi} e^{-if(\theta)l}\, I_l(t,x-y,\theta)\, d\theta,$$

where

$$I_l(t,x-y,\theta) = \mathcal{X}(\theta)(1+2h(\theta))(1+2\overline{h}(\theta))\int_{-\infty}^{\infty} e^{it\zeta^2}e^{i(x-y+l)\zeta}\frac{d\zeta}{\alpha_{\zeta-f(\theta)}(\theta)}$$

verifies

$$|\partial\theta^k I_l(t,x-y,\theta)| \le C \text{ for all } k\in\mathbb{N}.$$

The only critical point of $f|_{(0,2\pi)}$ is π, which is nondegenerate, so we can apply the stationary phase formula for large l. In view of the definition of $\alpha_\zeta(\pi)$, $J_t(x,y)$ is, modulo an $\mathbb{L}^\infty$ function,

$$J_t(x,y) = \sum_{l\in\mathbb{Z}^*}\left(\frac{e^{-if(\pi)l}}{\sqrt{|l|}}\, I_l(t,x-y)\,\frac{1}{2}\mathcal{X}(\pi)(1+2h(\pi))(1+2\overline{h}(\pi)) + O(|l|^{-\frac{3}{2}})\right)$$

with

$$I_l(t,x-y) = \int_{-\infty}^{\infty} e^{it\zeta^2}e^{i(x-y+l)\zeta}\frac{d\zeta}{\beta(\pi)+\frac{\gamma(\pi)}{\zeta}}.$$

We have used the known result that the sum of exponentials

$$F(\alpha) = \sum_{l\in\mathbb{Z}^*}\frac{e^{-i\alpha l}}{\sqrt{|l|}} \tag{13}$$

blows up as

$$\frac{1}{\sqrt{|\alpha|}}$$

if α tends to zero, and otherwise the sum is finite. Here $f(\pi)\in(0,\pi)$.

By changing ζ in $\frac{x-y+l}{\sqrt{t}}$ and by considering that (x,y) lies in a compact set, we have

$$I_l(t,x-y) = \frac{x-y+l}{\sqrt{t}}\int_{-\infty}^{\infty} e^{i(x-y+l)^2(\zeta^2+\frac{\zeta}{\sqrt{t}})}\,\frac{d\zeta}{\beta(\pi)+\frac{\gamma(\pi)\sqrt{t}}{(x-y+l)\zeta}}.$$

The stationary phase formula applied again for $\zeta = -\frac{1}{2\sqrt{t}}$ gives

$$I_l(t,x-y) = \frac{1}{\sqrt{t}} e^{-i\frac{(x-y+l)^2}{4t}} \frac{1}{\beta(\pi) - \frac{2\gamma(\pi)t}{x-y+l}} + \frac{O((x-y+l)^{-2})}{\sqrt{t}}.$$

Thus, modulo an $\mathbb{L}^\infty$ function, we obtain that

$$J_t(x,y) = \frac{C}{\sqrt{t}} \sum_{l\in\mathbb{Z}^*} \frac{e^{-if(\pi)l}}{\sqrt{|l|}}\, e^{-i\frac{(x-y+l)^2}{4t}},$$

with $C \neq 0$. Let t verify

$$\frac{1}{4t} \in 2\pi\mathbb{Z}.$$

Note that t can be chosen arbitrary small. Also,

$$J_t(x,y) = \frac{Ce^{-i\frac{(x-y)^2}{4t}}}{\sqrt{t}} \sum_{l\in\mathbb{Z}^*} \frac{e^{-i(\frac{x-y}{2t}+f(\pi))l}}{\sqrt{|l|}}.$$

It follows then that $K_t(x,y)$ is, modulo an $\mathbb{L}^\infty$ function,

$$\frac{e^{-i\frac{(x-y)^2}{4t}}}{\sqrt{t}} \left(C_1\, F\left(\frac{x-y}{2t} + f(\pi)\right) + C_2\, F\left(-\frac{x-y}{2t} + f(\pi)\right)\right)$$

$$+\frac{e^{-i\frac{(x+y)^2}{4t}}}{\sqrt{t}} \left(C_3\, F\left(\frac{x+y}{2t} + f(\pi)\right) + C_4\, F\left(-\frac{x+y}{2t} + f(\pi)\right)\right).$$

Since $f(\pi) \neq 0$, in view of the behavior of F presented above (see (13)), the kernel $K_t(x,y)$ is not in $\mathbb{L}^\infty(x,y)$. Therefore the local dispersion for the Schrödinger equation (S) fails and Theorem 1.2 is proved.

Acknowledgment. I thank my advisor Patrick Gérard for having guided this work.

REFERENCES

[1] M. Avellaneda, C. Bardos, and J. Rauch, *Contrôlabilité exacte, homogénéisation et localisation d'ondes dans un milieu non-homogène*, Asymptot. Anal., 5 (1992), pp. 481–494.

[2] N. Burq, P. Gérard, and N. Tzvetkov, *Strichartz inequalities and the nonlinear Schrödinger equation on compact manifolds*, Amer. J. Math., to appear.

[3] C. Castro and E. Zuazua, *Concentration and lack of observability of waves in highly heterogeneous media*, Arch. Ration. Mech. Anal., 164 (2002), pp. 39–72.

[4] W. Craig, T. Kappeler, and W. Strauss, *Microlocal dispersive smoothing for the Schrödinger equation*, Comm. Pure Appl. Math., 48 (1995), pp. 769–860.

[5] S. Doi, *Remarks on the Cauchy problem for Schrödinger-type equations*, Comm. Partial Differential Equations, 21 (1996), pp. 163–178.

[6] I.M. Gelfand, D.A. Raikov, and G.E. Chilov, *Les anneaux normés commutatifs*, Monographies internationales de mathématiques modernes, Gauthier-Villars, Paris, 1964.

[7] J. Ginibre and G. Velo, *Generalized Strichartz inequalities for the wave equation*, J. Funct. Anal., 133 (1995), pp. 50–68.

[8] L. Kapitanski and Y. Safarov, *Dispersive smoothing for Schrödinger equations*, Math. Res. Lett., 3 (1996), pp. 77–91.

[9] W. Magnus and S. Winkler, *Hill's Equation*, Interscience Tracts in Pure and Applied Mathematics 20, Interscience–John Wiley, New York, 1966.
[10] G. Staffilani and D. Tataru, *Strichartz estimates for a Schrödinger operator with nonsmooth coefficients*, Comm. Partial Differential Equations, 27 (2002), pp. 1337–1372.
[11] R.S. Strichartz, *Restrictions of Fourier transforms to quadratic surfaces and decay of solutions of wave equations*, Duke Math. J., 44 (1977), pp. 705–714.
[12] P. Tomas, *A restriction theorem for the Fourier transform*, Bull. Amer. Math. Soc., 81 (1975), pp. 177–178.

SIAM J. MATH. ANAL.
Vol. 35, No. 4, pp. 884–921

STABILITY OF SELF-SIMILAR SOLUTIONS OF THE DAFERMOS REGULARIZATION OF A SYSTEM OF CONSERVATION LAWS*

XIAO-BIAO LIN† AND STEPHEN SCHECTER†

Abstract. In contrast to a viscous regularization of a system of n conservation laws, a Dafermos regularization admits many self-similar solutions of the form $u = u(\frac{X}{T})$. In particular, it is known in many cases that Riemann solutions of a system of conservation laws have nearby self-similar smooth solutions of an associated Dafermos regularization. We refer to these smooth solutions as *Riemann–Dafermos solutions*. In the coordinates $x = \frac{X}{T}$, $t = \ln T$, Riemann–Dafermos solutions become stationary, and their time-asymptotic stability as solutions of the Dafermos regularization can be studied by linearization. We study the stability of Riemann–Dafermos solutions near Riemann solutions consisting of n Lax shock waves. We show, by studying the essential spectrum of the linearized system in a weighted function space, that stability is determined by eigenvalues only. We then use asymptotic methods to study the eigenvalues and eigenfunctions. We find there are fast eigenvalues of order $\frac{1}{\epsilon}$ and slow eigenvalues of order 1. The fast eigenvalues correspond to eigenvalues of the viscous profiles for the individual shock waves in the Riemann solution; these have been studied by other authors using Evans function methods. The slow eigenvalues are related to inviscid stability conditions that have been obtained by various authors for the underlying Riemann solution.

Key words. conservation law, Riemann problem, Dafermos regularization, stability, spectrum, singular perturbation

AMS subject classifications. 35L65, 35L67, 35C20

DOI. 10.1137/S0036141002405029

1. Introduction.

Consider a system of *viscous conservation laws* in one space dimension, i.e., a partial differential equation of the form

$$u_T + f(u)_X = (B(u)u_X)_X, \tag{1.1}$$

where $X \in \mathbb{R}$, $T \in [0, \infty)$, $u \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}^n$, and $B(u)$ is an $n \times n$ matrix for which all eigenvalues have positive real part. We are interested in the behavior, as $T \to \infty$, of solutions of (1.1) that satisfy the constant boundary conditions

$$u(-\infty, T) = u^\ell, \quad u(+\infty, T) = u^r, \quad 0 \le T < \infty, \tag{1.2}$$

and some initial condition $u(X, 0) = u^0(X)$. Our interest is not in the solution for any particular initial condition, but in the possible asymptotic behavior of solutions as $T \to \infty$.

It is believed that as $T \to \infty$, solutions of such initial-boundary-value problems typically approach Riemann solutions for the system of conservation laws

$$u_T + f(u)_X = 0 \tag{1.3}$$

obtained from (1.1) by dropping the viscous term. In numerical simulations, the convergence is seen when the solution is viewed in the rescaled spatial variable $x = \frac{X}{T}$; the rescaling counteracts the tendency of the solution to spread as time increases. The

*Received by the editors April 3, 2002; accepted for publication (in revised form) May 30, 2003; published electronically November 4, 2003. This work was supported in part by the National Science Foundation under grant DMS-9973105.

http://www.siam.org/journals/sima/35-4/40502.html

†Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (xblin@math.ncsu.edu, schecter@math.ncsu.edu).

shock waves in the observed Riemann solution satisfy the viscous profile criterion for the viscosity $B(u)$. Speaking very roughly, Riemann solutions are believed to play the same role for (1.1)–(1.2) that constant solutions (equilibria) play for ordinary differential equations (ODEs): they are the simplest asymptotic states. An important difference, however, is that Riemann solutions are not solutions of (1.1) but only of the related equation (1.3). We recall that a *shock wave* is a weak solution with a jump discontinuity of the system of conservation laws (1.3). The simplest such solutions are

$$u(X,T) = \begin{cases} u^- & \text{for } X < sT, \\ u^+ & \text{for } X > sT. \end{cases} \tag{1.4}$$

For (1.4) to be a weak solution of (1.3), the triple (u^-, s, u^+) must satisfy the *Rankine–Hugoniot condition*

$$f(u^+) - f(u^-) - s(u^+ - u^-) = 0. \tag{1.5}$$

A shock wave (1.4) satisfies the *viscous profile criterion* for the viscosity $B(u)$, provided (1.1) has a traveling wave solution $u(X - sT)$ that satisfies the boundary conditions

$$u(-\infty) = u^-, \quad u(+\infty) = u^+. \tag{1.6}$$

A traveling wave solution of (1.1) that satisfies these boundary conditions exists if and only if the *traveling wave ODE*

$$\dot{u} = B(u)^{-1}(f(u) - f(u^-) - s(u - u^-)) \tag{1.7}$$

has an equilibrium at u^+ (it automatically has one at u^-) and a connecting orbit from u^- to u^+. The condition that (1.7) have an equilibrium at u^+ is just the Rankine–Hugoniot condition (1.5).

A *Riemann problem* for the system of conservation laws (1.3) is an initial value problem of the form

$$u(X,0) = \begin{cases} u^\ell & \text{for } X < 0, \\ u^r & \text{for } X > 0. \end{cases} \tag{1.8}$$

Since (1.3), (1.8) is invariant under the transformations $(X,T) \to (aX, aT)$, to avoid one-parameter families of solutions, a solution $u(X,T)$ of (1.3), (1.8) should have the form $u(X,T) = \hat{u}(x)$, $x = \frac{X}{T}$. Then $\hat{u}(x)$ satisfies

$$(Df(u) - xI)u_x = 0, \quad -\infty < x < \infty; \qquad u(-\infty) = u^\ell, \quad u(\infty) = u^r. \tag{1.9}$$

Notice that even though a Riemann problem in the form (1.3), (1.8) is an initial value problem, in the form (1.9) it is a boundary value problem.

Normally one looks for a solution of (1.9) consisting of constant parts, continuously changing parts (*rarefaction waves*), and jump discontinuities (shock waves). Shock waves occur when

$$\lim_{x \to s-} \hat{u}(x) = u^- \neq u^+ = \lim_{x \to s+} \hat{u}(x).$$

We shall require that each such triple (u^-, s, u^+) satisfy the viscous profile criterion for a given $B(u)$.

It is known that even with the viscous profile criterion, Riemann problems can have multiple solutions. This is disconcerting if the Riemann problem is regarded as an initial value problem. There is no such difficulty, however, when Riemann problems are regarded as boundary value problems whose solutions represent asymptotic states of (1.1)–(1.2). Indeed, in this context, multiple solutions of a Riemann problem represent multiple asymptotic states of (1.1)–(1.2), which are approached for different initial conditions $u^0(X)$. For a model initial-boundary-value problem (1.1)–(1.2) whose associated Riemann problem has three solutions, Azevedo et al. [2] have done careful numerical work that indicates that this is in fact the case. Two of the Riemann solutions appear to be attractors, while the third appears to attract a codimension-one set of initial conditions.

The study of the stability of Riemann solutions as asymptotic states of (1.1)–(1.2) is not easy. If the Riemann solution is a single shock wave, then it corresponds to a traveling wave solution of (1.1), and one can use a moving coordinate system to convert the traveling wave solution to a steady state solution. One can then study stability by studying the spectrum of the linearization at this solution. There is always a zero eigenvalue, which corresponds to shifts of the traveling wave. An additional difficulty is that the continuous spectrum touches the imaginary axis. For a single conservation law, Sattinger [39] dealt with this difficulty by using an exponentially weighted norm, which shifts the continuous spectrum to the left. For systems, the gap lemma of Gardner and Zumbrun [14] (see also [19]) allows one to study eigenvalues of the linearization near the origin despite the continuous spectrum. A series of papers by Liu, Zumbrun, and Howard justifies the passage from linear to nonlinear stability [28], [29], [27], [50].

Alternatively, one can study stability of viscous shock waves by energy methods [34], [15]. A relation between the two approaches is that energy methods can be used to verify that the spectrum of the linearization is contained in the left half plane.

Riemann solutions other than a single shock wave do not correspond to traveling wave solutions of (1.1). Thus one cannot determine their stability by finding the spectrum of a linear operator. In some situations one can construct an approximate solution of (1.1)–(1.2) near the Riemann solution and show that solutions of (1.1)–(1.2) that start near the approximate solution approach it. See [26] for Riemann solutions consisting of weak Lax shock waves and [45] for Riemann solutions consisting of a single rarefaction.

Riemann solutions are functions of $\frac{X}{T}$ only, and it is in the variables (x, T) with $x = \frac{X}{T}$ that the convergence of solutions of (1.1)–(1.2) to Riemann solutions is observed. With this motivation, in (1.1) we make the change of variables

$$x = \frac{X}{T}, \quad t = \ln T. \tag{1.10}$$

(The substitution $t = \ln T$ is simply for convenience. Decay that is algebraic in T becomes exponential in t.) We obtain

$$u_t + (Df(u) - xI)u_x = e^{-t}(B(u)u_x)_x. \tag{1.11}$$

Thus in the (x, t) variables, which are natural for the study of the large-time behavior of solutions of (1.1), (1.1) becomes a system that is both spatially dependent and nonautonomous. In studying nonautonomous systems, it is natural to first freeze

the time variable and study the resulting autonomous system. In this case one sets $\epsilon = e^{-t}$; for large t, ϵ is small. One obtains

$$u_t + (Df(u) - xI)u_x = \epsilon(B(u)u_x)_x. \tag{1.12}$$

Returning to (X, T) variables, (1.12) becomes

$$u_T + f(u)_X = \epsilon T(B(u)u_X)_X. \tag{1.13}$$

Equation (1.13) is the *Dafermos regularization* of the system of conservation laws (1.3) associated with the viscosity $B(u)$ ([8]; see also [46], [47]). It is usually regarded as an artificial, nonphysical equation because of the factor T in the viscous term. As we have seen, however, if one is interested in the behavior of solutions of (1.1)–(1.2) for large T and uses the appropriate variables (1.10) for large T, the Dafermos regularization is actually a natural simplification of the physical equations. Like the Riemann problem, but unlike (1.1), (1.13) has many solutions of the form $u(X, T) = \hat{u}(x)$, $x = \frac{X}{T}$. (This is why it was originally introduced.) They satisfy a *Dafermos ODE*

$$(Df(u) - xI)u_x = \epsilon(B(u)u_x)_x. \tag{1.14}$$

Corresponding to the Riemann data (1.8) we have the boundary conditions

$$u(-\infty) = u^\ell, \quad u(+\infty) = u^r. \tag{1.15}$$

We shall refer to a solution $u_\epsilon(x)$ of (1.14)–(1.15) as a *Riemann–Dafermos solution* of (1.13) for the boundary data (u^ℓ, u^r). A Riemann–Dafermos solution of (1.13) is just a stationary solution of (1.12). The boundary value problem (1.14)–(1.15) is a viscous regularization of the Riemann boundary value problem (1.9).

Actually, Dafermos always used $B(u) \equiv I$. For this case, he conjectured that Riemann–Dafermos solutions of the boundary value problem (1.14)–(1.15) converge to a corresponding Riemann solution as $\epsilon \to 0$. This conjecture has been proved for u^r close to u^ℓ by Tzavaras [48]. His proof relies on showing that the Riemann–Dafermos solutions are of uniformly bounded variation and oscillation.

Recently, Szmolyan [44] studied the boundary value problem (1.14)–(1.15) with $B(u) \equiv I$ using geometric singular perturbation theory [18]. The idea is to think of a Riemann solution, with shock waves that satisfy the viscous profile criterion for $B(u) \equiv I$, as a singular solution ($\epsilon = 0$), and then show by geometric singular perturbation theory that, for small $\epsilon > 0$, there is a nearby Riemann–Dafermos solution.

A Riemann solution is *structurally stable* if the number and types of its waves do not change when the flux function or boundary data are varied slightly [40]. (This use of the term "structurally stable" is consistent with its use in dynamical systems theory, but differs from Majda's use of the term in [32].) For $B(u) \equiv I$, Szmolyan proved that, for small $\epsilon > 0$, structurally stable classical Riemann solutions, which consist of n rarefactions and Lax shock waves, have Riemann–Dafermos solutions of (1.14)–(1.15) nearby. There is no requirement that u^ℓ and u^r be close.

A valuable feature of the Dafermos regularization is that it works equally well for general $B(u)$. Schecter [41] makes this point explicit and shows that any structurally stable Riemann solution consisting entirely of shock waves that satisfy the viscous profile criterion for a given $B(u)$ has, for small $\epsilon > 0$, a Riemann–Dafermos solution of (1.14)–(1.15) nearby. Undercompressive shock waves, whose existence and location are very dependent on $B(u)$, are explicitly allowed.

It is likely that any structurally stable Riemann solution whose shock waves satisfy the viscous profile criterion for a given $B(u)$ has Riemann–Dafermos solutions of the corresponding Dafermos regularization nearby. Some nonstructurally stable Riemann solutions are treated in [30].

In this paper we shall study the Dafermos system (1.13) in the transformed form (1.12), with boundary conditions

$$u(-\infty,t)=u^\ell,\quad u(\infty,t)=u^r,\quad 0\le t<\infty. \tag{1.16}$$

Our goal is to begin the study of the asymptotic stability of Riemann–Dafermos solutions (i.e., steady state solutions) of (1.12), (1.16). We will consider (1.12), (1.16) on the time interval $t\ge 0$, which corresponds to considering (1.13) on $T\ge 1$.

The possible usefulness of this study for the study of the stability of Riemann solutions as asymptotic states of (1.1)–(1.2) is as follows. Let

$$u(x,t)=u_\epsilon(x) \text{ with } \epsilon=e^{-t},$$

where the $u_\epsilon(x)$ are Riemann–Dafermos solutions of (1.12) that converge, as $\epsilon\to 0$, to a Riemann solution $\hat{u}(x)$ of (1.3), (1.8). Then for large t, $u(x,t)$ is almost a solution of (1.11) and converges as $t\to\infty$ to $\hat{u}(x)$. With a good enough understanding of the stability of the $u_\epsilon(x)$ as solutions of (1.12), one can perhaps show that near $u(x,t)$ is a true solution of (1.11) with the same stability that $u_\epsilon(x)$ has as a solution of (1.12) for small ϵ.

Tzavaras [48] gives a different argument for the relevance of the Dafermos regularization to understanding Riemann solutions as asymptotic states of (1.1). We now preview the remainder of the paper. For simplicity, we shall take $B(u)\equiv I$. Then (1.12) becomes

$$u_t+(Df(u)-xI)u_x=\epsilon u_{xx}. \tag{1.17}$$

We consider a structurally stable Riemann solution of (1.3) that consists of exactly n Lax shock waves with speeds $\bar{s}^1<\bar{s}^2<\cdots<\bar{s}^n$. We assume that each Lax shock wave satisfies the viscous profile criterion for $B(u)=I$. Precise definitions are given in section 2. We do not assume that u^ℓ and u^r are close.

We write the Riemann solution as a piecewise constant function $u_0(x)$ that is undefined at $x=\bar{s}^i$, $i=1,\dots,n$, where $u_0(x)$ has jumps. From [44] or [41], near it there is, for small $\epsilon>0$, a Riemann–Dafermos solution $u_\epsilon(x)$ of (1.17). It has sharp transition layers near $x=\bar{s}^i$, $i=1,\dots,n$.

In section 3, we construct an asymptotic expansion of $u_\epsilon(x)$ in powers of ϵ. In the *regular layer*, which is $\mathbb{R}$ with $\bar{s}^i$, $i=1,\dots,n$, removed, $u_\epsilon(x)$ has an expansion of the form

$$u_\epsilon^R(x)=\sum_{j=0}^{\infty}\epsilon^j u_j^R(x),$$

in which $u_0^R(x)$ is just the piecewise constant Riemann solution $u_0(x)$.

We shall refer to a small neighborhood of $\bar{s}^i$ as the *ith singular layer* and denote it S^i, $i=1,\dots,n$. The Riemann–Dafermos solution $u_\epsilon(x)$ has sharp transition layers at

$$x^i(\epsilon)=\sum_{j=0}^{\infty}\epsilon^j x_j^i,\quad i=1,\dots,n,$$

with $x^i(0) = \bar{s}^i$. Near $x^i(\epsilon)$ we use the stretched variable $\xi = \frac{x - x^i(\epsilon)}{\epsilon}$. In terms of this variable, the solution has an expansion

$$u^i_\epsilon(\xi) = \sum_{j=0}^{\infty} \epsilon^j u^i_j(\xi) \quad \text{in the singular layer } S^i.$$

It turns out that $u^i_0(\xi)$ is a traveling wave of (1.1) with speed $\bar{s}^i$.

This description of $u_\epsilon(x)$ is consistent with its construction by geometric singular perturbation theory.

Let $C(\gamma, \mathbb{R}_x)$, $\gamma \geq 0$, be the Banach space of uniformly continuous functions $U(x)$ such that the weighted norm $|U|_\gamma := \sup_x |U(x)| e^{\gamma |x|} < \infty$. Let

$$C^2(\gamma, \mathbb{R}_x) := \{U : U, U', U'' \in C(\gamma, \mathbb{R}_x)\}.$$

On $C^2(\gamma, \mathbb{R}_x)$ we will use the equivalent norms $|U|_{2,\gamma,\epsilon} := |U|_\gamma + \epsilon |U'|_\gamma + \epsilon^2 |U''|_\gamma$, where $\epsilon > 0$ is the small parameter in (1.17). This family of norms was used by Fife [12]; the ϵ scales out when the stretched variable $\xi = \frac{x - x^i(\epsilon)}{\epsilon}$ is used instead of x. An advantage of this family of norms is that one can have a family of functions $U_\epsilon(x)$ for which $\sup_x |U'_\epsilon(x)| = O(\frac{1}{\epsilon})$ and $\sup_x |U''_\epsilon(x)| = O(\frac{1}{\epsilon^2})$ but $|U_\epsilon|_{2,\gamma,\epsilon}$ remains bounded as $\epsilon \to 0$.

Let X_γ denote the affine space of functions $u(x) = u_\epsilon(x) + U(x)$ with $U \in C^2(\gamma, \mathbb{R}_x)$. This function space includes the most important perturbations of $u_\epsilon(x)$. We shall study (1.17) together with the boundary conditions (1.16) in the space X_γ. In section 4 we show that for $\gamma \geq 0$, (1.17), (1.16) is well-posed in a neighborhood of $u_\epsilon(x)$ in X_γ. The size of the neighborhood is uniform in the norm $|\cdot|_{2,\gamma,\epsilon}$ as $\epsilon \to 0$. Thus, for small $\epsilon > 0$, perturbations with large derivatives are allowed.

An argument like that of Evans [10] shows that linearized stability of $u_\epsilon(x)$ in X_γ implies nonlinear stability in X_γ. Therefore we consider the linearized stability of $u_\epsilon(x)$ in X_γ.

In section 5 we show that for γ sufficiently large, using the exponentially weighted norm moves the essential spectrum of the linearization of (1.17) about $u_\epsilon(x)$ to the left of the imaginary axis, as in [39], [38]. Thus linearized stability of $u_\epsilon(x)$ in X_γ is determined by the eigenvalues.

In sections 6 and 7 we study eigenvalues for $\gamma > 0$ using asymptotic expansions in ϵ. We assume the eigenvalues have asymptotic expansions of the form

$$\lambda = \sum_{j=-1}^{\infty} \epsilon^j \lambda_j$$

and the corresponding eigenfunctions have similar expansions. Section 6 is devoted to eigenvalues with $\lambda_{-1} \neq 0$. The corresponding eigenfunctions are *local*; i.e., their expansions are nonzero only in singular layers. These eigenvalues reflect the fast convergence of the solution to traveling waves in the singular layers. Section 7 is devoted to eigenvalues with $\lambda_{-1} = 0$, which we discuss in more detail below. The fact that there are both $O(\frac{1}{\epsilon})$ and $O(1)$ eigenvalues is consistent with the description of solutions at the beginning of section 6.

The fast eigenvalues $\lambda = \frac{\lambda_{-1}}{\epsilon} + O(1)$, with $\lambda_{-1} \neq 0$, correspond to the nonzero eigenvalues λ_{-1} of the individual traveling waves that are found by Evans function methods [14], [3]. However, a nondegeneracy condition is needed to ensure that a zero of the Evans function can be continued to a fast eigenvalue $\lambda = \frac{\lambda_{-1}}{\epsilon} + O(1)$; see

section 6. Thus, roughly speaking, a necessary condition for stability of the Riemann–Dafermos solution is that the Evans function for each individual viscous shock wave in the Riemann solution have no zero with positive real part. Slow eigenvalues have the form $\lambda = \lambda_0 + O(\epsilon)$. It turns out that $\lambda_0 = 0$ is never an eigenvalue, while $\lambda_0 = -1$ is always among the $O(1)$ eigenvalues. Its multiplicity is n. The corresponding eigenfunctions are local. To lowest order they are just the derivatives of the individual traveling waves in the n singular layers and correspond to shifts of the traveling waves.

Other $O(1)$ eigenvalues are nonlocal: The corresponding eigenfunctions asymptotically satisfy a piecewise continuous system of ODEs in x, along with jump conditions at $x = \bar{s}^i$, $i = 1, \ldots, n$. To lowest order, these $O(1)$ eigenvalues and eigenfunctions can be interpreted as eigenvalues and eigenfunctions for a system of first-order hyperbolic equations. This system has been used by many authors to study perturbations of Riemann solutions of the inviscid equation (1.3) that contain only shock waves. There are two types of treatment of this equation of which we are aware: (1) One can show that if a nondegeneracy condition (Majda's stability condition) holds for each shock wave, the system can be solved by characteristics for all time [32]. (2) Assuming the same nondegeneracy condition, one can interpret the system as describing the scattering of incoming small shock waves by the large shock waves that comprise the original Riemann solution, and one can find sufficient conditions that guarantee that, in some norm, the total weight of the scattered shocks is smaller than the total weight of the incoming shocks [42], [4], [5], [49], [22], [21]. A condition of this type can then be used in Glimm's scheme to show the existence of solutions of (1.3) for initial data close to the original Riemann data. For a Riemann solution with $n = 2$ that consists of two Lax shocks, this approach yields a simple computable inviscid stability condition.

The system that determines the $O(1)$ eigenvalues to lowest order is also related to the SLEP system used by Nishiura and Fujii [35] for reaction-diffusion equations to study the stability of solutions with several sharp layers.

In this paper we study only the possible values of λ_0 for slow eigenvalues. The study of conditions under which λ_0 can actually be continued to a slow eigenvalue $\lambda = \lambda_0 + O(\epsilon)$ of the Riemann–Dafermos solution $u_\epsilon(x)$ is deferred to a later paper.

A necessary condition for stability of the Riemann–Dafermos solution is that no slow eigenvalue have positive real part. For $n = 2$, we show that to lowest order in ϵ, the $O(1)$ eigenvalues, other than -1, of a Riemann–Dafermos solution with two Lax shock waves all have the same real part. They are evenly spaced along a line in the complex plane. We compute the real part of these eigenvalues; the condition that it be negative turns out to be the $n = 2$ inviscid stability condition mentioned above. For $n > 2$, the relationship between the $O(1)$ eigenvalues and the known sufficient conditions for inviscid stability remains to be determined.

In section 9 we calculate slow eigenvalues other than -1 for Riemann solutions of the p-system that consist of two Lax shocks. They all have real part -2, independent of the Riemann solution. The calculation is essentially the same as the calculation of the inviscid stability criterion for these Riemann solutions in [4].

Thus, for Riemann–Dafermos solutions whose underlying Riemann solution consists of n Lax shock waves, our analysis suggests that they should be asymptotically stable if (1) each viscous shock wave is linearly stable, a matter that is determined by the wave's Evans function, and (2) the Riemann solution is stable in the inviscid sense, sufficient conditions for which have been determined by studying the scattering of small shock waves by large ones. The stability analysis of Riemann–Dafermos solutions thus unites two distinct lines of research. These relationships are explored in a little more detail in section 8.

A shortcoming of our analysis is that we do not address the possible existence of eigenvalues intermediate between fast and slow. This issue is discussed at the end of section 6. Its resolution may well involve Majda's stability condition, which is known to be related to the derivative of the Evans function at the origin [14], [3].

It should not be difficult to extend the results of this paper to more general diffusion matrices $B(u)$ or to general structurally stable Riemann solutions consisting entirely of shock waves, including undercompressive shock waves. However, we do not see how to deal with rarefactions, for which the asymptotic expansions are much more difficult due to loss of normal hyperbolicity in the underlying geometric singular perturbation problem [44].

2. Riemann solutions. In this section we define precisely the notion of a structurally stable Riemann solution consisting of Lax shock waves. A *Lax i-shock* for (1.3) that satisfies the viscous profile criterion for $B(u) \equiv I$ is a function

$$u(x) = \begin{cases} u^- & \text{for } x < s, \\ u^+ & \text{for } x > s, \end{cases} \tag{2.1}$$

with $x = \frac{X}{T}$, together with a solution $q(\xi)$ of the traveling wave ODE

$$\dot{u} = f(u) - f(u^-) - s(u - u^-), \tag{2.2}$$

such that the following hold:

(L1) $f(u^+) - f(u^-) - s(u^+ - u^-) = 0$.

(L2) The eigenvalues $\nu_1^- < \cdots < \nu_n^-$ of $Df(u^-)$ are real and distinct and satisfy $\nu_{i-1}^- < s < \nu_i^-$.

(L3) The eigenvalues $\nu_1^+ < \cdots < \nu_n^+$ of $Df(u^+)$ are real and distinct and satisfy $\nu_i^+ < s < \nu_{i+1}^+$.

(L4) $q(\xi)$ approaches u^- as $\xi \to -\infty$ and u^+ as $\xi \to \infty$.

Notice that (L1), (L2), and (L3) imply that for (2.2), $u^\pm$ are hyperbolic equilibria, the unstable manifold of u^- has dimension $n - i + 1$, and the stable manifold of u^+ has dimension i. Assumption (L4) says that these manifolds intersect. Because of the dimensions of the manifolds, generically, if they intersect, they do so in curves.

Remark 2.1. The function $q(\xi)$ is also a solution of

$$(Df(u) - sI)u_\xi = u_{\xi\xi} \tag{2.3}$$

and satisfies the boundary conditions (1.6).

A solution of the Riemann problem (1.3), (1.8) that consists of n Lax shock waves, each satisfying the viscous profile criterion for $B(u) \equiv I$, is a piecewise constant function

$$u_0(x) = \begin{cases} \bar{u}^0 & \text{for } x < \bar{s}^1, \\ \bar{u}^i & \text{for } \bar{s}^i < x < \bar{s}^{i+1}, \quad i = 1, \dots, n-1, \\ \bar{u}^n & \text{for } x > \bar{s}^n, \end{cases} \tag{2.4}$$

with $x = \frac{X}{T}$, together with $\mathbb{R}^n$-valued functions $q^i(\xi)$, $i = 1, \dots, n$, such that

(R1) $\bar{u}^0 = u^\ell$ and $\bar{u}^n = u^r$;

(R2) for each $i = 1, \dots, n$, the triple $(\bar{u}_{i-1}, \bar{s}_i, \bar{u}_i)$, together with the function $q^i(\xi)$, defines a Lax i-shock.

Define a mapping $G : \mathbb{R}^{n^2+2n} \to \mathbb{R}^{n^2}$ by

$$G(u^0, s^1, u^1, \dots, u^{n-1}, s^n, u^n) \\ = (f(u^1) - f(u^0) - s^1(u^1 - u^0), \dots, f(u^n) - f(u^{n-1}) - s^n(u^n - u^{n-1})).$$

Notice that

$$G(\bar{u}^0, \bar{s}^1, \bar{u}^1, \dots, \bar{u}^{n-1}, \bar{s}^n, \bar{u}^n) = 0. \tag{2.5}$$

The Riemann solution just defined is *structurally stable*, provided

(S1) $DG(\bar{u}^0, \bar{s}^1, \bar{u}^1, \dots, \bar{u}^{n-1}, \bar{s}^n, \bar{u}^n)$, restricted to the n^2-dimensional space of vectors $(U^0, S^1, U^1, \dots, U^{n-1}, S^n, U^n)$ with $U^0 = U^n = 0$, is invertible;

(S2) for each $i = 1, \dots, n$, the unstable manifold of $\bar{u}^{i-1}$ and the stable manifold of $\bar{u}^i$ for the traveling wave ODE $\dot{u} = f(u) - f(\bar{u}^{i-1}) - \bar{s}^i(u - \bar{u}^{i-1})$ meet transversally along $q^i(\xi)$.

If (S1) and (S2) are satisfied, then for each set of Riemann data (u^0, u^n) near $(\bar{u}^0, \bar{u}^n)$, there is a Riemann solution near the original one. Condition (S1) can be restated as follows:

(S1′) If we set $(U^0, U^n) = (0, 0)$, then the system of linear equations

$$(Df(\bar{u}^i) - \bar{s}^i I)U^i - (Df(\bar{u}^{i-1}) - \bar{s}^i I)U^{i-1} - S^i(\bar{u}^i - \bar{u}^{i-1}) = 0, \quad i = 1, \dots, n,$$

has only the trivial solution

$$(S^1, U^1, \dots, U^{n-1}, S^n) = (0, 0, \dots, 0, 0).$$

A condition equivalent to (S2) is the following:

(S2′) For each $i = 1, \dots, n$, the linear differential equation

$$(Df(q^i(\xi)) - \bar{s}^i I)U_\xi = U_{\xi\xi}$$

has, up to scalar multiplication, a unique solution that approaches zero as $\xi \to \pm\infty$. It is $q^i_\xi(\xi)$.

3. Stationary solutions. Consider the Riemann problem (1.3), (1.8). Assume that it has a solution (2.4) that consists of n Lax shock waves and is structurally stable. We shall study (1.17) together with the boundary conditions

$$u(-\infty, t) = u^\ell, \quad u(\infty, t) = u^r, \quad 0 \le t < \infty. \tag{3.1}$$

Stationary solutions $u_\epsilon(x)$ of (1.17), (3.1) satisfy

$$(Df(u) - xI)u_x = \epsilon u_{xx} \tag{3.2}$$

and the boundary conditions

$$u(-\infty) = u^\ell, \quad u(\infty) = u^r. \tag{3.3}$$

We shall look for $u_\epsilon(x)$ that lie near the given structurally stable Riemann solution (2.4). Such stationary solutions are known to exist, and to approach 0 exponentially as $x \to \pm\infty$, from the geometric singular perturbation arguments of [44].

In the regular layer, which is $\mathbb{R}$ with $\bar{s}^i$, $i = 1, \dots, n$, removed, $u_\epsilon(x)$ has an expansion of the form

$$u^R_\epsilon(x) \sim \sum \epsilon^j u^R_j(x), \tag{3.4}$$

in which $u_0^R(x)$ is just the piecewise constant Riemann solution (2.4). The regular layer is divided by the points $\bar{s}^i$ into $n+1$ connected sublayers

$$\begin{aligned} R^0 &= (-\infty, \bar{s}^1), \\ R^i &= (\bar{s}^i, \bar{s}^{i+1}), \quad i = 1, \dots, n-1, \\ R^n &= (\bar{s}^n, \infty). \end{aligned}$$

Each $u_j^R(x)$ is defined and piecewise C^∞ in the regular layer. At the jump points $\bar{s}^i$, we assume that each $u_j^R(x)$ has one-sided limits $u_j^R(x_0^i\pm) := \lim_{x\to x_0^i\pm} u_j^R(x)$. We assume that the same is true for all derivatives of the $u_j^R(x)$.

As explained in the introduction, we shall refer to a small neighborhood of $\bar{s}^i$ as the *ith singular layer* and denote it by S^i, $i = 1, \dots, n$. The Riemann–Dafermos solution $u_\epsilon(x)$ has sharp transition layers at

$$x^i(\epsilon) = \sum \epsilon^j x_j^i, \quad i = 1, \dots, n, \tag{3.5}$$

with $x^i(0) = \bar{s}^i$. Near $x^i(\epsilon)$ we use the stretched variable $\xi = \frac{x - x^i(\epsilon)}{\epsilon}$. In terms of this variable, the solution has an expansion

$$u_\epsilon^i(\xi) = \sum \epsilon^j u_j^i(\xi) \quad \text{in the singular layer } S^i. \tag{3.6}$$

The expansions $u_\epsilon^R(x)$ and $u_\epsilon^i(\xi)$ satisfy, respectively,

$$(Df(u^R) - xI)u_x^R = \epsilon u_{xx}^R, \tag{3.7}$$

$$(Df(u^i) - xI)u_\xi^i = u_{\xi\xi}^i, \quad x = x^i(\epsilon) + \epsilon\xi. \tag{3.8}$$

We first consider the regular layer. We substitute (3.4) into (3.7) and expand in powers of ϵ. At order ϵ^0 we obtain

$$(Df(u_0^R(x)) - xI)u_{0x}^R = 0.$$

We shall set $u_0^R(x)$ equal to the Riemann solution (2.4), which satisfies this equation.

In the regular layer, at order ϵ^1,

$$(Df(u_0^R(x)) - xI)u_{1x}^R = u_{0xx}^R = 0.$$

Thus $u_1^R(x)$ is constant on each regular sublayer. By induction, we can show that $u_j^R(x)$ is constant on each regular sublayer for all j.

We denote the constant value of $u_j^R(x)$ in R^i by $\bar{u}_j^i$. Thus

$$\bar{u}_0^i = \bar{u}^i, \quad i = 0, \dots, n.$$

From the boundary condition (3.3),

$$\bar{u}_j^0 = 0 \text{ for } j = 1, \dots, \infty, \quad \bar{u}_j^n = 0 \text{ for } j = 1, \dots, \infty. \tag{3.9}$$

Next, we consider the ith singular layer S^i. We substitute (3.6) and (3.5) into (3.8) and expand in powers of ϵ. At order ϵ^0, we obtain

$$(Df(u_0^i) - x_0^i I)u_{0\xi}^i = u_{0\xi\xi}^i. \tag{3.10}$$

To match the solutions at order ϵ^0 in regular and singular layers, we must have

$$u_0^i(-\infty) = \bar{u}_0^{i-1} = \bar{u}^{i-1} \quad \text{and} \quad u_0^i(\infty) = \bar{u}_0^i = \bar{u}^i. \tag{3.11}$$

We set

$$x_0^i = \bar{s}_i, \quad i = 1, \dots, n.$$

Then by (S2′) in section 2, (3.10), (3.11) has the solution $u_0^i(\xi) = q^i(\xi)$. As $\xi \to \pm\infty$, $q^i(\xi)$ approaches the limits exponentially fast. By (S2), the solution q^i is locally unique up to a shift in ξ.

In S^i, at order ϵ^1, we have

$$u_{1\xi\xi}^i - ((Df(q^i) - \bar{s}^i I)u_1^i)_\xi = -(x_1^i + \xi)q_\xi^i. \tag{3.12}$$

We look for u_1^i that satisfies the matching conditions

$$u_1^i(-\infty) = \bar{u}_1^{i-1} \quad \text{and} \quad u_1^i(\infty) = \bar{u}_1^i. \tag{3.13}$$

By (3.9), $\bar{u}_1^0 = \bar{u}_1^n = 0$. The other $\bar{u}_1^i$ and the x_1^i are to be determined.

Integrating (3.12) from $\xi = -\infty$ to $\xi = \infty$, we obtain

$$\begin{aligned}(Df(\bar{u}_0^i) - \bar{s}^i I)\bar{u}_1^i - (Df(\bar{u}_0^{i-1}) - \bar{s}^i I)\bar{u}_1^{i-1} - x_1^i(\bar{u}_0^i - \bar{u}_0^{i-1}) \\ = \int_{-\infty}^{\infty} \xi q_\xi^i d\xi, \quad i = 1, \dots, n.\end{aligned} \tag{3.14}$$

After making the substitutions $\bar{u}_1^0 = \bar{u}_1^n = 0$, (3.14) becomes a system of n^2 linear equations in the n^2 unknowns x_1^i, $i = 1, \dots, n$, and $\bar{u}_1^i$, $i = 1, \dots, n-1$. By (S1) there is a unique solution.

The space of bounded solutions of the adjoint system to the homogeneous part of (3.12), $\psi_{\xi\xi} + (Df(q^i) - \bar{s}^i I)\psi_\xi = 0$, is n-dimensional and consists of constant solutions. Therefore, using lemmas from [6], [24], condition (3.14) is necessary and sufficient for the existence of solutions $u_1^i(\xi)$ to (3.12) that satisfy the boundary conditions (3.13). For completeness, we state this fact as a lemma and present a simpler proof, taking advantage of the fact that (3.12) is in conservation form.

LEMMA 3.1. *Consider the system*

$$U_{\xi\xi} - ((Df(q^i(\xi)) - \bar{s}^i I)U)_\xi = g(\xi), \tag{3.15}$$

where $g(\xi)$ approaches zero exponentially as $\xi \to \pm\infty$. There is a solution U such that $U(\xi) \to U^\pm$ exponentially as $\xi \to \pm\infty$ if and only if

$$(Df(q^i(\infty)) - \bar{s}^i I)U^+ - (Df(q^i(-\infty)) - \bar{s}^i I)U^- + \int_{-\infty}^{\infty} g(s)ds = 0. \tag{3.16}$$

Proof. It is easy to see that the condition is necessary. We prove only that the condition is sufficient. The system (3.15) is equivalent to the system

$$U_\xi - (Df(q^i(\xi)) - \bar{s}^i I)U(\xi) + (Df(q^i(-\infty)) - \bar{s}^i I)U^- = G(\xi), \tag{3.17}$$

where $G(\xi) := \int_{-\infty}^{\xi} g(s)ds$ is bounded, $G(\xi) \to 0$ exponentially as $\xi \to -\infty$, and $G(\xi) \to \int_{-\infty}^{\infty} g(s)ds$ exponentially as $\xi \to \infty$.

From the definition of a Lax i-shock, $Df(q^i(\pm\infty)) - \bar{s}^i I$ is hyperpolic, so system (3.17) has exponential dichotomies [7] on $\mathbb{R}^{\pm}$. Therefore there exist nonunique bounded solutions $U_L(\xi)$ and $U_R(\xi)$ that solve (3.17) on $\mathbb{R}^-$ and $\mathbb{R}^+$, respectively.

For the dichotomy on $\mathbb{R}^-$, let $P_u(0-)$ denote projection onto the unstable subspace at $x = 0$, with kernel the stable subspace. Similarly, for the dichotomy on $\mathbb{R}^+$, let $P_s(0+)$ denote projection onto the stable subspace at $x = 0$, with kernel the unstable subspace. Then the definition of a Lax i-shock implies that $\mathcal{R}P_u(0-)+\mathcal{R}P_s(0+) = \mathbb{R}^n$. Therefore there exists a (nonunique) pair (ϕ_u, ϕ_s) such that

$$\begin{aligned} &U_L(0-) + \phi_u = U_R(0+) + \phi_s, \\ &\phi_u \in \mathcal{R}P_u(0-), \quad \phi_s \in \mathcal{R}P_s(0+). \end{aligned}$$

Let $\Phi(\xi, \zeta)$ be the principle matrix solution to (3.17). The solution $U(\xi), \xi \in \mathbb{R}$, can be obtained by letting

$$\begin{aligned} U(\xi) &= U_L(\xi) + \Phi(\xi, 0)\phi_u, \quad \xi \ \leq 0, \\ U(\xi) &= U_R(\xi) + \Phi(\xi, 0)\phi_s, \quad \xi \ \geq 0. \end{aligned}$$

From (3.17) and (3.16), using the limits of $G(\xi)$ as $\xi \to \pm\infty$, it is easy to show that $U(\xi) \to U^-$ as $\xi \to -\infty$ and $U(\xi) \to U^+$ as $\xi \to \infty$. ∎

Proceeding inductively, we can solve for all x_j^i and $\bar{u}_j^i$.

Our asymptotic expansions are justified by the fact that $u_\epsilon(x)$ is known to exist from the geometric singular perturbation theory arguments of [44]. Alternatively, a proof of existence of the exact stationary solutions $u_\epsilon(x)$ can be based on the existence of the formal asymptotic expansions (3.4)–(3.5). For this approach to singular perturbation theory, see [25]. The same assumptions (S1) and (S2) are used in both types of arguments.

We summarize the results about stationary solutions in the following.

PROPOSITION 3.2. *In the regular layer, to all orders of* ϵ, $u_\epsilon^R(x)$ *is piecewise constant with jumps at* $x_0^i(\epsilon)$, $i = 1, \ldots, n$, *only. At lowest order,* $u_0^R(x)$ *is the Riemann solution* (2.4). *In the* i*th singular layer, at lowest order,* $u_0^i(\xi) = q^i(\xi)$, *a heteroclinic solution connecting the states* $\bar{u}_0^{i-1}$ *and* $\bar{u}_0^i$. *Higher order terms* $u_j^R(x)$, $u_j^i(\xi)$, *and* x_j^i *can be obtained recursively, using the matching of regular and singular layers and Lemma* 3.1.

4. Well-posedness. To show the well-posedness of initial value problems with initial conditions near a Riemann–Dafermos solution, it is convenient to use the stretched variables $\xi = \frac{x}{\epsilon}$ and $\tau = \frac{t}{\epsilon}$. We shall translate the results back to (x, t) variables at the end of the section.

Using the stretched variables, (1.17) becomes

$$u_\tau + (Df(u) - \epsilon\xi I)u_\xi = u_{\xi\xi}. \tag{4.1}$$

Let $u_\epsilon(x)$ be a stationary solution of (1.17), (3.1). Then $u_\epsilon(\epsilon\xi)$ is a stationary solution of (4.1). A solution of (4.1) near $u_\epsilon(\epsilon\xi)$ can be expressed as $u_\epsilon(\epsilon\xi) + U(\xi, \tau)$ with U satisfying

$$U_\tau + (Df(u_\epsilon + U) - \epsilon\xi I)U_\xi + (Df(u_\epsilon + U) - Df(u_\epsilon))u_{\epsilon\,\xi} = U_{\xi\xi}. \tag{4.2}$$

For any $\rho \geq 0$, let $C(\rho, \mathbb{R}_\xi)$ be the Banach space of uniformly continuous functions $U(\xi), \xi \in \mathbb{R}$, such that the weighted norm $|U|_\rho := \sup_\xi |U(\xi)|e^{\rho|\xi|} < \infty$. Let $C^k(\rho, \mathbb{R}_\xi)$

be the space of functions $U(\xi)$ such that $U, U', \dots, U^{(k)} \in C(\rho, \mathbb{R}_\xi)$. On $C^k(\rho, \mathbb{R}_\xi)$ we use the norm $|U|_{k,\rho} := |U|_\rho + |U'|_\rho + \cdots + |U^{(k)}|_\rho$. One can define $C(\rho, \mathbb{R}^\pm_\xi)$ and $C^k(\rho, \mathbb{R}^\pm_\xi)$ similarly.

We shall show that for any $\rho \geq 0$, (4.2) is well-posed for small initial data in $C^2(\rho, \mathbb{R}_\xi)$. The intuitive reason is that for the underlying Riemann problem, the characteristics on the two unbounded regular layers head inward. This keeps a space of exponentially decaying profiles invariant.

Stronger results can be obtained using fractional powers of Banach spaces or intermediate spaces [13], [17], [36], [9], [31], [23]. We choose to use $C^2(\rho, \mathbb{R}_\xi)$ for simplicity.

We rewrite (4.2) as

$$U_\tau + (Df(u_\epsilon) - \epsilon\xi I)U_\xi + D^2 f(u_\epsilon)u_{\epsilon\,\xi}U + g_\epsilon(U, U_\xi, \xi) = U_{\xi\xi}, \tag{4.3}$$

with

$$\begin{aligned} &g_\epsilon(U, U_\xi, \xi) \\ &\quad = (Df(u_\epsilon + U) - Df(u_\epsilon))U_\xi + (Df(u_\epsilon + U) - Df(u_\epsilon) - D^2 f(u_\epsilon)U)u_{\epsilon\,\xi}. \end{aligned} \tag{4.4}$$

Note that because of the dependence on U_ξ in (4.4), if $U \in C^2(\rho, \mathbb{R}_\xi)$, then $g_\epsilon \in C^1(\rho, \mathbb{R}_\xi)$. Moreover, we have

$$\begin{aligned} |g_\epsilon(U)|_{1,\rho} &\leq C|U|^2_{2,\rho}, \\ |g_\epsilon(U_1) - g_\epsilon(U_2)|_{1,\rho} &\leq C \max\{|U_1|_{2,\rho}, |U_2|_{2,\rho}\}|U_1 - U_2|_{2,\rho}. \end{aligned} \tag{4.5}$$

We first consider the inhomogeneous linear system

$$U_\tau + (Df(u_\epsilon) - \epsilon\xi I)U_\xi + D^2 f(u_\epsilon)u_{\epsilon\,\xi}U + h_\epsilon(\xi, \tau) = U_{\xi\xi}. \tag{4.6}$$

The hypotheses on h in the following lemma are motivated by the observations just made about g.

PROPOSITION 4.1. *Let $\tau_0 > 0$, $\epsilon_0 > 0$, and $\rho \geq 0$. Assume that*

(1) *for each $0 < \epsilon \leq \epsilon_0$, $h_\epsilon(\cdot, \tau)$ is a continuous mapping from $0 \leq \tau \leq \tau_0$ to $C^1(\rho, \mathbb{R}_\xi)$;*
(2) *there is a constant M such that $|h_\epsilon(\cdot, \tau)|_{1,\rho} \leq M$ on $\{(\tau, \epsilon) : 0 \leq \tau \leq \tau_0, 0 < \epsilon \leq \epsilon_0\}$.*

Let

$$U(\xi, 0) = \phi(\xi), \tag{4.7}$$

with $\phi \in C^2(\rho, \mathbb{R}_\xi)$. Then there exists τ_1, $0 < \tau_1 \leq \tau_0$, such that for each $0 < \epsilon \leq \epsilon_0$, the initial value problem (4.6), (4.7) has a solution $U(\xi, \tau)$, $0 \leq \tau \leq \tau_1$. The mapping $\tau \to U(\cdot, \tau)$ is continuous from $[0, \tau_1]$ to $C^2(\rho, \mathbb{R}_\xi)$, and there is a constant C such that, for each (τ, ϵ),

$$|U|_{2,\rho} \leq C(|\phi|_\rho + |h|_{1,\rho}).$$

The numbers τ_1 and C depend on ϵ_0 but are independent of ρ and M.

Proof. Let $y = e^{\epsilon\tau}\xi$ and define $v(y, \tau) := U(e^{-\epsilon\tau}y, \tau)$. Then

$$v_\tau + e^{\epsilon\tau}(Df(u_\epsilon)v_y + D^2 f(u_\epsilon)u_{\epsilon\,y}v) + h = e^{2\epsilon\tau}v_{yy}.$$

Let $s = \frac{e^{2\epsilon\tau}-1}{2\epsilon}$, so that $\tau = \tau(s) = (1+2\epsilon s)^{\frac{1}{2\epsilon\tau}}$. Let $w(y,s) := v(y,\tau(s))$. Then

$$\begin{aligned} &w_s + \frac{1}{\sqrt{2\epsilon s+1}}(Df(u_\epsilon)w_y + D^2 f(u_\epsilon)u_{\epsilon y}w) + \frac{1}{2\epsilon s+1}h = w_{yy}, \\ &w(y,0) = \phi(y). \end{aligned} \tag{4.8}$$

Moreover, if τ_1 is sufficiently small, then for each $0 < \epsilon \le \epsilon_0$, h_ϵ defines a continuous function from $0 \le s \le s_1(\epsilon)$ to $C^1(\rho, \mathbb{R}_\xi)$, where $s_1(\epsilon) = \frac{e^{2\epsilon\tau_1}-1}{2\epsilon} \approx \tau_1$. In (4.8) the coefficients of w and w_y, and the inhomogeneous term, are bounded on

$$\{(y,s,\epsilon) : y \in \mathbb{R}, 0 \le s \le s_1(\epsilon), 0 < \epsilon \le \epsilon_0\}.$$

Let $\Phi(y,s) := \frac{1}{2\sqrt{\pi s}}e^{-y^2/4s}$ be the fundamental solution of the heat equation $w_s = w_{yy}$. The solution of (4.8) is the fixed point of the integral equation

$$\begin{aligned} \bar{w}(y_0,s_0) &= \int_{-\infty}^{\infty} \Phi(y_0-y,s_0)\phi(y)dy - \int_0^s \int_{-\infty}^{\infty} \Phi(y_0-y,s_0-s)\frac{1}{2\epsilon s+1}h_\epsilon(y,s)dyds \\ &- \int_0^s \int_{-\infty}^{\infty} \Phi(y_0-y,s_0-s)\frac{1}{\sqrt{2\epsilon s+1}}(Df(u_\epsilon)w_y(y,s) + D^2 f(u_\epsilon)u_{\epsilon y}w(y,s))dyds. \end{aligned}$$

If $w(y,s)$ defines a continuous function from $0 \le s \le s_1(\epsilon)$ to $C^2(\rho,\mathbb{R}_\xi)$, then it is easy to show that $\bar{w}$ defines a continuous function from $0 \le s \le s_1(\epsilon)$ to $C^2(\rho,\mathbb{R}_\xi)$. Moreover, if τ_1 is sufficiently small, then, independent of ρ, the mapping $w \to \bar{w}$ is a contraction mapping in the space of continuous functions from $0 \le s \le s_1(\epsilon)$ to $C^2(\rho,\mathbb{R}_\xi)$. Therefore, there exists a unique fixed point $w(y,s)$ in $C^2(\rho,\mathbb{R}_\xi)$, which is the solution of (4.8).

Then

$$|U(\xi,\tau)| = |v(y,\tau(s))| = |w(y,s)| \le C(|\phi|_\rho + |h|_{1,\rho})e^{-\rho|y|} \le C_1(|\phi|_\rho + |h|_{1,\rho})e^{-\rho|\xi|}.$$

Similar estimates for $|U_\xi|$ and $|U_{\xi\xi}|$ can also be obtained from the integral equation for w. The proof that $w : [0,\tau_1] \to C^2(\rho,\mathbb{R}_\xi)$ is continuous uses a well-known technique from the theory of evolution equations in abstract Banach spaces [17] and will be omitted. □

Using Proposition 4.1, the estimates (4.5), and the contraction mapping theorem in $C^2(\rho,\mathbb{R}_\xi)$, we can easily prove the following proposition.

PROPOSITION 4.2. *Consider the initial value problem* (4.2), (4.7), *with* $\phi \in C^2(\rho,\mathbb{R}_\xi)$ *and* $\rho \ge 0$. *There exist positive constants* τ_1, ϵ_1, *and* δ_1, *all independent of* ρ, *such that if* $|\phi|_{2,\rho} \le \delta_1$, *then for each* $0 < \epsilon \le \epsilon_1$, *the initial value problem has a unique solution* $U(\xi,\tau)$, $0 \le \tau \le \tau_1$, *such that* $\tau \to U(\cdot,\tau)$ *is a continuous mapping from* $[0,\tau_1]$ *to* $C^2(\rho,\mathbb{R}_\xi)$.

We can apply Proposition 4.2 repeatedly until the maximal time interval of existence is reached.

We recall from the introduction that for a C^k function $\psi(x)$, we define

$$|\psi|_{k,\gamma,\epsilon} := |\psi|_\gamma + \epsilon|\psi'|_\gamma + \cdots + \epsilon^k|\psi^{(k)}|_\gamma.$$

LEMMA 4.3. *Let* k *be a nonnegative integer. Let* $\psi \in C^k(\gamma,R_x)$. *Define* $\phi(\xi) = \psi(\epsilon\xi)$. *Then* $\phi \in C^k(\epsilon\gamma,R_\xi)$, *and* $|\phi|_{k,\epsilon\gamma} = |\psi|_{k,\gamma,\epsilon}$.

Proof. We have

$$|\psi(x)|e^{\gamma|x|} = |\psi(\epsilon\xi)|e^{\gamma\epsilon|\xi|} = |\phi(\xi)|^{\epsilon\gamma|\xi|},$$
$$\epsilon|\psi_x(x)|e^{\gamma|x|} = \epsilon|\psi_x(\epsilon\xi)|e^{\gamma\epsilon|\xi|} = |\phi_\xi(\xi)|e^{\epsilon\gamma|\xi|}, \tag{4.9}$$

etc. The result follows. □

In the original variables $x = \epsilon\xi$ and $t = \epsilon\tau$, (4.2) becomes

$$V_t + (Df(u_\epsilon + V) - xI)V_x + (Df(u_\epsilon + V) - Df(u_\epsilon))u_{\epsilon\, x} = V_{xx}, \tag{4.10}$$

and the initial condition (4.7) becomes

$$V(x, 0) = \psi(x). \tag{4.11}$$

COROLLARY 4.4. *Consider the initial value problem* (4.10), (4.11), *with* $\psi \in C^2(\gamma, \mathbb{R}_x)$ *and* $\gamma \geq 0$. *There exist positive constants* τ_1, ϵ_1, *and* δ_1, *all independent of* γ, *such that if* $|\psi|_{2,\gamma,\epsilon} \leq \delta_1$, *then for each* $0 < \epsilon \leq \epsilon_1$, *there is a unique solution* $V(x, t)$, $0 \leq t \leq \epsilon\tau_1$, *such that* $t \to V(\cdot, t)$ *is a continuous mapping from* $[0, \epsilon\tau_1]$ *to* $C^2(\gamma, \mathbb{R}_x)$.

Proof. The constants τ_1, ϵ_1, and δ_1 are those of Proposition 4.2. Suppose $|\psi|_{2,\gamma,\epsilon} \leq \delta_1$. Let $\phi(\xi) = \psi(\epsilon\xi)$. By Lemma 4.3, $|\phi|_{2,\epsilon\gamma} = |\psi|_{2,\gamma,\epsilon}$. By Proposition 4.2, the initial value problem (4.2), (4.7) has a solution $U(\xi, \tau)$, $0 \leq \tau \leq \tau_1$. Let $V(x, t) = U(\frac{x}{\epsilon}, \frac{t}{\epsilon})$. □

As noted in the introduction, the condition $|\psi|_{2,\gamma,\epsilon} \leq \delta_1$ allows, for small $\epsilon > 0$, initial perturbations of the Riemann–Dafermos solution $u_\epsilon(x)$ with very large derivatives.

5. Essential spectrum. In the space of uniformly bounded functions, a traveling wave (viscous shock) solution of (1.1) has an essential spectrum that touches the imaginary axis. This is the main difficulty in proving stability of the traveling wave. The same difficulty occurs for a Riemann–Dafermos solution u_ϵ of the Dafermos regularization. Following an idea of Sattinger [39], we use weighted function spaces to move the essential spectrum to the left.

Let $\tilde{\delta} > 0$ be given. For sufficiently large $\gamma > 0$, we shall show that, for small $\epsilon > 0$, in the space $C^2(\gamma, \mathbb{R}_x)$, the essential spectrum of the linearization of (1.17) about a Riemann–Dafermos solution $u_\epsilon(x)$ lies in the region $\mathrm{Re}\tilde{\lambda} \leq -\tilde{\delta}$. Therefore the stability of the Riemann–Dafermos solution is determined by the eigenvalues.

Let $T(\xi, \zeta)$ be the fundamental matrix solution for a first-order system

$$W_\xi = B(\xi)W, \quad \xi \in J. \tag{5.1}$$

DEFINITION 5.1. *Let* $\beta < \alpha$ *be real numbers. System* (5.1) *has a* pseudoexponential dichotomy on J with spectral gap $\beta < \alpha$ *if there is a real number* $C \geq 0$ *and projections* $P(\xi)$, $\xi \in J$, *such that*

(1) $T(\xi, \zeta)P(\zeta) = P(\xi)T(\xi, \zeta)$;

(2) *if* $w_s \in \mathcal{R}P(\zeta)$, *and* $\xi > \zeta$ *in* J, *then*

$$|T(\xi, \zeta)w_s| \leq Ce^{\beta(\xi-\zeta)}|w_s|;$$

(3) *if* $w_u \in \mathcal{R}(I - P(\zeta))$, *and* $\xi < \zeta$ *in* J, *then*

$$|T(\xi, \zeta)w_u| \leq Ce^{\alpha(\xi-\zeta)}|w_u|;$$

(4) $P(\xi)$ *is continuous with respect to* ξ.

Notice that $P(\xi)$ is not assumed to be uniformly bounded.

The linearization of (1.17) about $u_\epsilon(x)$ is

$$U_t + (Df(u_\epsilon) - xI)U_x + D^2 f(u_\epsilon) u_{\epsilon\, x} U = \epsilon U_{xx}. \tag{5.2}$$

The complex number $\tilde{\lambda}$ is in the resolvent set of (5.2), provided the spectral equation

$$\tilde{\lambda} U + (Df(u_\epsilon) - xI)U_x + D^2 f(u_\epsilon) u_{\epsilon\, x} U + \tilde{h} = \epsilon U_{xx} \tag{5.3}$$

can be solved for U in terms of $\tilde{h}$, and the mapping $\tilde{h} \to U$ is bounded.

In (5.3) let $\lambda = \epsilon\tilde{\lambda}$, $\xi = \frac{x}{\epsilon}$, and $h = \epsilon\tilde{h}$. Then (5.3) becomes

$$\lambda U + (Df(u_\epsilon) - \epsilon\xi I)U_\xi + D^2 f(u_\epsilon) u_{\epsilon\, \xi} U + h = U_{\xi\xi}. \tag{5.4}$$

Let $\tilde{\delta} > 0$ be given. We shall show that for $\epsilon > 0$ sufficiently small and $\mathrm{Re}\lambda \geq -\epsilon\tilde{\delta}$, (5.4) with $h = 0$ has, for an appropriate $a > 0$, pseudoexponential dichotomies on the intervals $\xi \leq -\frac{a}{\epsilon}$ and $\xi \geq \frac{a}{\epsilon}$. Although the projection operators $P(\lambda, \epsilon, \xi)$ of the pseudoexponential dichotomies are not uniformly bounded, even for fixed (λ, ϵ), we will show that the restriction of $P(\lambda, \epsilon, \xi)$ to the subspace of $\mathbb{R}^{2n}$ defined by setting the first n coordinates equal to zero is uniformly bounded. Based on these results we will show that for $\epsilon > 0$ sufficiently small, the essential spectrum of (5.3) is in the region $\mathrm{Re}\tilde{\lambda} \leq -\tilde{\delta}$.

Let $W = (U, V)$ and let

$$\tilde{B}(\lambda, \epsilon, x) := \begin{pmatrix} 0 & I \\ \lambda + \epsilon D^2 f(u_\epsilon) u_{\epsilon\, x} & Df(u_\epsilon) - xI \end{pmatrix}. \tag{5.5}$$

Let

$$B(\lambda, \epsilon, \xi) := \tilde{B}(\lambda, \epsilon, \epsilon\xi) = \begin{pmatrix} 0 & I \\ \lambda + D^2 f(u_\epsilon) u_{\epsilon\, \xi} & Df(u_\epsilon) - \epsilon\xi I \end{pmatrix}. \tag{5.6}$$

Then (5.4) can be recast as

$$W_\xi = B(\lambda, \epsilon, \xi) W + (0,\ h)^\top. \tag{5.7}$$

Our proof that $W_\xi = BW$ has pseudoexponential dichotomies on the intervals $\xi \leq -\frac{a}{\epsilon}$ and $\xi \geq \frac{a}{\epsilon}$ is motivated by the proof of Coppel's Proposition 1 [7, p. 50]. This result says, roughly speaking, that if the matrices $B(\xi)$, $\xi \in J$, are uniformly bounded and uniformly hyperbolic, and vary slowly with ξ, then (5.1) has an exponential dichotomy on J. Our case differs in that the matrices $B(\lambda, \epsilon, \xi)$ are not uniformly bounded, even for fixed (λ, ϵ). In addition, they have eigenvalues near 0 for small ϵ, so we are interested in pseudoexponential dichotomies rather than exponential dichotomies.

Let

$$\tilde{A}(\lambda, x) := \begin{pmatrix} 0 & I \\ \lambda & Df(u^r) - xI \end{pmatrix}.$$

LEMMA 5.1. *For $\delta > 0$ sufficiently small, there are numbers $\beta(\delta) < \alpha(\delta) < 0$ such that if $\mathrm{Re}\lambda \geq -\delta$ and $x_0^n \leq x$, then $\tilde{A}(\lambda, x)$ has n eigenvalues with real parts less than $\beta(\delta)$ and n eigenvalues with real parts between $\alpha(\delta)$ and 0. As $\delta \to 0$, $\beta(\delta)$ approaches a negative limit, and $\alpha(\delta)$ is $O(\delta)$.*

Proof. Since (1.3) is strictly hyperbolic, the eigenvalues of $Df(u^r)$ are real and distinct. Denote them by $\nu_1 < \cdots < \nu_n$ and denote the corresponding eigenvectors by $\mathbf{r}_1, \ldots, \mathbf{r}_n$.

Let μ be an eigenvalue of $\tilde{A}(\lambda, x)$. It is easily verified that

$$\det(\lambda + \mu(Df(u^r) - (x+\mu)I)) = 0.$$

Therefore one of the following equations must hold:

$$\mu^2 + (x - \nu_j)\mu - \lambda = 0, \quad j = 1, \ldots, n.$$

Thus there are two eigenvalues of $\tilde{A}(\lambda, x)$ for each j,

$$\mu_j^{\pm} = -\frac{-x - \nu_j}{2} \pm \sqrt{\left(\frac{x - \nu_j}{2}\right)^2 + \lambda},$$

with corresponding eigenvectors

$$(\mathbf{r}_j, \mu_j^{\pm}\mathbf{r}_j)^{\top}.$$

For each x with $x_0^n \le x$, we have $\nu_n < x$. Let $p = \frac{1}{2}(x_0^n - \nu_n) > 0$. Let δ be such that $0 < \delta < p^2$. Let

$$\beta(\delta) = -p - \sqrt{p^2 - \delta}, \quad \alpha(\delta) = -p + \sqrt{p^2 - \delta} = \frac{-\delta}{p + \sqrt{p^2 - \delta}}.$$

Notice that $\beta(\delta) < \alpha(\delta) < 0$, $\lim_{\delta \to 0} \beta(\delta) = -2p < 0$, and $\alpha(\delta)$ is $O(\delta)$.

Let $1 \le j \le n$, let $\mathrm{Re}\lambda \ge -\delta$, and let $x_0^n \le x$. From Corollary 5.6 at the end of this section, with $r = p_j = \frac{1}{2}(x - \nu_j)$, $\mu_j^{\pm}$ must satisfy

$$\mathrm{Re}\mu_j^- \le -p_j - \sqrt{p_j^2 - \delta} \le \beta(\delta)$$

and

$$\mathrm{Re}\mu_j^+ \ge -p_j + \sqrt{p_j^2 - \delta} = \frac{-\delta}{p_j + \sqrt{p_j^2 - \delta}} \ge \alpha(\delta). \tag{5.8}$$ ∎

We shall refer to the μ_j^-, $j = 1, \ldots, n$, as *pseudostable eigenvalues* and the μ_j^+, $j = 1, \ldots, n$, as *pseudounstable eigenvalues*.

We now construct projections associated to the pseudostable and pseudounstable eigenvalues.

Let $\mathbf{R} = (\mathbf{r}_1 \ldots \mathbf{r}_n)$ and $\mathcal{M}^{\pm}(\lambda, x) = \mathrm{diag}(\mu_1^{\pm} \ldots \mu_n^{\pm})$ be $n \times n$ matrices. The eigenvectors of $\tilde{A}(\lambda, x)$ form a $2n \times 2n$ matrix

$$H(\lambda, x) := \begin{pmatrix} \mathbf{R} & 0 \\ 0 & \mathbf{R} \end{pmatrix} \begin{pmatrix} I_n & I_n \\ \mathcal{M}^- & \mathcal{M}^+ \end{pmatrix}.$$

The first n columns of H are eigenvectors $(\mathbf{r}_j, \mu_j^-\mathbf{r}_j)^{\top}$ for the corresponding μ_j^-, and the last n columns are eigenvectors $(\mathbf{r}_j, \mu_j^+\mathbf{r}_j)^{\top}$ for the corresponding μ_j^+. Let $D(\lambda, x) = \mathcal{M}^+ - \mathcal{M}^- = \mathrm{diag}(\mu_j^+ - \mu_j^-)$. Then

$$H^{-1} = \begin{pmatrix} \mathcal{M}^+ D^{-1} & -D^{-1} \\ -\mathcal{M}^- D^{-1} & D^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{R}^{-1} & 0 \\ 0 & \mathbf{R}^{-1} \end{pmatrix}.$$

Let $\tilde{P} = \begin{pmatrix} I_n & 0 \\ 0 & 0 \end{pmatrix}$. The projection to the space spanned by the pseudostable eigenvectors is

$$P(\lambda, x) = H\tilde{P}H^{-1} = \begin{pmatrix} \mathbf{R} & 0 \\ 0 & \mathbf{R} \end{pmatrix} \begin{pmatrix} \mathcal{M}^+ D^{-1} & -D^{-1} \\ -\lambda D^{-1} & -\mathcal{M}^- D^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{R}^{-1} & 0 \\ 0 & \mathbf{R}^{-1} \end{pmatrix}.$$

Here we have used $\mathcal{M}^-\mathcal{M}^+ = -\lambda I_n$.

PROPOSITION 5.2. *Let $\tilde{\delta} > 0$. Let $a > \max_{i=1,\dots,n} |x_0^i|$. Then for $\epsilon > 0$ sufficiently small and $\mathrm{Re}\lambda \geq -\epsilon\tilde{\delta}$, $W_\xi = BW$ has pseudoexponential dichotomies with n-dimensional pseudostable and pseudounstable spaces on $\xi \leq -\frac{a}{\epsilon}$ and on $\xi \geq \frac{a}{\epsilon}$. The spectral gaps are $0 < \beta_1\epsilon < \alpha_1$ for $\xi \leq -\frac{a}{\epsilon}$ and $\beta_2 < \alpha_2\epsilon < 0$ for $\xi \geq \frac{a}{\epsilon}$. The numbers α_j and β_j, $j = 1, 2$, are independent of λ. The constant C in the definition of pseudoexponential dichotomy is independent of (λ, ϵ).*

Proof. We will consider only the interval $\xi \geq \frac{a}{\epsilon}$, since the interval $\xi \leq -\frac{a}{\epsilon}$ can be handled similarly.

From section 3, on the interval $x \geq a$, $u_\epsilon(x) - u^r$ is 0 to any finite order in ϵ. Thus on the interval $\xi \geq \frac{a}{\epsilon}$, $W_\xi = BW$ is approximately $W_\xi = AW$, with $W = (U, V)$ and

$$A(\lambda, \epsilon, \xi) := \tilde{A}(\lambda, \epsilon\xi) = \begin{pmatrix} 0 & I \\ \lambda & Df(u^r) - \epsilon\xi I \end{pmatrix}.$$

Let $\delta = \delta(\epsilon) = \epsilon\tilde{\delta}$. Choose $\tilde{\epsilon} > 0$ such that $\tilde{\epsilon}\tilde{\delta}$ is small enough that Lemma 5.1 applies. In the following we consider only ϵ with $0 < \epsilon < \tilde{\epsilon}$.

Let $\mathcal{M}(\lambda, x) := \mathrm{diag}(\mathcal{M}^-, \mathcal{M}^+)$. Then $\tilde{A} = H\mathcal{M}H^{-1}$. Consider the (λ, x)-dependent change of variables $W = HZ$. After making the substitution $x = \epsilon\xi$, $W_\xi = AW$ becomes

$$Z_\xi = \mathcal{M}Z - H^{-1}H_\xi Z. \tag{5.9}$$

The differential equation (5.9) is a perturbation of the diagonalized system

$$Z_\xi = \mathcal{M}Z. \tag{5.10}$$

That is, $z_j' = \mu_j^- z_j$ if $1 \leq j \leq n$ and $z_j' = \mu_{j-n}^+ z_j$ if $n+1 \leq j \leq 2n$. For $0 < \epsilon < \tilde{\epsilon}$, system (5.10) has a pseudoexponential dichotomy with projection $\tilde{P}$ and spectral gap $\beta(\delta) < \alpha(\delta) < 0$, with $\delta = \epsilon\tilde{\delta}$.

It is easily verified that there is a constant C, independent of δ for δ sufficiently small, such that

$$\frac{1 + |\mu_j^-| + |\mu_j^+|}{|\sqrt{(x - \nu_j)^2 + 4\lambda}|} \leq C$$

for all (j, λ, x), with $j = 1, \dots, n$, $\mathrm{Re}\lambda \geq -\delta$ and $x_0^n \leq x$. Therefore $|H^{-1}| \leq C$ uniformly with respect to (λ, x). Moreover, using $x = \epsilon\xi$, we have

$$\partial\mu_j^\pm/\partial\xi = \frac{-\epsilon \pm \epsilon(x - \nu_j)((x - \nu_j)^2 + 4\lambda)^{-\frac{1}{2}}}{2} = O(\epsilon)$$

for all (j, λ, x). Therefore $H^{-1}H_\xi = O(\epsilon)$. From this, one can show by an argument similar to the proof of roughness of exponential dichotomies that for sufficiently small

ϵ, (5.9) also has a pseudoexponential dichotomy on $\xi \geq \frac{a}{\epsilon}$. The projection, which we denote by $\tilde{Q}(\lambda, \epsilon, \xi)$, is $O(\epsilon)$ close to $\tilde{P}$. For appropriate negative constants α_2 and β_2, the spectral gap is $\beta_2 < \alpha_2\epsilon < 0$. The constant C in the definition of pseudoexponential dichotomy is independent of (λ, ϵ).

Because the system $W_\xi = AW$ is just (5.9) after a linear change of variables, it also has a pseudoexponential dichotomy on $\xi \geq \frac{a}{\epsilon}$ with spectral gap $\beta_2 < \alpha_2\epsilon < 0$.

The matrices A and B differ by $O(\epsilon)$ terms that are in the last n rows only. Existence of a pseudoexponential dichotomy on $\xi \geq \frac{a}{\epsilon}$ for $W_\xi = BW$ then follows by an argument similar to the proof of roughness of exponential dichotomies. □

The pseudoexponential dichotomy for $W_\xi = AW$ has the projection $\bar{Q} := H\tilde{Q}H^{-1}$ $= H(\tilde{P} + O(\epsilon))H^{-1} = O(1 + \epsilon|x| + \sqrt{|\lambda|})$, which can be large for large ξ and $|\lambda|$.

LEMMA 5.3. *Let $Q(\lambda, \xi)$ be the projection for the pseudoexponential dichotomy for $W_\xi = BW$. Then $|Q(\lambda, \xi)(I - \tilde{P})|$ is uniformly bounded for all (λ, ξ) with $\mathrm{Re}\lambda \geq -\delta$ and $|\xi| \geq \frac{a}{\epsilon}$.*

Proof. We will show the result for $W_\xi = AW$. The result for $W_\xi = BW$ then follows by an argument similar to the proof of roughness of exponential dichotomies.

Observe that

$$\bar{Q}(I - \tilde{P}) = H\tilde{Q}H^{-1}(I - \tilde{P}),$$
$$|\bar{Q}(I - \tilde{P})| \leq |H||\tilde{Q}||H^{-1}(I - \tilde{P})|.$$

Using the facts

$$|H| \leq C(1 + |\mathcal{M}^-| + |\mathcal{M}^+|),$$
$$|\tilde{Q}| \leq C,$$
$$|H^{-1}(I - \tilde{P})| \leq C|(\mathcal{M}^+ - \mathcal{M}^-)^{-1}|,$$

we obtain that $|\bar{Q}(I - \tilde{P})|$ is uniformly bounded with respect to (λ, ϵ, ξ) in the domain of consideration. □

Let γ be a constant such that $\gamma > \max\{-\alpha_2, \beta_1\}$. We now show that in the function space $C(\gamma, \mathbb{R}_x)$, the region $\mathrm{Re}\tilde{\lambda} \geq -\tilde{\delta}$ consists of normal points only. Observe that in the ξ-coordinate, the space is $C(\epsilon\gamma, \mathbb{R}_\xi)$.

Without loss of generality, assume that $x = 0$ is between x_0^1 and x_0^n. Consider the nonhomogeneous equation (5.4), where $h \in C(\epsilon\gamma, \mathbb{R}_\xi)$. This is equivalent to the first-order system

$$W_\xi = BW + (0,\ h)^\top. \tag{5.11}$$

By Proposition 5.2, the associated homogeneous system of (5.11) has pseudoexponential dichotomies on $\xi \leq -\frac{a}{\epsilon}$ and $\xi \geq \frac{a}{\epsilon}$. These dichotomies can be extended from $(-\infty, -\frac{a}{\epsilon}]$ to $\mathbb{R}^-$ and from $[\frac{a}{\epsilon}, \infty)$ to $\mathbb{R}^+$. The constants of the extended dichotomies are ϵ dependent and may approach ∞ as $\epsilon \to 0$, but the exponents remain the same. If, for certain λ, the n-dimensional pseudounstable space at $\xi = 0-$ has a nontrivial intersection with the n-dimensional pseudostable space at $\xi = 0+$, then λ is obviously an eigenvalue.

Next assume that for some λ, the n-dimensional pseudounstable space at $\xi = 0-$ has trivial intersection with the n-dimensional pseudostable space at $\xi = 0+$, so that

$$\mathcal{R}Q(0^+) \oplus \mathcal{R}(I - Q(0^-)) = \mathbb{R}^n. \tag{5.12}$$

Let $w_s \in \mathcal{R}Q(0^+)$ and $w_u \in \mathcal{R}(I - Q(0^-))$. Then the solution of (5.11) can be expressed as

$$
\begin{aligned}
w(\xi) &= T(\xi,0)w_s + \int_0^\xi T(\xi,\zeta)Q(\zeta)(0,\ h(\zeta))^\top d\zeta \\
&\quad + \int_\infty^\xi T(\xi,\zeta)(I-Q(\zeta))(0,\ h(\zeta))^\top d\zeta, \quad \xi > 0, \\
w(\xi) &= T(\xi,0)w_u + \int_0^\xi T(\xi,\zeta)(I-Q(\zeta))(0,\ h(\zeta))^\top d\zeta \\
&\quad + \int_{-\infty}^\xi T(\xi,\zeta)Q(\zeta)(0,\ h(\zeta))^\top d\zeta, \quad \xi < 0.
\end{aligned} \tag{5.13}
$$

Using Lemma 5.3 and the fact that $(0,\ h(\zeta))^\top = (I-\tilde{P})(0,\ h(\zeta))^\top$, it is easy to show that the integrals in (5.13) are convergent and define functions in $C(\epsilon\gamma, \mathbb{R}^+_\xi)$ for $\xi > 0$ and in $C(\epsilon\gamma, \mathbb{R}^-_\xi)$ for $\xi < 0$.

It remains to find $w_s \in \mathcal{R}Q(0^+)$ and $w_u \in \mathcal{R}(I-Q(0^-))$ such that $w(0^-) = w(0^+)$. From (5.13),

$$
w_u - w_s = \int_\infty^0 T(0,\zeta)(I-Q(\zeta))(0,\ h(\zeta))^\top d\zeta - \int_{-\infty}^0 T(0,\zeta)Q(\zeta)(0,\ h(\zeta))^\top d\zeta. \tag{5.14}
$$

By (5.12), there exist unique $w_s \in \mathcal{R}Q(0^+)$ and $w_u \in \mathcal{R}(I - Q(0^-))$ such that (5.14) holds.

Thus the spectral equation (5.4) has a unique solution U for each h. From (5.13), we see that $|U|_{\epsilon\gamma} \le C_\epsilon |h|_{\epsilon\gamma}$. This shows that λ is in the resolvent of the linear partial differential equation (5.4).

We have proved the following.

THEOREM 5.4. *Let $\tilde{\delta}$ be a positive constant. Let $\gamma > \max\{-\alpha_2, \beta_1\}$. Then for $\epsilon > 0$ sufficiently small, system* (5.3) *on the space $C^2(\gamma, \mathbb{R}_x)$ (resp., system* (5.4) *on the space $C^2(\epsilon\gamma, \mathbb{R}_\xi)$) has only normal points in the region* $\mathrm{Re}\tilde{\lambda} \ge -\tilde{\delta}$ *(resp., in the region* $\mathrm{Re}\lambda \ge -\delta := -\epsilon\tilde{\delta}$*).*

We end this section by stating a lemma that will also be used in the next section and a corollary that was used in the proof of Lemma 5.1.

LEMMA 5.5. *Let $\lambda = \sigma + \omega i$ and $z = x + yi$ be complex variables, with $\sigma, \omega, x, y \in \mathbb{R}$. For a given real $r \ne 0$, consider the mapping*

$$
z = \sqrt{r^2 + \lambda}
$$

and its inverse

$$
\lambda = z^2 - r^2.
$$

(1) *For any $a > 0$, the mapping $\lambda = z^2 - r^2$ takes each vertical line* $\mathrm{Re}z = \pm a$ *bijectively onto the parabola*

$$
\sigma = a^2 - r^2 - \frac{\omega^2}{4a^2}.
$$

The regions $\mathrm{Re}z \ge a$ *and* $\mathrm{Re}z \le -a$ *are each mapped bijectively onto the closure of the region to the right of the parabola, i.e., onto*

$$
\sigma \ge a^2 - r^2 - \frac{\omega^2}{4a^2}.
$$

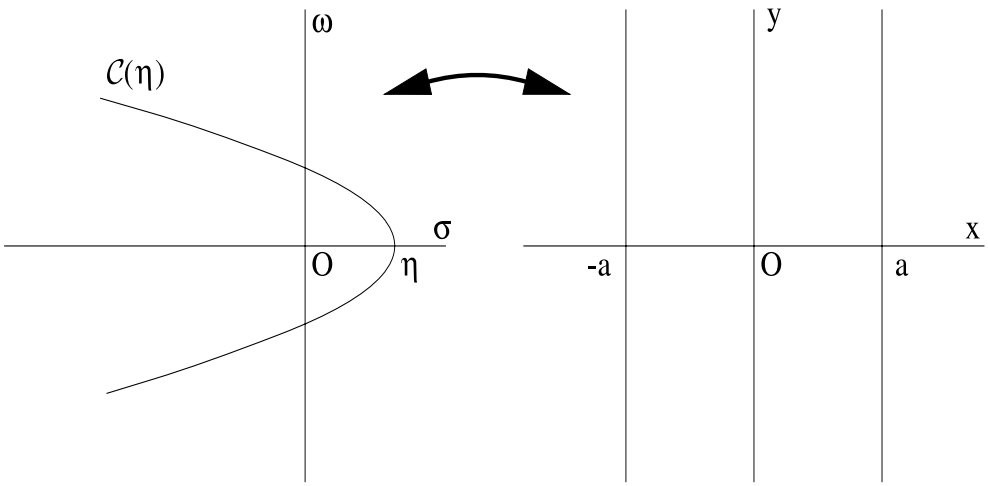


FIG. 5.1. *The mapping* $\lambda = z^2 - r^2$ *takes each vertical line* $\mathrm{Re}z = \pm a$ *bijectively onto the parabola* $\mathcal{C}(\eta)$.

(2) *For any* $\eta > -r^2$, *let*

$$\mathcal{C}(\eta) := \left\{ (\sigma, \omega) : \sigma = \eta - \frac{\omega^2}{4(r^2+\eta)} \right\},$$

a parabola with vertex at $(\eta, 0)$ *that opens to the left. Then the mapping* $z = \sqrt{r^2+\lambda}$ *takes* $\mathcal{C}(\eta)$ *onto the vertical lines* $\mathrm{Re}z = \pm a = \pm\sqrt{r^2+\eta}$. *The closure of the region to the right of* $\mathcal{C}(\eta)$, *denoted* $\mathcal{R}(\eta)$, *is mapped onto* $|z| \geq a = \sqrt{r^2+\eta}$.

(3) *If* $\eta > 0$, *then* $a > |r|$; *if* $-r^2 < \eta < 0$, *then* $0 < a < |r|$.

See Figure 5.1.

COROLLARY 5.6. *For any* $0 < \delta < r^2$, *let* $\eta = -\delta$. *Then the region* $\mathrm{Re}\lambda \geq -\eta$ *is in* $\mathcal{R}(-\delta)$ *and is mapped by* $z = \sqrt{r^2+\lambda}$ *into* $|\mathrm{Re}z| \geq \sqrt{r^2+\eta} = \sqrt{r^2-\delta}$.

6. $O(\frac{1}{\epsilon})$ Eigenvalues. Let us first consider a time-dependent solution $u_\epsilon(x,t)$ of (1.17) with initial data $u_\epsilon(x,0) = \phi_\epsilon(x)$ near the Riemann–Dafermos solution $u_\epsilon(x)$. Thus, $\phi_\epsilon(x)$ has n sharp transition layers at $\bar{x}^i_\epsilon$, with $\bar{x}^i_\epsilon$ near $\bar{s}^i$. Then we expect that $u_\epsilon(x,t)$ has n sharp jumps near curves $\bar{x}^i_\epsilon(t)$, with $\bar{x}^i_\epsilon(0) = \bar{x}^i_\epsilon$. (If the Riemann–Dafermos solution is stable, we expect that $\bar{x}^i_\epsilon(t) \to x^i(\epsilon)$ as $t \to \infty$.) Near the curve $\bar{x}^i_\epsilon(t)$ we use the fast spatial variable $\xi = \frac{x - \bar{x}^i_\epsilon(t)}{\epsilon}$. Then (1.17) becomes

$$\epsilon u_t = u_{\xi\xi} - \left(Df(u) - \bar{x}^i_\epsilon(t) - \frac{d}{dt}\bar{x}^i_\epsilon(t) - \epsilon\xi \right) u_\xi.$$

Unless ϕ_ϵ is a stationary solution of (1.17), we have $u_t = O(\frac{1}{\epsilon})$ near $\bar{x}^i_\epsilon$; i.e., the system exhibits very fast motion near $\bar{x}^i_\epsilon$. It is common in singular perturbation problems to have an *initial layer* in which there is motion with speed of order $\frac{1}{\epsilon}$ for time of order ϵ. Thus we expect the existence of eigenvalues of order $\frac{1}{\epsilon}$, with the support of the eigenfunctions concentrated near the points $\bar{x}^i_\epsilon$.

Assume now that in the singular layers, the solution quickly converges to traveling-wave-like solutions. Then after the initial time layer, the solution behaves like convection in the regular layer coupled with traveling waves in singular layers. This motion

occurs for $t > O(\epsilon)$ and has $u_t = O(1)$. Thus we expect to find eigenvalues of order 1 and related eigenfunctions.

We discuss fast eigenvalues of order $\frac{1}{\epsilon}$ in this section. Slow eigenvalues of order 1 will be studied in the next section.

We recall that the linear variational system at a Riemann–Dafermos solution $u_\epsilon(x)$ is

$$U_t + (Df(u_\epsilon) - xI)U_x + D^2 f(u_\epsilon) u_{\epsilon\, x} U = \epsilon U_{xx}.$$

We shall study this equation in the space $C^2(\gamma, \mathbb{R}_x)$, $\gamma > 0$.

Eigenvalues $\tilde{\lambda}$ and corresponding eigenfunctions $U(x)$ satisfy

$$\tilde{\lambda} U + (Df(u_\epsilon) - xI)U_x + D^2 f(u_\epsilon) u_{\epsilon\, x} U = \epsilon U_{xx}. \tag{6.1}$$

In section 3 we found an expansion for $u_\epsilon(x)$ in the regular layer. We also found expansions for the jump positions $x^i(\epsilon)$, and for $u^i_\epsilon(\xi)$ in singular layers centered around $x^i(\epsilon)$, in the stretched coordinate $\xi = \frac{x - x^i(\epsilon)}{\epsilon}$. We shall use these expansions in what follows.

We shall look for eigenvalues

$$\tilde{\lambda} = \sum_{j=-1}^{\infty} \epsilon^j \lambda_j. \tag{6.2}$$

Fast eigenvalues have $\lambda_{-1} \neq 0$; slow eigenvalues have $\lambda_{-1} = 0$. We shall look for corresponding eigenfunctions with expansions

$$U^R_\epsilon(x) = \sum_{j=0}^{\infty} \epsilon^j U^R_j(x) \qquad \text{in the regular layer,} \tag{6.3}$$

$$U^i_\epsilon(\xi) = \sum_{j=0}^{\infty} \epsilon^j U^i_j(\xi) \qquad \text{in the singular layer } S^i. \tag{6.4}$$

In this section we look for fast eigenvalues, which have the form (6.2) with $\lambda_{-1} \neq 0$.

We shall show that under certain conditions, fast eigenvalues have eigenfunctions that are localized in a single singular layer. These eigenvalues correspond to zeros of Evans functions on each singular layer.

We first consider the regular layer. We substitute (3.4), (6.2), and (6.3) into (6.1) and expand in powers of ϵ. At order ϵ^{-1} (the lowest order) we obtain

$$\lambda_{-1} U^R_0 = 0. \tag{6.5}$$

Since $\lambda_{-1} \neq 0$, $U^R_0 = 0$.

At order ϵ^0 we obtain

$$\lambda_{-1} U^R_1 = \text{terms involving } U^R_0 = 0.$$

Since $\lambda_{-1} \neq 0$, $U^R_1 = 0$. Similarly, higher-order expansions of eigenvalues and the corresponding eigenfunctions are determined by a system of algebraic equations. In particular, we find that $U^R_j = 0$ for all j.

In the ith singular layer, we rewrite (6.1) as

$$\epsilon(\tilde{\lambda} + 1)U + ((Df(u_\epsilon) - xI)U^i)_\xi = U^i_{\xi\xi}, \quad \text{with } x = x_i(\epsilon) + \epsilon\xi. \tag{6.6}$$

We substitute (3.6), (6.2), and (6.4) into (6.6) and expand in powers of ϵ. At order ϵ^0 (the lowest order) we obtain

$$\lambda_{-1}U_0^i + ((Df(q^i) - x_0^i I)U_0^i)_\xi = U_{0\xi\xi}^i. \tag{6.7}$$

Since $U_0^R = 0$, we must have $U_0^i(\xi) \to 0$ as $\xi \to \pm\infty$. We note that (6.7) also arises in the study of the stability of the traveling wave solution $u(X,T) = q^i(X - x_0^i T)$ of the system of viscous conservation laws (1.1); it determines the eigenvalues and eigenfunctions of the linearization of (1.1) at the traveling wave. Let us assume the following:

(H1) For the complex number $\lambda_{-1} \neq 0$, there is exactly one i, $1 \leq i \leq n$, such that (6.7) has a nontrivial solution $U_0^i(\xi)$ that satisfies the boundary conditions $U_0^i(\xi) \to 0$ as $\xi \to \pm\infty$.

(H2) For that i, λ_{-1} is a semisimple eigenvalue [20, p. 41] of the linear differential operator

$$U_{0\xi\xi}^i - ((Df(q^i) - x_0^i I)U_0^i)_\xi \tag{6.8}$$

on the Banach space of uniformly continuous functions that approach 0 as $\xi \to \pm\infty$, with the sup norm.

Consider first the index i of assumption (H1). Let $\lambda_{-1}^i := \lambda_{-1}$. Let the multiplicity of λ_{-1}^i as an eigenvalue of (6.8) be m_i. Let $\phi_j^i(\xi)$, $j = 1, \ldots, m_i$, be a basis for the eigenspace. Then to lowest order, an eigenfunction associated to $\tilde{\lambda} = \sum_{j=-1}^{\infty} \lambda_j^i \epsilon^j$ has the form $U_0^i(\xi) = \sum_{j=1}^{m_i} c_j^i \phi_j^i(\xi)$ in the ith singular layer for some constants $\{c_j^i\}_{j=1}^{m_i}$ and is zero in the regular layer and other singular layers.

We now show how to determine the possible values of λ_0^i and $\{c_j^i\}_{j=1}^{m_i}$ using the expansions to order ϵ^1.

Later, we will show that in certain regions of λ-space, the limiting systems of (6.7) at $\xi = \pm\infty$ have exponential dichotomies with n-dimensional unstable and stable subspaces. The eigenfunction U_0^i corresponds to a nontrivial intersection of the unstable subspace at $\xi = -\infty$ and the stable subspace at $\xi = \infty$.

By [33], the adjoint system to (6.7) must also have an m_i-dimensional space of bounded solutions. Let $\{\psi_\ell^i\}_{\ell=1}^{m_i}$ be a basis for this space.

In the ith singular layer, at order ϵ^1, we have

$$\begin{aligned}(\lambda_0^i + 1)U_0^i + ((D^2 f(q^i)u_1^i - (x_1^i + \xi)I)U_0^i)_\xi \\ + \lambda_{-1}^i U_1^i + ((Df(q^i) - x_0^i I)U_1^i)_\xi = U_{1\xi\xi}^i.\end{aligned} \tag{6.9}$$

The solvability condition of (6.9) can be obtained from Fredholm's alternative [33]:

$$\langle \psi_\ell^i, (\lambda_0^i + 1)U_0^i + ((D^2 f(q^i)u_1^i - (x_1^i + \xi)I)U_0^i)_\xi \rangle = 0, \quad \ell = 1, \ldots, m_i. \tag{6.10}$$

Recall that $U_0^i = \sum_{j=1}^{m_i} c_j^i \phi_j^i(\xi)$. Since λ_{-1}^i is semisimple, without loss of generality, we assume that $\langle \psi_\ell^i, \phi_j^i \rangle = \delta_j^\ell$.

Let $\mathcal{B}^i = \{b_{\ell,j}^i\}$ be the $m_i \times m_i$ matrix whose entries are

$$b_{\ell,j}^i := \langle \psi_\ell^i, ((D^2 f(q^i)u_1^i - (x_1^i + \xi)I)\phi_j^i)_\xi \rangle.$$

The solvability condition (6.10) becomes

$$((\lambda_0^i + 1)I - \mathcal{B}^i)\mathbf{c}^i = \mathbf{0}, \tag{6.11}$$

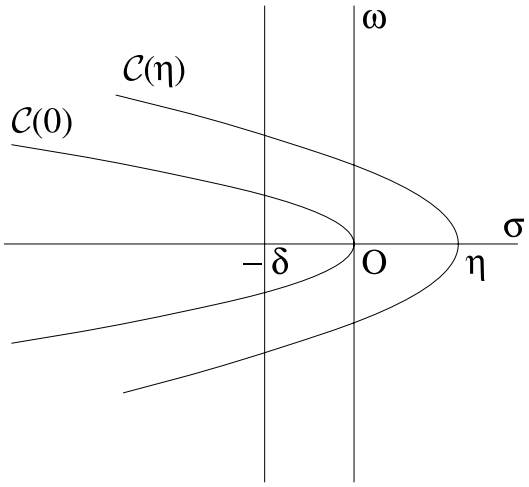


FIG. 6.1. *Eigenvalues to the right of* $\mathcal{C}(\eta)$ *and* $\mathcal{C}(0)$.

where $\mathbf{c}^i = (c^i_1, \ldots, c^i_{m_i})$, and I is the $m_i \times m_i$ identity matrix. Therefore $\lambda^i_0 + 1$ is an eigenvalue of the matrix $\mathcal{B}$ and $(c^i_1, \ldots, c^i_{m_i})$ is the corresponding eigenvector. The algebraic system (6.11) determines the possible values of λ^i_0 and the corresponding $\mathbf{c}^i$.

We assume the following:

(H3) The eigenvalues of the matrix $\mathcal{B}^i$ are distinct.

Of course, (H3) holds automatically in the most common case, $m_i = 1$.

From (H3), we have m_i distinct eigenvalues $\lambda^i_0 + 1$, each with an eigenvector $\mathbf{c}^i$ corresponding to an eigenfunction $U^i_0 = \sum c^i_j \phi^i_j$. Thus, for $\epsilon > 0$, λ^i_1 splits into m_i distinct eigenvalues.

Assuming (H3), higher-order expansions of eigenvalues and the corresponding eigenfunctions in singular layers can be obtained by a straightforward formal procedure, which will not be presented here.

Next, we consider i other than the one specified in assumption (H1). It is clear that $U^i_0 = 0$. From (6.9) we find that $U^i_1 = 0$. Similarly, all $U^i_j = 0$.

We refer to the $O(\frac{1}{\epsilon})$ eigenvalues as *local eigenvalues* since the asymptotic expansions of their associated eigenfunctions are localized in a single singular layer.

Our next object is to define, for the ith singular layer, an Evans function $E^i(\lambda)$ [11] whose zeros are complex numbers λ^i_{-1} for which (6.7) has solutions that approach 0 as $\xi \to \pm\infty$. For an arbitrary $\eta > 0$, we will define a parabola $\mathcal{C}(\eta)$ that opens to the left and has its vertex at $(\eta, 0), \eta > 0$, in the complex plane. The parabolas $\mathcal{C}(\eta)$ do not intersect. As $\eta \to 0+$, they approach a parabola $\mathcal{C}(0)$ with vertex at $(0, 0)$. See Figure 6.1 Let the region to the right of $\mathcal{C}(\eta)$ be $\mathcal{R}(\eta)$. The Evans function $E^i(\lambda)$ can be defined on $\mathcal{R}(0)$. For each small $\eta > 0$, if λ^i_{-1} is a zero of the Evans function defined in $\mathcal{R}(\eta)$, then (6.7) has a nontrivial solution that satisfies $U^i_0(\xi) = O(e^{-\eta|\xi|})$.

As in section 5, let $x^i_0 = \bar{s}^i$, $i = 1, \ldots, n$. Let $N > \max\{|x^1_0|, |x^n_0|\}$. Thus the compact interval $[-N, N]$ contains all the points x^i_0, $i = 1, \ldots, n$. Let $x^0_0 = -N$ and $x^{n+1}_0 = N$. For $\lambda \in \mathbb{C}$ and $i = 0, \ldots, n$, define

$$\tilde{A}^i(\lambda, x) = \begin{pmatrix} 0 & I \\ \lambda I & Df(\bar{u}^i_0) - xI \end{pmatrix}, \quad x \in [x^i_0, x^{i+1}_0],$$

where $\bar{u}^i_0$ was defined in section 5.

LEMMA 6.1. *For each $\eta > 0$, there exist $\beta(\eta) < 0 < \alpha(\eta)$ such that, for all $\lambda \in \mathcal{R}(\eta)$, for all $i = 0, \dots, n$, and for all x in $[x_0^i, x_0^{i+1}]$, $\tilde{A}^i(\lambda, x)$ has n eigenvalues less than $\beta(\eta)$ and n eigenvalues greater than $\alpha(\eta)$. As $\eta \to 0$, α and β are $O(\eta)$.*

Proof. Fix an index i between 0 and n. Let $\nu_1^i < \cdots < \nu_n^i$ denote the eigenvalues of $Df(\bar{u}_0^i)$. Let μ be an eigenvalue of $\tilde{A}^i(\lambda, x)$. Then one of the following equations must hold:

$$\mu^2 + (x - \nu_j^i)\mu - \lambda = 0, \quad j = 1, \dots, n. \tag{6.12}$$

Let $p_j^i(x) := \frac{1}{2}(x - \nu_j^i)$, $x \in [x_0^i, x_0^{i+1}]$. The solutions of (6.12) are

$$\mu_j^{i\pm}(\lambda, x) := -p_j^i \pm \sqrt{{p_j^i}^2 + \lambda}.$$

The main branch of the square root is used.

Define

$$\mathcal{C}_j^i(\eta) := \left\{ (\sigma, \omega) : \sigma = \eta - \frac{\omega^2}{4({p_j^i}^2 + \eta)} \right\},$$

$$\mathcal{R}_j^i(\eta) := \left\{ (\sigma, \omega) : \sigma \geq \eta - \frac{\omega^2}{4({p_j^i}^2 + \eta)} \right\},$$

$$\alpha_j^i := -p_j^i + \sqrt{{p_j^i}^2 + \eta},$$

$$\beta_j^i := -p_j^i - \sqrt{{p_j^i}^2 + \eta}.$$

The vertex of the parabola $\mathcal{C}_j^i(\eta)$ is at $(\sigma, \omega) = (\eta, 0)$. The parabola opens to the left.

Using Lemma 5.5 with $p = p_j^i$, we have that if $\lambda \in \mathcal{R}_j^i(\eta)$, then

$$\mathrm{Re}\mu_j^{i-} \leq \beta_j^i < 0 < \alpha_j^i \leq \mathrm{Re}\mu_j^{i+}, \quad j = 1, \dots, n.$$

Define

$$p := \max |p_j^i(x)|, \quad \alpha := \min \alpha_j^i, \quad \beta := \max \beta_j^i,$$

$$\mathcal{C}(\eta) := \{(\sigma, \omega) | \sigma = \eta - \frac{\omega^2}{4(p^2 + \eta)}, \tag{6.13}$$

$$\mathcal{R}(\eta) := \cap_{i,j} \mathcal{R}_j^i(\eta) = \{(\sigma, \omega) | \sigma \geq \eta - \frac{\omega^2}{4(p^2 + \eta)}. \tag{6.14}$$

If $\lambda \in \mathcal{R}(\eta)$, then $\mu_j^{i-} < \beta < 0 < \alpha < \mu_j^{i+}$ for all i and j and for all $x \in [x_0^i, x_0^{i+1}]$. From their definitions, $\alpha_j^i = O(\eta)$ if $p_j^i > 0$ and $\beta_j^i = O(\eta)$ if $p_j^i < 0$. Notice that p_j^i can be both positive and negative. It follows that α and β are $O(\eta)$. ∎

Let $x = \epsilon\xi$, and let $A^i(\lambda, \epsilon, \xi) := \tilde{A}^i(\lambda, \epsilon\xi)$. From the roughness theory of exponential dichotomies [7] and Lemma 6.1, we derive the following proposition.

PROPOSITION 6.2. *For each $i = 0, \dots, n$ and for each $\lambda \in \mathcal{R}(\eta)$, the slowly varying system*

$$W_\xi = A^i(\lambda, \epsilon, \xi)W, \quad \xi \in \left[-\frac{x_0^i}{\epsilon}, \frac{x_0^{i+1}}{\epsilon} \right],$$

has an exponential dichotomy with exponents $\beta(\eta) < 0 < \alpha(\eta)$. The unstable subspace of the exponential dichotomy in each subinterval is n-dimensional. As $\eta \to 0$, α and β are $O(\eta)$.

Using the information from Lemma 6.1, for each internal layer S^i and for each $\eta > 0$, we can define an Evans function $E^i(\lambda)$ for $\lambda \in \mathcal{R}(\eta)$. More precisely, rewrite (6.7) as

(6.15)
$$\begin{pmatrix} U_\xi \\ V_\xi \end{pmatrix} = B(\lambda,\xi)\begin{pmatrix} U \\ V \end{pmatrix}, \text{ where } B(\lambda,\xi) := \begin{pmatrix} 0 & I \\ \lambda I + D^2 f(q^i(\xi))q^i_\xi & Df(q^i(\xi)) - x^i_0 I \end{pmatrix}.$$

The coefficient matrix approaches $\tilde{A}(\lambda, x^i_0\pm)$ as $\xi \to \pm\infty$ exponentially. By Lemma 6.1, the limiting matrices $\tilde{A}(\lambda, x^i_0\pm)$ have n eigenvalues with real parts less than $\beta(\eta) < 0$ and the other n eigenvalues with real parts greater than $\alpha(\eta) > 0$. We conclude that for the system (6.15), there exist n linearly independent solutions $\{\phi^+_j(\lambda,\xi)\}^n_{j=1}$ such that each approaches zero as $\xi \to \infty$ and n linearly independent solutions $\{\phi^-_j(\lambda,\xi)\}^n_{j=1}$ such that each approaches zero as $\xi \to -\infty$.

The Evans function for the internal layer S^i is defined as

$$E^i(\lambda) := e^{-\int_0^\xi \operatorname{tr} B(\lambda,\zeta)d\zeta} a(\lambda,\xi) \wedge b(\lambda,\xi) = a(\lambda,0) \wedge b(\lambda,0). \tag{6.16}$$

Here $a(\lambda,\xi)$ and $b(\lambda,\xi)$ are n-forms associated to $\{\phi^-_j : i = 1,\dots,n\}$ and $\{\phi^+_j : i = 1,\dots,n\}$, respectively [11], [1], [14].

Since formula (6.16) is independent of η, the Evans function is actually defined on $\mathcal{R}(0)$. A zero of the Evans function corresponds to a complex number λ^i_{-1} for which (6.7) admits a nontrivial solution U^i_0 that approaches zero as $\xi \to \pm\infty$. The same Evans function arises in the study of the stability of the traveling wave solution $u(X,T) = q^i(X - x^i_0 T)$ of the system of viscous conservation laws (1.1).

According to [14], the Evans function extends analytically to a neighborhood of the origin. We always have $E^i(0) = 0$; an eigenfunction is q^i_ξ. By analyticity, there are no other zeros of $E^i(\lambda)$ near $\lambda = 0$. Therefore for any sufficiently small $\eta > 0$ and $\delta > 0$, all zeros of $E^i(\lambda)$ in $\{\lambda : \mathrm{Re}\lambda \geq -\delta\}$ are contained in $\mathcal{R}(\eta) \cap \{\lambda : \mathrm{Re}\lambda \geq -\delta\}$.

THEOREM 6.3. *Let $\eta > 0$ be given and let λ^i_{-1} be a zero of E^i in the region $\mathcal{R}(\eta) \cap \{\lambda : \mathrm{Re}\lambda \geq -\delta\}$. Assume that conditions* (H1)–(H3) *are satisfied. Then there exists $\epsilon_0(\eta) > 0$ such that if $0 < \epsilon < \epsilon_0(\eta)$, then the root λ^i_{-1} of E^i is associated to a finite number of curves of fast eigenvalues* (6.2).

To all orders in ϵ, the corresponding eigenfunction is zero in the regular layer and in singular layers other than the ith. The pair (λ^i_{-1}, U^i_0) satisfies (6.7) *and the boundary condition $U^i_0 \to 0$ as $\xi \to \pm\infty$. If the eigenspace of λ^i_{-1} for* (6.7) *is m_i-dimensional, then $U^i_0 = \sum_{j=1}^{m_i} c^i_j \phi^i_j$, where $\{\phi^i_j\}^{m_i}_{j=1}$ is a basis for the eigenspace. The m_i possible values of λ^i_0 and the corresponding vectors $\mathbf{c}^i$ are determined by the eigenvalue-eigenvector problem* (6.11).

Proof. Sketch of the proof: The procedure for finding the correction terms $\Delta\lambda$ and ΔU^i is similar to that for finding λ^i_0 and $\mathbf{c}^i$, followed by a contraction mapping argument. The necessary dichotomies in regular sublayers and singular layers come from Lemma 6.1 and Proposition 6.2. ∎

Remark 6.1. (1) We emphasize that Theorem 6.3 does not apply to $\lambda_{-1} = 0$. Indeed, by Proposition 6.2, as η decreases, the exponential dichotomy weakens, so the ϵ-interval on which the contraction mapping argument is valid shrinks. Thus,

as $\eta \to 0$, $\epsilon_0(\eta) \to 0$. Moreover, as we shall see in the next section, there can be an infinite number of curves of eigenvalues (6.2) whose asymptotic expansion begins with $\lambda_{-1} = 0$; in the case $n = 2$, at least, this is typical.

(2) We also emphasize that we have not shown that for a fixed small $\epsilon > 0$, all eigenvalues near $\lambda_{-1} = 0$ are given by expansions of the form (6.2) with $\lambda_{-1} = 0$. We note, however, that $E'(0)$ is the product of two terms, one of which is nonzero if and only if Majda's inviscid stability condition holds [14], [3]. We expect that in the case $E'(0) \neq 0$, all eigenvalues near $\lambda_{-1} = 0$ are in fact given by such expansions.

7. $O(1)$ Eigenvalues. We look for eigenvalues of (6.1) of the form

$$\tilde{\lambda} = \sum_{j=0}^{\infty} \epsilon^j \lambda_j \tag{7.1}$$

and the corresponding eigenfunctions $U(x)$. We continue to work in the space $C^2(\gamma, \mathbb{R}_x)$, $\gamma > 0$. We rewrite (6.1) as

$$(\tilde{\lambda} + 1)U^R + ((Df(u_\epsilon) - xI)U^R)_x = \epsilon U^R_{xx} \qquad \text{in the regular layer,} \tag{7.2}$$

$$\epsilon(\tilde{\lambda} + 1)U^i + ((Df(u_\epsilon) - x^i(\epsilon) - \epsilon\xi I)U^i)_\xi = U^i_{\xi\xi} \qquad \text{in the singular layer } S^i. \tag{7.3}$$

PROPOSITION 7.1. *To any order of ϵ, $\tilde{\lambda} = -1$ is an eigenvalue of* (7.2) *and* (7.3). *The corresponding eigenfunctions form an n-dimensional eigenspace. The ith basis vector is a homoclinic solution to 0 that, to lowest order in ϵ, equals q^i_ξ in the ith singular layer and is zero in other singular layers and in the regular layer.*

Proof. We need to find expansions of $U^R_\epsilon(x)$ and $U^i_\epsilon(\xi)$ to the following system:

$$((Df(u_\epsilon) - xI)U^R)_x = \epsilon U^R_{xx} \qquad \text{in the regular layer,} \tag{7.4}$$

$$((Df(u_\epsilon) - x^i(\epsilon) - \epsilon\xi I)U^i)_\xi = U^i_{\xi\xi} \qquad \text{in the singular layer } S^i. \tag{7.5}$$

By Lemma 7.2, proved below, for any $j \geq 0$, $U^R_j(x) = 0$ in the regular sublayer R^0.

Let $1 \leq i \leq n$. Assume that for all $j \geq 0$, $U^R_j(x) = 0$ in the regular sublayer R^{i-1}. We shall show that $U^i_0(\xi)$ is a constant multiple of q^i_ξ and that for every $j \geq 0$, $U^R_j(x) = 0$ in the regular sublayer R^i. Then, by induction on i, the proposition is proved.

In the singular layer S^i, in order to match the solution in R^{i-1}, we look for a bounded solution of (7.5) that approaches 0 as $\xi \to -\infty$. Integrating (7.5) from $-\infty$ to ξ, we have

$$U_\xi - (Df(u_\epsilon) - x^i(\epsilon) - \epsilon\xi I)U = 0. \tag{7.6}$$

By the definition of a Lax i-shock, at order ϵ^0, this system has exponential dichotomies for $\xi \in \mathbb{R}^\pm$. By the definition of a structurally stable Riemann solution, the unstable space of the dichotomy on $\mathbb{R}^-$ intersects the stable space of the dichotomy on $\mathbb{R}^+$ transversely at $\xi = 0$. The intersection is a one-dimensional space spanned by q^i_ξ. To have a bounded solution, we must set $U^i_0(\xi)$ equal to a constant multiple of q^i_ξ. Then $U^i_0(\xi)$ approaches zero exponentially as $\xi \to \pm\infty$.

At order ϵ^1, (7.6) becomes

$$U^i_{1\xi} - (Df(u^i_0(\xi)) - x^i_0)U^i_1 = (D^2 f(u^i_0(\xi))u^i_1 - (x^i_1 + \xi)I)U^i_0. \tag{7.7}$$

Since $U_0^i(\xi) = O(e^{-\alpha|\xi|})$, the nonhomogeneous term of (7.7) is $O((|\xi|+1)e^{-\alpha|\xi|})$, which approaches zero as $\xi \to \pm\infty$. Observe that the homogeneous part of (7.7) has exponential dichotomies in $\mathbb{R}^{\pm}$, and the unstable space of the dichotomy on $\mathbb{R}^-$ intersects the stable space of the dichotomy on $\mathbb{R}^+$ transversely at $\xi = 0$. Thus, if we assume that $U_1^i(0) \perp U_0^i(0)$, a unique solution $U_1^i = O((|\xi|+1)e^{-\alpha|\xi|})$ can be constructed using integral equations on $\mathbb{R}^{\pm}$ and the matching at $\xi = 0$. See [33], [25].

Proceeding inductively, at order ϵ^j, $j > 1$, we solve a nonhomogeneous system for U_j^i, with a nonhomogeneous term that is $O((1+|\xi|)^j e^{-\alpha|\xi|})$. The solution $U_j^i = O((1+|\xi|)^j e^{-\alpha|\xi|})$ approaches zero as $\xi \to \pm\infty$.

We now consider the solution in the regular sublayer R^i. By matching, for all $j \geq 0$, $U_j^R(x_0^i+) = U_j^i(\infty) = 0$.

We show inductively that for all $j \geq 0$, $U_j^R(x) = 0$ in R^i. At order ϵ^0, from (7.4), $(Df(u_0)-xI)U_0^R(x)$ is constant in R^i. Since it is zero at x_0^i+, $(Df(u_0)-xI)U_0^R(x) = 0$ in R^i. Since $Df(u_0) - xI$ is nonsingular in each regular sublayer, we see that $U_0^R = 0$ in R^i.

Next, at order ϵ^1, because $U_0^R = 0$, we can show similarly $(Df(u_0) - xI)U_1^R(x)$ is constant in R^i, and hence that $U_1^R = 0$ in this sublayer. Proceeding inductively, we see that for all $j \geq 0$, $U_j^R = 0$ in R^i. ∎

We remark that in the viscous regularization (1.1) of a system of conservation laws (1.3), traveling wave solutions always have a zero eigenvalue with eigenfunctions $U_0^i = c_0^i q_\xi^i$. Such an eigenfunction corresponds to a shift of the shock position from X_0^i to $X_0^i + c_0^i$. Using the self-similar variable $x = X/T$, the shock position is at $(X_0^i + c_0^i)/T$, which differs from X_0^i/T by a decay term c_0^i/T. Changing to the new time $t = \ln T$, the deviation of the shock position is $c_0^i e^{-t}$. This explains why (6.1) always has an eigenvalue (7.1) with $\lambda_0 = -1$, and why the eigenspace is as stated in Proposition 7.1.

To look for other slow eigenvalues, in the regular layer we substitute (3.4), (7.1), and (6.3) into (6.1) and expand in powers of ϵ. In singular layers, we substitute (3.6), (3.5), (7.1), and (6.4) into (6.6) and expand in powers of ϵ. For a fixed $\gamma > 0$, we shall look for solutions such that

$$|U(x)| \leq Ce^{-\gamma|x|} \text{ in the sublayers } R^0 = (-\infty, x_0^1) \text{ and } R^n = (x_0^n, \infty) \tag{7.8}$$

for some constant C.

At order ϵ^0 (the lowest order) we obtain

$$(\lambda_0 + 1)U_0^R + ((Df(\bar{u}_0^i) - xI)U_0^R)_x = 0 \qquad \text{in the sublayer } R^i, \tag{7.9}$$

$$((Df(q^i) - x_0^i I)U_0^i)_\xi = U_{0\xi\xi}^i \qquad \text{in the singular layer } S^i. \tag{7.10}$$

LEMMA 7.2. *To all orders of ϵ, eigenfunctions $U(x)$ that satisfy* (7.8) *are zero in the regular sublayers $R^0 = (-\infty, x_0^1)$ and $R^n = (x_0^n, \infty)$.*

Proof. First consider the lowest order ϵ^0. U_0^R satisfies (7.9). We consider only R^0. Let ν_j, $j = 1, \ldots, n$, be the eigenvalues of $Df(\bar{u}_0^0)$. Notice that for each $j = 1, \ldots, n$, $\nu_j - x > 0$ in R^0. Let $\mathbf{l}_j$, $j = 1, \ldots, n$, be corresponding left eigenvectors. Let $v_j(x) = \langle \mathbf{l}_j, U_0(x) \rangle$, $x \in R^0$. Equation (7.9) becomes

$$\lambda_0 v_j + (\nu_j - x)v_{jx} = 0, \quad j = 1, \ldots, n.$$

The general solution is $v_j = C_j(\nu_j - x)^{\lambda_0}$. Since $v_j = O(e^{-\gamma|x|})$, $\gamma > 0$, we must have $C_j = 0$ for all j. Therefore $U_0^R(x) = 0$ for all $x \in R^0$.

By an easy induction argument, we can show that $U_j^R = 0$ for all j on $R^0 \cup R^n$. □

In the ith singular layer, we look for a solution U_0^i of (7.10) connecting the adjacent sublayers. Integration from $\xi = -\infty$ to $\xi = \infty$, together with matching, yields jump conditions that must be satisfied by U_0^R:

$$(Df(\bar{u}_0^i) - x_0^i I)U_0^R(x_0^i+) - (Df(\bar{u}_0^{i-1}) - x_0^i I)U_0^R(x_0^i-) = 0, \quad i = 1, \dots, n. \tag{7.11}$$

By Lemma 7.2, $U_0^R(x_0^1-) = 0$. Then setting $i = 1$ in (7.11) yields

$$U_0^R(x_0^1+) = 0. \tag{7.12}$$

Solving the ODE (7.9) on the sublayer R^1 with the initial condition (7.12) yields $U_0^R(x) = 0$ for all $x \in R^1$. By induction, we have the following.

PROPOSITION 7.3. *Any solution of* (7.9)–(7.10) *that satisfies* (7.8) *has* $U_0^R(x) = 0$ *for all* x *in the regular layer.*

Proposition 7.3 implies that $U_0^i(\xi)$ approaches 0 as $\xi \to \pm\infty$ for all $i = 1, \dots, n$. Then assumption (S2′) implies the following proposition.

PROPOSITION 7.4. *Any solution of* (7.9)–(7.10) *that satisfies* (7.8) *has, for* $i = 1, \dots, n$, $U_0^i(\xi) = c_0^i q_\xi^i(\xi)$, $i = 1, \dots, n$, *for some constants* c_0^i.

The possible values of λ_0, along with the corresponding values of c_0^i, are determined at the ϵ^1-order expansion.

At order ϵ^1, we have

$$\lambda_0 U_1 + (Df(\bar{u}_0^i) - xI)U_{1x} = 0 \quad \text{in the regular layer,} \tag{7.13}$$

$$U_1^R(x) = 0 \quad \text{for } x \in R^0 \cup R^n, \tag{7.14}$$

$$\begin{aligned}(\lambda_0 + 1)U_0^i + ((D^2 f(q^i)u_1^i - (x_1^i + \xi)I)U_0^i)_\xi \\ + ((Df(q^i) - x_0^i I)U_1^i)_\xi = U_{1\xi\xi}^i \quad \text{in the singular layer } S^i.\end{aligned} \tag{7.15}$$

In (7.15), $U_0^i(\xi) = c_0^i q_\xi^i(\xi)$, $i = 1, \dots, n$, for some constants c_0^i by Proposition 7.4.

In order to match with $U_1^R(x)$ in the regular layer, $U_1^i(\xi)$ must satisfy the following boundary conditions: $U_1^i(\xi) \to U_1^R(x_0^i-)$ exponentially as $\xi \to -\infty$ and $U_1^i(\xi) \to U_1^R(x_0^i+)$ exponentially as $\xi \to \infty$. Then, integrating (7.15) from $\xi = -\infty$ to $\xi = \infty$ and using $U_0^i = c_0^i q_\xi^i$, we have the jump condition

$$\begin{aligned}(\lambda_0 + 1)c_0^i(\bar{u}_0^i - \bar{u}_0^{i-1}) + (Df(\bar{u}_0^i) - x_0^i I)U_1^R(x_0^i+) \\ - (Df(\bar{u}_0^{i-1}) - x_0^i I)U_1^R(x_0^i-) = 0, \quad i = 1, \dots, n.\end{aligned} \tag{7.16}$$

By Lemma 3.1, condition (7.16) is sufficient for the existence of a solution $U_1^i(\xi)$ of (7.15) that approaches the desired limits exponentially as $\xi \to \pm\infty$. Thus if

$$(\lambda_0, c_0^1, \dots, c_0^n, U_1^R(x)) \tag{7.17}$$

satisfies (7.13) with auxiliary conditions (7.14) and (7.16), then there exist $U_1^i(\xi)$, $1 \le i \le n$, that satisfy (7.15). More precisely, if we write

$$U_1^i(\xi) = U_1^{i\perp}(\xi) + c_1^i q_\xi^i(\xi),$$

where $U_1^{i\perp}(0)$ is orthogonal to $q_\xi^i(0)$, then (7.15) uniquely determines $U_1^{i\perp}(\xi)$, but the values of c_1^i are determined at the ϵ^2-order expansion. In general, for each $j \ge 1$, we

write $U_j^i(\xi) = U_j^{i\perp}(\xi) + c_j^i q_\xi^i(\xi)$ with $U_j^{i\perp}(0)$ orthogonal to $q_\xi^i(0)$. Then the ϵ^j-order expansion determines

$$(\lambda_{j-1}, c_{j-1}^1, \dots, c_{j-1}^n, U_j^{1\perp}(\xi), \dots, U_j^{n\perp}(\xi)),$$

leaving $(\lambda_j, c_j^1, \dots, c_j^n)$ to be determined at the ϵ^{j+1}-order expansion. In order to continue the expansion past the determination of (7.17), it is necessary to assume that $\lambda_0 + 1$ is a semisimple eigenvalue of a certain operator. This will be described in a later paper. See [25], [16] for related work on reaction-diffusion systems.

PROPOSITION 7.5. *For $\lambda_0 = 0$ there is no nontrivial solution of* (7.13) *with auxiliary conditions* (7.14) *and* (7.16).

Proof. If $\lambda_0 = 0$, then from (7.13), U_1^R is constant in each regular sublayer R^i, $i = 1, \dots, n-1$. Then (7.14) and assumption (S1) imply that the only solution of the system (7.16) is $U_1^R(x) \equiv 0$ for $x \in R^i$, $i = 1, \dots, n-1$, and $c_0^i = 0$ for all i. □

Let $V^i(x) = (Df(\bar{u}_0^i) - xI)U_1^R(x)$, $x \in R^i$ for $i = 0, \dots, n$. Let $s^i := (\lambda_0 + 1)c_0^i$ and $\Delta^i = \bar{u}_0^i - \bar{u}_0^{i-1}$ for $i = 1, \dots, n$. Each Δ^i is nonzero. Equations (7.13), (7.16), and (7.14) become

$$V_x^i + (\lambda_0 + 1)(Df(\bar{u}_0^i) - xI)^{-1}V^i = 0, \quad i = 1, \dots, n-1, \tag{7.18}$$

$$V^i(x_0^i) - V^{i-1}(x_0^i) = -s^i\Delta^i, \quad i = 1, \dots, n, \tag{7.19}$$

$$V^0(x) \equiv 0 \quad \text{and} \quad V^n(x) \equiv 0. \tag{7.20}$$

PROPOSITION 7.6. *For $\lambda_0 \neq -1$, there is a nontrivial solution* (7.17) *of* (7.13), (7.14), (7.16) *if and only if there is a nontrivial solution*

$$(s^1, \dots, s^n, V^1, \dots, V^{n-1})$$

of the system (7.18)–(7.20).

In contrast to the $O(\frac{1}{\epsilon})$ eigenvalues, which reflect the dynamics in a single internal layer, the $O(1)$ eigenvalues reflect the dynamics of the first-order linear ODE (7.18) in the regular layer. Equations (7.19) and (7.20) provide boundary and interface conditions for (7.18).

We remark that the system (7.18)–(7.20) is similar to the SLEP system introduced by Nishiura and Fujii [35] to study the stability of internal layer solutions of reaction-diffusion systems. We now derive the analogue of the SLEP matrix of Nishiura and Fujii.

Let $X(x, y, \lambda_0)$ be the principal matrix solution of (7.18). Although the differential equation (7.18) has jumps at x_0^i, $i = 1, \dots, n$, the principal matrix solution $X(x, y, \lambda_0)$ does not have jumps. If, for example, $y < x_0^j < x_0^{j+1} < \cdots < x_0^i < x$, then

$$X(x, y, \lambda_0) = X(x, x_0^i, \lambda_0) \cdot X(x_0^i, x_0^{i-1}, \lambda_0) \cdot \ \cdots \ \cdot X(x_0^j, y, \lambda_0).$$

If we integrate (7.18) from x_0^1- to x_0^n+ and use the jump conditions (7.19) and the initial and terminal conditions (7.20), we obtain

$$\sum_{j=1}^{n} X(x_0^n, x_0^j, \lambda_0)s^j\Delta^j = 0. \tag{7.21}$$

Let $\mathcal{M}(\lambda_0)$ be the $n \times n$ matrix whose jth column is the n-vector $X(x_0^n, x_0^j, \lambda_0)\Delta^j$, and let $\mathbf{s} = (s^1, \dots, s^n)$. The matrix $\mathcal{M}(\lambda_0)$ is the analogue of the SLEP matrix. Finding

the lowest order expansion of slow eigenvalues is equivalent to finding solutions of

$$\mathcal{M}(\lambda_0)\mathbf{s} = 0. \tag{7.22}$$

Note that Proposition 7.5 implies that $\mathcal{M}(0)$ is nonsingular.

We shall consider the existence of slow eigenvalues λ_0 other than -1 and 0 in more detail only for the case $n = 2$. In this case system (7.18)–(7.20) becomes

$$V_x + (\lambda_0 + 1)(Df(\bar{u}_0^1) - xI)^{-1}V = 0, \quad x_0^1 \le x \le x_0^2, \tag{7.23}$$

$$V(x_0^1) = -s^1\Delta^1, \tag{7.24}$$

$$V(x_0^2) = -s^2\Delta^2. \tag{7.25}$$

Since (7.23) is linear and Δ^1 and Δ^2 are nonzero, the system (7.23)–(7.25) has a nontrivial solution if and only if the following boundary value problem has a solution:

$$V_x + (\lambda_0 + 1)(Df(\bar{u}_0^1) - xI)^{-1}V = 0, \quad x_0^1 \le x \le x_0^2, \tag{7.26}$$

$$V(x_0^1) = \Delta^1, \tag{7.27}$$

$$V(x_0^2) = \text{ a nonzero multiple of } \Delta^2. \tag{7.28}$$

Let the eigenvalues of $Df(\bar{u}_0^1)$ be $\nu_1 < \nu_2$, with corresponding eigenvectors $\mathbf{r}_1$ and $\mathbf{r}_2$. Let

$$V(x) = \sum_{j=1}^{2} a_j(x)\mathbf{r}_j,$$

where $a_j(x)$ is a scalar function. The function $a_j(x)$ satisfies

$$a_j'(x) + \frac{\lambda_0 + 1}{\nu_j - x}a_j(x) = 0. \tag{7.29}$$

Therefore the subspaces of $\mathbb{R}^2$ spanned by $\mathbf{r}_1$ and $\mathbf{r}_2$ are invariant under (7.26).

PROPOSITION 7.7. *For $n = 2$, if Δ^1 or Δ^2 is a multiple of one of the $\mathbf{r}_j$, then there is no λ_0 such that the system* (7.26)–(7.28) *has a solution.*

Proof. Without loss of generality, suppose Δ^1 is a multiple of one of the $\mathbf{r}_j$. Then Δ^2 cannot be a multiple of the same $\mathbf{r}_j$, since it is easy to check that in the case $n = 2$, the Riemann solution $u_0(x)$ satisfies condition (S1) for structural stability if and only if Δ^1 and Δ^2 are linearly independent. Therefore, since the subspaces of $\mathbb{R}^2$ spanned by $\mathbf{r}_1$ and $\mathbf{r}_2$ are invariant under (7.26), the system (7.26)–(7.28) cannot have a solution. ∎

The case in which neither Δ^1 nor Δ^2 is a multiple of one of the $\mathbf{r}_j$ is covered by the following result.

PROPOSITION 7.8. *For $n = 2$, let*

$$\Delta^i = \sum_{j=1}^{2} d_j^i\mathbf{r}_j, \quad i = 1, 2,$$

with all d_j^i nonzero. Then there is a countably infinite set of λ_0 for which (7.26)–(7.28) *has a solution. All such λ_0 have the same real part and have nontrivial U_1^R (hence they are nonlocal). Explicit formulas for λ_0 are given in* (7.36).

Proof. The solution of the initial value problem (7.23), (7.24) is

$$a_j(x) = \left(\frac{x-\nu_j}{x_0^1-\nu_j}\right)^{\lambda_0+1} d_j^1, \quad j=1,2. \tag{7.30}$$

Notice that $x-\nu_j$ and $x_0^1-\nu_j$ have the same sign in the interval $x_0^1 \le x \le x_0^2$, so the number being raised to a power is positive. The function t^{λ_0+1} used in (7.30) is in general multivalued. Since we must have $a_j(x_0^1) = d_j^1$, $j=1,2$, the branch used must be the one for which $1^{\lambda_0+1}=1$.

The boundary condition (7.28) implies that

$$\det \begin{pmatrix} a_1(x) & d_1^2 \\ a_2(x) & d_2^2 \end{pmatrix} = 0 \quad \text{when } x = x_0^2, \tag{7.31}$$

which reduces to

$$\left(\frac{(x-\nu_1)(x_0^1-\nu_2)}{(x_0^1-\nu_1)(x-\nu_2)}\right)^{\lambda_0+1} = \frac{d_2^1 d_1^2}{d_1^1 d_2^2} \quad \text{when } x = x_0^2. \tag{7.32}$$

Again, the branch of t^{λ_0+1} used in (7.32) is the one for which $1^{\lambda_0+1}=1$. In fact, let us define a change of variables by

$$t = \frac{(x-\nu_1)(x_0^1-\nu_2)}{(x_0^1-\nu_1)(x-\nu_2)}, \quad x_0^1 \le x \le x_0^2. \tag{7.33}$$

Then t is an increasing function of x on the interval $x_0^1 \le x \le x_0^2$, and $t(x_0^1)=1$. Let

$$b = t(x_0^2) = \frac{(x_0^2-\nu_1)(x_0^1-\nu_2)}{(x_0^1-\nu_1)(x_0^2-\nu_2)} > 1, \quad d = \frac{d_2^1 d_1^2}{d_1^1 d_2^2} \ne 0. \tag{7.34}$$

Then (7.32) reduces to $b^{\lambda_0+1} = d$ or

$$(\lambda_0+1)\log b = \log d. \tag{7.35}$$

Let the main branch of logarithm for which $\log 1 = 0$ be denoted $\ln x$. We must use the main branch $\log b = \ln b$ in order to have $1^{\lambda_0+1}=1$ for all complex λ_0 . However, in calculating $\log d$, we may use any branch of the natural logarithm.

Since $b>1$ is real and d is real and nonzero, there are two cases.

1. $d>0$. Then $\log d = \ln d + 2n\pi i, \quad n \in \mathbb{Z}$.
2. $d<0$. Then $\log d = \ln|c| + (2n+1)\pi i, \quad n \in \mathbb{Z}$.

Substituting $\log d$ into (7.35), we find

$$\begin{aligned} &\mathrm{Re}\lambda_0 = -1 + \frac{\ln|d|}{\ln b} \quad \text{for } d \ne 0, \\ &\mathrm{Im}\lambda_0 = \begin{cases} \frac{2n\pi}{\ln b} & \text{if } d>0, \\ \frac{(2n+1)\pi}{\ln b} & \text{if } d<0. \end{cases} \end{aligned} \tag{7.36}$$

□

Remark 7.1. With $n=2$, consider a Riemann solution that consists of two weak Lax shocks connecting the states $\bar{u}_0^1$, $\bar{u}_0^2$, and $\bar{u}_0^3$. For the corresponding Riemann–Dafermos solution, Proposition 7.8 implies that the nonlocal slow eigenvalues are stable. In fact, for $i=1,2$, $\bar{u}_0^i - \bar{u}_0^{i-1}$ is approximately parallel to $\mathbf{r}_i$. Therefore $|d_2^1| << |d_1^1|$ and $|d_1^2| << |d_2^2|$, so $|d| << 1$. Hence $\mathrm{Re}\lambda_0 < -1$.

8. Slow eigenvalues and inviscid stability conditions. Let us consider the inviscid system (1.3) and its Riemann solution (2.4). In studying the linearized stability of (2.4) as a solution of (1.3), one considers the following system [22]:

$$U_T + \begin{cases} Df(\bar{u}^0)U_X = 0 & \text{for } X < \bar{s}^1 T, \\ Df(\bar{u}^i)U_X = 0 & \text{for } \bar{s}^i T < X < \bar{s}^{i+1} T, \quad i = 1, \dots, n-1, \\ Df(\bar{u}^n)U_X = 0 & \text{for } \bar{s}^n T < X, \end{cases} \tag{8.1}$$

$$\begin{aligned} (Df(\bar{u}^i) - \bar{s}^i I)U(\bar{s}^i T+, T) - (Df(\bar{u}^{i-1}) - \bar{s}^i I)U(\bar{s}^i T-, T) \\ - S^i(T)(\bar{u}^i - \bar{u}^{i-1}) = 0, \quad i = 1, \dots, n, \end{aligned} \tag{8.2}$$

where

$$U(\bar{s}^i T+, T) = \lim_{X \to \bar{s}^i T+} U(X, T), \quad U(\bar{s}^i T-, T) = \lim_{X \to \bar{s}^i T-} U(X, T). \tag{8.3}$$

In each sector, the matrix $Df(\bar{u}^i)$ is constant, so solutions (which may include discontinuities) propagate along straight-line characteristics. Along the lines $X = \bar{s}^i T$, data arrive from both sides along incoming characteristics, and one uses (8.2) to solve for S^i and for the continuation of the solution along outgoing characteristics. Majda's stability condition—which is that for each $i = 1, \dots, n$, the eigenvectors for the largest $i-1$ eigenvalues at $\bar{u}^{i-1}$, the eigenvectors for the smallest $n-i$ eigenvalues at $\bar{u}^i$, and the vector $\bar{u}^i - \bar{u}^{i-1}$ should constitute a basis for $\mathbb{R}^n$—is just the condition upon which one can do this.

In (8.1) and (8.2), let us make the change of variables $x = \frac{X}{T}$, $t = \ln T$. We obtain

$$U_t + \begin{cases} (Df(\bar{u}^0) - xI)U_x = 0 & \text{for } x < \bar{s}^1, \\ (Df(\bar{u}^i) - xI)U_x = 0 & \text{for } \bar{s}^i < x < \bar{s}^{i+1}, \quad i = 1, \dots, n-1, \\ (Df(\bar{u}^n) - xI)U_x = 0 & \text{for } \bar{s}^n < x, \end{cases} \tag{8.4}$$

$$\begin{aligned} (Df(\bar{u}^i) - \bar{s}^i I)U(\bar{s}^i+, t) - (Df(\bar{u}^{i-1}) - \bar{s}^i I)U(\bar{s}^i-, t) \\ - S^i(t)(\bar{u}^i - \bar{u}^{i-1}) = 0, \quad i = 1, \dots, n, \end{aligned} \tag{8.5}$$

where

$$U(\bar{s}^i+, t) = \lim_{x \to \bar{s}^i+} U(x, t), \quad U(\bar{s}^i-, t) = \lim_{x \to \bar{s}^i-} U(x, t). \tag{8.6}$$

The characteristics are no longer straight lines, but the lines $X = \bar{s}^i T$ become $x = \bar{s}^i$, so it is reasonable to look for eigenvalues and eigenfunctions. A solution of (8.4), (8.5) of the form $U(x,t) = e^{\lambda t}U(x)$, $S^i(t) = e^{\lambda t}S^i$ satisfies

$$\lambda U + \begin{cases} (Df(\bar{u}^0) - xI)U_x = 0 & \text{for } x < \bar{s}^1, \\ (Df(\bar{u}^i) - xI)U_x = 0 & \text{for } \bar{s}^i < x < \bar{s}^{i+1}, \quad i = 1, \dots, n-1, \\ (Df(\bar{u}^n) - xI)U_x = 0 & \text{for } \bar{s}^n < x, \end{cases} \tag{8.7}$$

$$\begin{aligned} (Df(\bar{u}^i) - \bar{s}^i I)U(\bar{s}^i+) - (Df(\bar{u}^{i-1}) - \bar{s}^i I)U(\bar{s}^i-) \\ - S^i(\bar{u}^i - \bar{u}^{i-1}) = 0, \quad i = 1, \dots, n, \end{aligned} \tag{8.8}$$

where

$$U(\bar{s}^i+) = \lim_{x\to\bar{s}^i+} U(x), \quad U(\bar{s}^i-) = \lim_{x\to\bar{s}^i-} U(x). \tag{8.9}$$

If we add the conditions $U(x) = 0$ for $x < \bar{s}^1$ and $\bar{s}^n < x$, then (8.7)–(8.8) is equivalent to the system (7.13)–(7.14), (7.16) that was studied in section 7.

Assuming Majda's stability condition, one can interpret (8.1)–(8.2) or (8.4)–(8.5) as describing the scattering of incoming small shock waves by the large shock waves that comprise the original Riemann solution. Several authors have found sufficient conditions that guarantee that, in some norm, the total weight of the scattered shocks is smaller than the total weight of the incoming shocks [42], [4], [5], [49], [22], [21]. For the case $n = 2$, the BV stability condition reads as follows in the notation of section 7 [49], [21]. Recall that $x_0^i = \bar{s}^i$ and $\bar{u}_0^i = \bar{u}^i$. Let

$$(\nu_1 I - Df(\bar{u}^1))^{-1}(\bar{u}^1 - \bar{u}^0) = a_1^1\mathbf{r}_1 + a_2^1\mathbf{r}_2, \tag{8.10}$$

$$(Df(\bar{u}^1) - \nu_2 I)^{-1}(\bar{u}^2 - \bar{u}^1) = a_1^2\mathbf{r}_1 + a_2^2\mathbf{r}_2. \tag{8.11}$$

Then

$$\left|\frac{a_1^2 a_2^1}{a_1^1 a_2^2}\right| < 1. \tag{8.12}$$

As in section 7, for $i = 1, 2$ let $\Delta^i = \bar{u}^i - \bar{u}^{i-1} = d_1^i\mathbf{r}_1 + d_2^i\mathbf{r}_2$, and define b and d by (7.34). Elementary computations show that

$$\frac{a_1^2 a_2^1}{a_1^1 a_2^2} = \frac{d_1^2 d_2^1(\bar{s}^1 - \nu_1)(\nu_2 - \bar{s}^2)}{d_1^1 d_2^2(\bar{s}^1 - \nu_2)(\nu_1 - \bar{s}^2)} = \frac{d}{b}, \tag{8.13}$$

and, since $b > 1$,

$$\frac{|d|}{b} < 1 \text{ if and only if } -1 + \frac{\ln|d|}{\ln b} < 0. \tag{8.14}$$

Thus the $n = 2$ BV inviscid stability condition holds if and only if all slow eigenvalues have negative real part.

9. Two Lax shocks in the *p*-system: An example. We consider the p-system

$$u_t - v_x = 0,$$

$$v_t + p(u)_x = 0,$$

with p a smooth function, $p'(u) < 0$ for all u, and $p''(u) \neq 0$ for all u.

The p-system has been used as a model for isentropic gas dynamics with $p(u) = ku^{-\gamma}$, $k > 0$, $\gamma \geq 1$ [37], [43]. The p-system is strictly hyperbolic with eigenvalues and eigenvectors

$$\nu_1(u,v) = -\sqrt{-p'(u)} < 0, \qquad \mathbf{r}_1(u,v) = (1, \sqrt{-p'(u)}),$$

$$\nu_2(u,v) = \sqrt{-p'(u)} > 0, \qquad \mathbf{r}_2(u,v) = (1, -\sqrt{-p'(u)}).$$

Consider a Riemann solution $(u_0, v_0)(x)$ that consists of two Lax shocks:

$$(u_0, v_0)(x) = \begin{cases} (\bar{u}^0, \bar{v}^0) & \text{for } x < \bar{s}^1, \\ (\bar{u}^1, \bar{v}^1) & \text{for } \bar{s}^1 < x < \bar{s}^2, \\ (\bar{u}^2, \bar{v}^2) & \text{for } \bar{s}^2 < x. \end{cases}$$

THEOREM 9.1. *To lowest order in ϵ, the corresponding Riemann–Dafermos solution has exactly the following slow eigenvalues:* (1) *a local eigenvalue with $\lambda_0 = -1$;* (2) *a family of nonlocal eigenvalues with $\lambda_0 = -2 + n\omega_0 i$, $n \in \mathbb{Z}$, $\omega_0 > 0$.*

Proof. We fix the middle state $(\bar{u}^1, \bar{v}^1)$ and look for (u, v) and s such that the Rankine–Hugoniot condition

$$-(\bar{v}^1 - v) - s(\bar{u}^1 - u) = 0, \quad p(\bar{u}^1) - p(u) - s(\bar{v}^1 - v) = 0$$

is satisfied. The solution set is two curves: Γ_1 given by

$$v = \phi(u) = \bar{v}^1 - \operatorname{sgn}(u - \bar{u}^1)\sqrt{(u - \bar{u}^1)(p(\bar{u}^1) - p(u))},$$

$$s = s^1(u) = -\sqrt{\frac{p(\bar{u}^1) - p(u)}{u - \bar{u}^1}},$$

and Γ_2 given by

$$v = \psi(u) = \bar{v}^1 + \operatorname{sgn}(u - \bar{u}^1)\sqrt{(u - \bar{u}^1)(p(\bar{u}^1) - p(u))},$$

$$s = s^2(u) = \sqrt{\frac{p(\bar{u}^1) - p(u)}{u - \bar{u}^1}}.$$

Γ_1 is a curve of 1-shocks, Γ_2 a curve of 2-shocks. Using Lax's condition for an i-shock, we easily check the following:

(1) If $(u, v, s^1) \in \Gamma_1$, then there is a 1-shock *from* (u, v) *to* $(\bar{u}^1, \bar{v}^1)$ with speed s^1 if and only if $u - \bar{u}^1 > 0$.
(2) If $(u, v, s^2) \in \Gamma_2$, then there is a 2-shock *from* $(\bar{u}^1, \bar{v}^1)$ *to* (u, v) with speed s^2 if and only if $u - \bar{u}^1 > 0$.

Therefore we have, in the notation of section 7,

$$\Delta^1 = (\bar{u}^1 - \bar{u}^0, \bar{v}^1 - \bar{v}^0) = (\bar{u}^1 - \bar{u}^0, \bar{v}^1 - \phi(\bar{u}^0)), \quad \bar{u}^0 - \bar{u}^1 > 0,$$
$$\Delta^2 = (\bar{u}^2 - \bar{u}^1, \bar{v}^2 - \bar{v}^1) = (\bar{u}^2 - \bar{u}^1, \psi(\bar{u}^2) - \bar{u}^1), \quad \bar{u}^2 - \bar{u}^1 > 0.$$

Let

$$q(u) = \sqrt{\frac{(u - \bar{u}^1)(p(\bar{u}^1) - p(u))}{-p'(\bar{u}^1)}}.$$

Then

$$\Delta^i = \sum_{j=1}^{2} d_j^i \mathbf{r}_j, \quad i = 1, 2,$$

with

$$d_1^1 = \frac{1}{2}(-(\bar{u}^0 - \bar{u}^1) - q(\bar{u}^0)), \quad d_2^1 = \frac{1}{2}(-(\bar{u}^0 - \bar{u}^1) + q(\bar{u}^0)),$$

$$d_1^2 = \frac{1}{2}(\bar{u}^2 - \bar{u}^1 - q(\bar{u}^2)), \quad d_2^2 = \frac{1}{2}(\bar{u}^2 - \bar{u}^1 + q(\bar{u}^2)).$$

Therefore

$$d = \frac{d_2^1 d_1^2}{d_1^1 d_2^2} = \frac{(\bar{u}^0 - \bar{u}^1 - q(\bar{u}^0))(\bar{u}^2 - \bar{u}^1 - q(\bar{u}^2))}{(\bar{u}^0 - \bar{u}^1 + q(\bar{u}^0))(\bar{u}^2 - \bar{u}^1 + q(\bar{u}^2))}.$$

By Lemma 9.2 below, the numerator of this fraction is positive. Therefore $d > 0$.

Let $\nu_i = \nu_i(\bar{u}^1, \bar{v}^1)$, $i = 1, 2$. Then

$$b = \frac{(\bar{s}^2 - \nu_1)(\bar{s}^1 - \nu_2)}{(\bar{s}^1 - \nu_1)(\bar{s}^2 - \nu_2)} > 1.$$

An easy computation shows that $b = \frac{1}{d}$. The result now follows from Proposition 7.8. ∎

LEMMA 9.2. *For $u > \bar{u}^1$, the sign of $u - \bar{u}^1 - q(u)$ is independent of u.*

Proof. We shall assume $p''(u) > 0$ for all u. The case $p''(u) < 0$ for all u is similar.

Let $u > \bar{u}^1$. Since $p'' > 0$ everywhere,

$$p'(\bar{u}^1) < \frac{p(\bar{u}^1) - p(u)}{\bar{u}^1 - u}.$$

Therefore

$$(u - \bar{u}^1)^2 > \frac{(u - \bar{u}^1)(p(\bar{u}^1) - p(u))}{-p'(\bar{u}^1)},$$

so $u - \bar{u}^1 > q(u)$. ∎

Acknowledgments. The original version of this paper did not relate slow eigenvalues to inviscid stability conditions. We thank the referees for pointing out this relationship, as well as the relationship between the calculation of slow eigenvalues for the p-system and the analogous calculation in [4]. We also thank Steven Schochet and Marta Lewicka for their generous help with these matters.

REFERENCES

[1] J. ALEXANDER, R. GARDNER, AND C. JONES, *A topological invariant arising in the stability analysis of traveling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.

[2] A. AZEVEDO, D. MARCHESIN, B. J. PLOHR, AND K. ZUMBRUN, *Nonuniqueness of solutions of Riemann problems*, Z. Angew. Math. Phys., 47 (1996), pp. 977–998.

[3] S. BENZONI-GAVAGE, D. SERRE, AND K. ZUMBRUN, *Alternate Evans functions and viscous shock waves*, SIAM J. Math. Anal., 32 (2001), pp. 929–962.

[4] A. BRESSAN AND R. COLOMBO, *Unique solutions of 2×2 conservation laws with large data*, Indiana Univ. Math. J., 44 (1995), pp. 677–725.

[5] A. BRESSAN AND A. MARSON, *A variational calculus for discontinuous solutions of systems of conservation laws*, Comm. Partial Differential Equations, 20 (1995), pp. 1491–1552.

[6] S.-N. CHOW AND X.-B. LIN, *Bifurcation of homoclinic orbits asymptotic to a saddle-node equilibrium*, Differential Integral Equations, 3 (1990), pp. 435–466.

[7] W. A. COPPEL, *Dichotomies in Stability Theory*, Lecture Notes in Math. 629, Springer-Verlag, Berlin, New York, 1978.

[8] C. M. DAFERMOS, *Solution of the Riemann problem for a class of hyperbolic systems of conservation laws by the viscosity method*, Arch. Ration. Mech. Anal., 52 (1973), pp. 1–9.

[9] G. DA PRATO AND P. GRISVARD, *Equations d'évolution abstraites de type parabolique*, Ann. Mat. Pura Appl., 120 (1979), pp. 329–396.

[10] J. EVANS, *Nerve axon equations* I: *Linear approximations*, Indiana Univ. Math. J., 21 (1972), pp. 877–885.

[11] J. EVANS, *Nerve axon equations* IV: *The stable and the unstable impulse*, Indiana Univ. Math. J., 24 (1974/75), pp. 1169–1190.

[12] P. C. FIFE, *Boundary and interior transition layer phenomena for pairs of second-order differential equations*, J. Math. Anal. Appl., 54 (1976), pp. 497–521.

[13] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.

[14] R. GARDNER AND K. ZUMBRUN, *The gap lemma and geometric criteria for instability of viscous shock profiles*, Comm. Pure Appl. Math., 51 (1998), pp. 789–847.

[15] J. GOODMAN, *Nonlinear asymptotic stability of viscous shock profiles for conservation laws*, Arch. Ration. Mech. Anal., 95 (1986), pp. 325–344.

[16] J. K. Hale and X.-B. Lin, *Multiple internal layer solutions generated by spatially oscillatory perturbations*, J. Differential Equations, 154 (1999), pp. 364–418.
[17] D. Henry, *Geometric theory of semilinear parabolic equations*, Lecture Notes in Math. 841, Springer-Verlag, Berlin, New York, 1981.
[18] C. K. R. T. Jones, *Geometric singular perturbation theory*, in Dynamical Systems (Montecatini Terme, 1994), Lecture Notes in Math. 1609, Springer-Verlag, Berlin, 1995, pp. 44–118.
[19] T. Kapitula and B. Sandstede, *Stability of bright solitary-wave solutions to perturbed nonlinear Schrödinger equations*, Phys. D, 124 (1998), pp. 58–103.
[20] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, New York, 1980.
[21] M. Lewicka, *L^1 stability of patterns of non-interacting large shock waves*, Indiana Univ. Math. J., 49 (2000), pp. 1515–1537.
[22] M. Lewicka, *Stability conditions for patterns of noninteracting large shock waves*, SIAM J. Math. Anal., 32 (2001), pp. 1094–1116.
[23] X.-B. Lin, *Exponential dichotomies in intermediate spaces with applications to a diffusively perturbed predator-prey model*, J. Differential Equations, 108 (1994), pp. 36–63.
[24] X.-B. Lin, *Homoclinic bifurcations with weakly expanding center manifolds*, in Dynamics Reported, Dynam. Report. Expositions Dynam. Systems (N.S.) 5, Springer-Verlag, Berlin, 1996, pp. 99–189.
[25] X.-B. Lin, *Construction and asymptotic stability of structurally stable internal layer solutions*, Trans. Amer. Math. Soc., 353 (2001), pp. 2983–3043.
[26] T.-P. Liu, *Nonlinear stability of shock waves for viscous conservation laws*, Mem. Amer. Math. Soc., 56 (1985), pp. 1–108.
[27] T.-P. Liu, *Pointwise convergence to shock waves for viscous conservation laws*, Comm. Pure Appl. Math., 50 (1997), pp. 1113–1182.
[28] T.-P. Liu and K. Zumbrun, *Nonlinear stability of an undercompressive shock for complex Burgers equation*, Comm. Math. Phys., 168 (1995), pp. 163–186.
[29] T.-P. Liu and K. Zumbrun, *On nonlinear stability of general undercompressive viscous shock waves*, Comm. Math. Phys., 174 (1995), pp. 319–345.
[30] W. Liu, *Multiple viscous wave fan profiles for Riemann solutions of hyperbolic systems of conservation laws*, Discrete Contin. Dynam. Systems Ser. B, to appear.
[31] A. Lunardi, *Bounded solutions of linear periodic abstract parabolic equations*, Proc. Roy. Soc. Edinburgh Sect. A, 110 (1998), pp. 135–159.
[32] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Appl. Math. Sci. 53, Springer-Verlag, New York, 1984.
[33] K. J. Palmer, *Exponential dichotomies and transversal homoclinic points*, J. Differential Equations, 55 (1984), pp. 225–256.
[34] A. Matsumura and K. Nishihara, *On the stability of travelling wave solutions of a one-dimensional model system for compressible viscous gas*, Japan J. Appl. Math., 2 (1985), pp. 17–25.
[35] Y. Nishiura and H. Fujii, *SLEP method to the stability of singularly perturbed solutions with multiple transition layers in reaction-diffusion systems*, in Dynamics of Infinite-Dimensional Systems (Lisbon, 1986), NATO Adv. Sci. Inst. Ser. F Comput. Systems Sci. 37, Springer-Verlag, Berlin, 1987, pp. 211–230.
[36] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, 1983.
[37] M. Renardy and R. Rogers, *An Introduction to Partial Differential Equations*, Springer-Verlag, New York, 1993.
[38] B. Sandstede, *Stability of travelling waves*, in Handbook of Dynamical Systems II: Towards Applications, B. Fiedler, ed., North-Holland, Amsterdam, 2002, pp. 983–1055.
[39] D. H. Sattinger, *On the stability of waves of nonlinear parabolic systems*, Adv. Math., 22 (1976), pp. 312–355.
[40] S. Schecter, D. Marchesin, and B. J. Plohr, *Structurally stable Riemann solutions*, J. Differential Equations, 126 (1996), pp. 303–354.
[41] S. Schecter, *Undercompressive shock waves and the Dafermos regularization*, Nonlinearity, 15 (2002), pp. 1361–1377.
[42] S. Schochet, *Sufficient conditions for local existence via Glimm's scheme for large BV data*, J. Differential Equations, 89 (1991), pp. 317–354.
[43] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.
[44] P. Szmolyan, in preparation.
[45] A. Szepessy and K. Zumbrun, *Stability of rarefaction waves in viscous media*, Arch. Ration. Mech. Anal., 133 (1996), pp. 249–298.
[46] V. A. Tupčiev, *On the splitting of an arbitrary discontinuity for a system of two first-order*

quasi-linear equations, USSR Comput. Math. Math. Phys., 4 (1964), pp. 36–48.

[47] V. A. Tupčiev, *The method of introducing a viscosity in the study of a problem of decay of a discontinuity*, Soviet Math. Dokl., 14 (1973), pp. 978–982.

[48] A. E. Tzavaras, *Wave interactions and variation estimates for self-similar zero-viscosity limits in systems of conservation laws*, Arch. Ration. Mech. Anal., 135 (1996), pp. 1–60.

[49] Y. Wang, *Highly oscillatory shock waves*, in Nonlinear Theory of Generalized Functions (Vienna, 1997), Chapman & Hall/CRC Res. Notes Math. 401, Chapman & Hall/CRC, Boca Raton, FL, 1999, pp. 153–161.

[50] K. Zumbrun and P. Howard, *Pointwise semigroup methods and stability of viscous shock waves*, Indiana Univ. Math. J., 47 (1998), pp. 741–871.

SIAM J. MATH. ANAL.
Vol. 35, No. 4, pp. 922–948

PERMANENT REGIMES FOR THE 1D VLASOV–POISSON SYSTEM WITH BOUNDARY CONDITIONS*

M. BOSTAN†

Abstract. We prove the existence of weak solutions for the Vlasov–Poisson problem with time periodic boundary conditions in one dimension. We consider boundary data with finite charge and current. This analysis is based upon the mild formulation for the regularized Vlasov–Poisson equations.

Key words. Vlasov–Poisson equations, weak/mild formulation, regularization

AMS subject classifications. 35Q99, 35L50

DOI. 10.1137/S0036141002416420

1. Introduction. Many studies in physics and applied physics are modeled by kinetic equations (Vlasov, Boltzmann, etc.) coupled with the electromagnetic equations (Poisson, Maxwell). A few application domains are semiconductors, particle accelerators, and electron guns. Various results were shown for free space systems. Weak solutions of the Vlasov–Poisson equations were constructed by Arseneev [2], Illner and Neunzert [16], and Horst and Hunze [15]. The existence of weak solutions of the Vlasov–Maxwell system was shown by DiPerna and Lions [11].

There are few mathematical works on boundary value problems. For the stationary case, results have been obtained by Greengard and Raviart [13] for the one dimensional (1D) Vlasov–Poisson system and by Poupaud [17] for the multidimensional Vlasov–Maxwell system. An asymptotic analysis of the Vlasov–Poisson system has been performed by Degond and Raviart in [10] in the case of the plane diode. Weak solutions of the initial-boundary value problem for the Vlasov–Poisson system are obtained by Abdallah in [1]. The regularity of the solution for the Vlasov–Maxwell system in a half line has been analyzed by Guo in [14].

The periodic case has been studied as well (see [5], [6], [7]), but existence results are available only under some restrictive hypothesis concerning the velocity support of the boundary incoming particle distribution and the potential data. Basically the above model does not handle charge flows with small incoming velocities. The main idea was to keep only the particles which are travelling through the domain in finite time, which makes it possible to get estimates for the charge and current densities.

In this paper we study the existence for the 1D Vlasov–Poisson problem with time periodic boundary conditions

$$\partial_t f + v \cdot \partial_x f + E(t,x) \cdot \partial_v f = 0, \quad (t,x,v) \in \mathbb{R}_t \times]0,1[\times \mathbb{R}_v,$$

$$f(t, x = 0, v > 0) = g_0(t, v > 0), \quad f(t, x = 1, v < 0) = g_1(t, v < 0), \ t \in \mathbb{R}_t,$$

$$E(t,x) = -\partial_x U, \quad -\partial_x^2 U = \rho(t,x) := \int_{\mathbb{R}_v} f(t,x,v)dv, \quad (t,x) \in \mathbb{R}_t \times]0,1[,$$

*Received by the editors October 21, 2002; accepted for publication (in revised form) May 9, 2003; published electronically November 4, 2003.
http://www.siam.org/journals/sima/35-4/41642.html
†Université de Franche-Comté, 16 route de Gray F-25030, Besançon Cedex, France (mbostan@descartes.univ-fcomte.fr).

$$U(t,x=0)=\varphi_0(t),\ \ U(t,x=1)=\varphi_1(t),\ \ t\in\mathbb{R}_t.$$

The function $f(t,x,v)$ denotes the particle distribution depending on time t, position x, and velocity v. The electric field $E(t,x)$ derives from an electrostatic potential U satisfying the Poisson equation with charge density ρ. The boundary conditions $g_0, g_1, \varphi_0, \varphi_1$ are supposed T-periodic in time for some $T>0$.

The main goal of this paper is to establish existence in the general case, under a minimal hypothesis, say, for incoming particle distribution with finite charge (as has been shown for the stationary case in [13]). In this case we prove that the solution f belongs to L^1. The major difficulty in studying this problem is the lack of a natural a priori estimate of the solution. In fact, since we are looking for permanent regimes, initial data are not available, and therefore directly applying conservation laws like

$$\int_0^1\!\!\int_{\mathbb{R}_v} f(t,x,v)dxdv \le \int_0^1\!\!\int_{\mathbb{R}_v} f(t_0,x,v)dxdv + \int_{t_0}^t\!\!\int_{v>0} vg_0(s,v)dsdv - \int_{t_0}^t\!\!\int_{v<0} vg_1(s,v)dsdv,\ \ t>t_0,$$

does not provide any estimate as long we don't have any information on $f(t_0)$. On the other hand, even if there is $t_0\in\mathbb{R}$ such that $f(t_0)\in L^1$, the previous inequality gives us only an estimate of the charge in terms of the incoming current, whereas the natural estimate would be in terms of the incoming charge. In fact we can prove that if the incoming particle distribution has finite current (resp., kinetic energy), then the solution verifies $|v|f\in L^1$ (resp., $|v|^2 f\in L^1$).

This work begins with the study of linear time periodic Vlasov equations (the electric field is assumed to be known and T-periodic). We introduce the weak and mild formulations and recall some usual computations for such solutions. We also introduce a perturbed Vlasov equation. In this section we present a very important lemma concerning bounds for the velocity change along the characteristics (see Lemma 2.11), which states that along all characteristics associated with a regular field the following inequality holds:

$$|\,V(s_1)-V(s_2)\,|\le C\cdot\|E\|_{L^\infty}^{1/2}\ \ \forall\, s_1,s_2,$$

where C is a constant depending only on the diameter of the spatial domain $\Omega=]0,1[$ here.

In section 3 the Vlasov–Poisson system is analyzed. The existence of a T-periodic solution will be obtained by application of the Schauder fixed point theorem. The nonuniqueness of the solution for the Vlasov problem does not allow us to directly apply the fixed point method. We need to introduce a perturbed problem by adding an absorption term αf in the Vlasov equation, where $\alpha>0$ is a small parameter, and to regularize the electric field, which allows us to use the mild formulation. The perturbed problem can be written as

$$\alpha f(t,x,v)+\partial_t f+v\cdot\partial_x f+E_\varepsilon(t,x)\cdot\partial_v f=0,\ \ (t,x,v)\in\mathbb{R}_t\times]0,1[\times\mathbb{R}_v,$$

$$f(t,x=0,v>0)=g_0(t,v>0),\ \ f(t,x=1,v<0)=g_1(t,v<0),\ t\in\mathbb{R}_t,$$

$$E_\varepsilon(t,x)=\int_{\mathbb{R}}\zeta_\varepsilon(t-s)ds\int_0^1\zeta_\varepsilon(x-y)E(s,y)dy,\ (t,x)\in\mathbb{R}_t\times]0,1[,$$

where E is the electric field given by the Poisson equation with source $\rho(t,x) = \int_{\mathbb{R}_v} f(t,x,v)dv$,

$$E(t,x) = \int_0^x \rho(t,y)dy - \int_0^1 (1-y)\rho(t,y)dy - \varphi_1(t) + \varphi_0(t),\ (t,x) \in \mathbb{R}_t \times]0,1[,$$

and $\zeta_\varepsilon(\cdot) = \frac{1}{\varepsilon}\zeta(\frac{\cdot}{\varepsilon}),\ \varepsilon > 0$ is a mollifier sequence. Clearly, the perturbed Vlasov problem has a unique T-periodic weak solution, and the existence for the nonlinear perturbed problem follows easily by fixed point argument. Indeed, in this case ($\alpha > 0$ fixed), we immediately obtain the following estimate for the T-periodic weak solution of the perturbed Vlasov problem:

$$\int_0^1 \int_{\mathbb{R}_v} f(t,x,v)dxdv \leq \left(\frac{1}{\alpha T} + 1\right)\left(\int_0^T \int_{v>0} vg_0(t,v)dtdv - \int_0^T \int_{v<0} vg_1(t,v)dtdv\right),\ t \in \mathbb{R}_t,$$

which allows us to define a fixed point application.

Obviously, the main difficulty consists of finding uniform estimates for the perturbed problems with $\alpha > 0, \varepsilon > 0$. In section 4 we obtain estimates for the total charge and current and the electric field. The main tool is Lemma 2.11 combined with the mild formulation. In fact the previous lemma allows us to get bounds on the particle lifetimes, at least for particles with initial velocities v large enough. Indeed, since along a characteristic we have $\mid V(s) - v \mid \leq C \cdot \|E\|_{L^\infty}^{1/2},\ \ s_{in} \leq s \leq s_{out}$, we deduce that $|V(s)|$ is bounded from below $|V(s)| \geq |v| - C \cdot \|E\|_{L^\infty}^{1/2}$ and therefore

$$s_{out} - s_{in} \leq \frac{1}{|v| - C \cdot \|E\|_{L^\infty}^{1/2}} \ \text{ if } \ |v| > C \cdot \|E\|_{L^\infty}^{1/2}.$$

These arguments work for bounded spatial domains.

In section 5 we prove the existence of the T-periodic weak solution for the Vlasov–Poisson system by passing $\alpha \to 0$. In order to pass to the limit in the nonlinear term $E_n \cdot \partial_v f_n$, we can combine the strong convergence in L^1 of E_n with the weak $\star$ convergence in L^∞ of f_n. Some generalizations are analyzed as well. Basically, for incoming data satisfying $|v|^p g \in L^1$ for some integer $p \geq 1$, we prove that $|v|^p f \in L^1$.

We end this paper with several remarks and conclusions. We investigate the Vlasov–Poisson system with several species of particles, as well as the case of attractive (gravitational) potentials.

2. The Vlasov equation. The equation which governs the transport of charged particles is called the Vlasov equation, and in one dimension it is given by

$$\partial_t f + v \cdot \partial_x f + E(t,x) \cdot \partial_v f = 0, \quad (t,x,v) \in \mathbb{R}_t \times]0,1[\times \mathbb{R}_v, \tag{2.1}$$

where $f(t,x,v)$ is the density of particles under the action of the electric field $E(t,x) = -\partial_x U$ and $U(t,x)$ is the potential. Charged particles are injected through the boundary

$$f(t,x,v) = g(t,x,v), \quad (t,x,v) \in \mathbb{R}_t \times \Sigma^-, \tag{2.2}$$

where Σ^- is the subset of the boundary of the phase space $]0,1[\times\mathbb{R}_v$ corresponding to the incoming velocities,

$$\Sigma^- = \{(0,v) \mid v>0\} \cup \{(1,v) \mid v<0\} = \Sigma_0^- \cup \Sigma_1^-.$$

Similarly we also define $\Sigma^+ = \{(0,v) \mid v<0\} \cup \{(1,v) \mid v>0\} = \Sigma_0^+ \cup \Sigma_1^+$, which corresponds to the outgoing velocities and $\Sigma^0 = \{(0,0),(1,0)\}$. With the notation $g|_{\mathbb{R}_t\times\Sigma_0^-} = g_0, g|_{\mathbb{R}_t\times\Sigma_1^-} = g_1$, the previous boundary condition (2.2) can be written as

$$f(t,x=0,v>0) = g_0(t,v>0), \quad f(t,x=1,v<0) = g_1(t,v<0). \tag{2.3}$$

The functions $g_0, g_1 \geq 0$, which describe the emission profiles of the injected charged particles, are supposed T-periodic in time, $T>0$. Now let us briefly recall the definition of weak and mild solutions for the Vlasov problem (2.1), (2.3).

2.1. Weak solution for the Vlasov problem. We introduce the spaces L_i^-, $L_{i,loc}^-$ of incoming data with bounded or locally bounded fluxes:

$$L_i^- = \{g(t,v) \mid v\cdot g(t,v) \in L^1(]0,T[\times\Sigma_i^-)\},$$

$$L_{i,loc}^- = \{g(t,v) \mid v\cdot g(t,v) \in L^1_{loc}(]0,T[\times\Sigma_i^-)\},$$

where $i=0,1$. We shall use also the following notation:

$$G_p := \frac{1}{T}\int_0^T\!\!\int_{v>0} |v|^p g_0(t,v)dtdv + \frac{1}{T}\int_0^T\!\!\int_{v<0} |v|^p g_1(t,v)dtdv, \;\; 0\leq p<+\infty,$$

and

$$G_\infty := \max\{\|g_0\|_{L^\infty(\mathbb{R}_t\times\Sigma_0^-)}, \|g_1\|_{L^\infty(\mathbb{R}_t\times\Sigma_1^-)}\}$$

when g_0, g_1 belong to the corresponding spaces.

DEFINITION 2.1. *Assume that $E\in L^\infty(\mathbb{R}_t\times]0,1[)$ and $g_0\in L_{0,loc}^-$, $g_1\in L_{1,loc}^-$ are T-periodic functions in time. We say that $f\in L^1_{loc}(]0,T[\times]0,1[\times\mathbb{R}_v)$ is a T-periodic weak solution for the Vlasov problem* (2.1), (2.3) *iff*

$$-\int_0^T\!\!\int_0^1\!\!\int_{\mathbb{R}_v} f(t,x,v)(\partial_t\varphi + v\cdot\partial_x\varphi + E(t,x)\cdot\partial_v\varphi)dtdxdv = \int_0^T\!\!\int_{v>0} vg_0(t,v)\varphi(t,0,v)dtdv - \int_0^T\!\!\int_{v<0} vg_1(t,v)\varphi(t,1,v)dtdv,$$

for all test functions $\varphi\in\mathcal{T}_w$, where

$$\mathcal{T}_w = \{\varphi\in W^{1,\infty}(\mathbb{R}_t\times]0,1[\times\mathbb{R}_v) \mid \varphi \text{ is } T\text{-periodic in time }, \varphi|_{\mathbb{R}_t\times\Sigma^+}=0,$$

$$\exists R>0 : \mathrm{supp}(\varphi)\subset\mathbb{R}_t\times[0,1]\times B_R\}.$$

2.2. Mild solution for the Vlasov problem. Throughout this paper we also need to consider some special solutions for (2.1), (2.3), which are called mild solutions or solutions by characteristics. These solutions require more regularity for

the electric field, and they are particular cases of weak solutions. Assume that $E \in L^\infty(\mathbb{R}_t; W^{1,\infty}(]0,1[))$ is T-periodic, and for $(t,x,v) \in \mathbb{R}_t \times]0,1[\times \mathbb{R}_v$ denote by $(X(s;t,x,v), V(s;t,x,v))$ the unique solution for the system of ordinary differential equations

$$\frac{d}{ds}X(s;t,x,v) = V(s;t,x,v), \quad \frac{d}{ds}V(s;t,x,v) = E(s, X(s;t,x,v)) \tag{2.4}$$

for $s \in (s_{in}, s_{out})$ which verifies the condition

$$X(s=t;t,x,v) = x, \quad V(s=t;t,x,v) = v. \tag{2.5}$$

Here $s_{in} = s_{in}(t,x,v)$ (resp., $s_{out} = s_{out}(t,x,v)$) represents the incoming (resp., outgoing) time of the particle in the domain $]0,1[$ defined by

$$s_{in}(t,x,v) = \sup\{s \le t \ : \ (X(s;t,x,v), V(s;t,x,v)) \in \Sigma^-\} \ge -\infty$$

and

$$s_{out}(t,x,v) = \inf\{s \ge t \ : \ (X(s;t,x,v), V(s;t,x,v)) \in \Sigma^+ \cup \Sigma^0\} \le +\infty.$$

Using the previous notation, the total travel time through the domain (lifetime) can be written as $\tau(t,x,v) = s_{out}(t,x,v) - s_{in}(t,x,v) \le +\infty$. Now we replace in Definition 2.1 the function $\partial_t\varphi + v \cdot \partial_x\varphi + E(t,x) \cdot \partial_v\varphi$ by ψ, which, after integration along the characteristics curves gives

$$\varphi(t,x,v) = -\int_t^{s_{out}(t,x,v)} \psi(s, X(s;t,x,v), V(s;t,x,v))\, ds,$$

and we define the mild solutions as follows.

Definition 2.2. *Assume that $E \in L^\infty(\mathbb{R}_t; W^{1,\infty}(]0,1[))$ and $g_0 \in L^-_{0,loc}$, $g_1 \in L^-_{1,loc}$ are T-periodic functions in time. We say that $f \in L^1_{loc}(]0,T[\times]0,1[\times\mathbb{R}_v)$ is a T-periodic mild solution for the Vlasov problem* (2.1), (2.3) *iff*

$$\begin{aligned}
&\int_0^T\int_0^1\int_{\mathbb{R}_v} f(t,x,v)\psi(t,x,v)dtdxdv \\
&\quad = \int_0^T\int_{v>0} v g_0(t,v) \int_t^{s_{out}(t,0,v)} \psi(s, X(s;t,0,v), V(s;t,0,v))dsdtdv \\
&\quad - \int_0^T\int_{v<0} v g_1(t,v) \int_t^{s_{out}(t,1,v)} \psi(s, X(s;t,1,v), V(s;t,1,v))dsdtdv,
\end{aligned}$$

for all test functions $\psi \in \mathcal{T}_m$, where

$$\mathcal{T}_m = \{\psi \in L^\infty(\mathbb{R}_t \times]0,1[\times\mathbb{R}_v) \mid \psi \text{ is } T\text{-periodic in time},$$

$$\exists R > 0 : \operatorname{supp}(\psi) \subset \mathbb{R}_t \times [0,1] \times B_R\}.$$

We shall sometimes use the notation

$$(X(s), V(s)) = (X(s;t,x,v), V(s;t,x,v)),$$

$$(X^0(s), V^0(s)) = (X(s;t,0,v), V(s;t,0,v)),$$

$$(X^1(s), V^1(s)) = (X(s;t,1,v), V(s;t,1,v)),$$

and

$$s_{out} = s_{out}(t,x,v), \quad s^0_{out} = s_{out}(t,0,v), \quad s^1_{out} = s_{out}(t,1,v),$$

$$s_{in} = s_{in}(t,x,v), \quad s^0_{in} = s_{in}(t,0,v), \quad s^1_{in} = s_{in}(t,1,v).$$

REMARK 2.3. *In fact the mild solution is given by* $f(t,x,v)=g_i(s_{in}, V(s_{in};t,x,v))$ *if* $s_{in}(t,x,v) > -\infty$ *and* $X(s_{in};t,x,v) = i$, *where* $i = 0,1$ *and* $f(t,x,v) = 0$ *otherwise.*

REMARK 2.4. *Since* E *is* T*-periodic, we have* $X(s+T;t+T,x,v) = X(s;t,x,v)$, $V(s+T;t+T,x,v) = V(s;t,x,v)$, $s_{in}(t+T,x,v) = s_{in}(t,x,v)+T$ *for all* $(s,t,x,v) \in \mathbb{R}_s \times \mathbb{R}_t \times]0,1[\times \mathbb{R}_v$, *and thus, by the periodicity of* g_0, g_1, *it follows that the mild solution is* T*-periodic.*

REMARK 2.5. *There is in general no uniqueness for the weak solution because* f *can take arbitrary values on the characteristics such that* $s_{in} = -\infty$. *However, it is possible to prove that the mild solution is the unique minimal solution for the transport equation (see* [17] *and* [4] *for definitions and proofs).*

2.3. Weak and mild solutions for the perturbed Vlasov problem. We intend to apply a fixed point procedure on the electric field. For example, let us define the following map:

$$\begin{aligned} E \to f_E \text{ solution of the Vlasov problem } &\to \rho_E \text{ charge density of } f_E \\ &\to E_1 \text{ solution of the Poisson problem with source } \rho_E. \end{aligned}$$

Unfortunately the above map is not well defined since we have no uniqueness for the Vlasov problem. In order to recover the uniqueness property, we need to introduce an absorption term αf, $\alpha > 0$. The perturbed Vlasov equation is now written as

$$\alpha f(t,x,v) + \partial_t f + v \cdot \partial_x f + E(t,x) \cdot \partial_v f = 0, \quad (t,x,v) \in \mathbb{R}_t \times]0,1[\times \mathbb{R}_v. \tag{2.6}$$

Obviously, the weak and mild formulations previously introduced for the Vlasov problem still hold for the perturbed problem with the corresponding modifications due to the term αf (when $\alpha = 0$ we recover Definitions 2.1 and 2.2).

DEFINITION 2.6. *Under the same hypothesis as in Definition* 2.1, *we say that* f *is a* T*-periodic weak solution for the perturbed Vlasov problem* (2.6), (2.3) *iff*

$$\begin{aligned} &-\int_0^T\int_0^1\int_{\mathbb{R}_v} f(t,x,v)(-\alpha\varphi + \partial_t\varphi + v\cdot\partial_x\varphi + E(t,x)\cdot\partial_v\varphi)dtdxdv \\ &\quad = \int_0^T\int_{v>0} vg_0(t,v)\varphi(t,0,v)dtdv - \int_0^T\int_{v<0} vg_1(t,v)\varphi(t,1,v)dtdv, \end{aligned}$$

for all test functions $\varphi \in \mathcal{T}_w$.

REMARK 2.7. *After multiplication by* f *and integration on* $]0,T[\times]0,1[\times\mathbb{R}_v$ *we can easily check that there is a unique weak solution for the perturbed Vlasov problem (see* [6, p. 657], [3], [12]*).*

DEFINITION 2.8. *Under the same hypothesis as in Definition* 2.2 *we say that* f *is a* T*-periodic mild solution for the perturbed Vlasov problem* (2.6), (2.3) *iff*

$$\begin{aligned}\int_0^T\int_0^1\int_{\mathbb{R}_v} f(t,x,v)\psi(t,x,v)dtdxdv \\ = \int_0^T\int_{v>0} vg_0(t,v)dtdv \int_t^{s_{out}^0} e^{-\alpha(s-t)}\psi(s,X(s;t,0,v),V(s;t,0,v))ds \\ - \int_0^T\int_{v<0} vg_1(t,v)dtdv \int_t^{s_{out}^1} e^{-\alpha(s-t)}\psi(s,X(s;t,1,v),V(s;t,1,v))ds\end{aligned}$$

for all test functions $\psi \in \mathcal{T}_m$.

REMARK 2.9. *We can easily check that, if* $g_0 \in L_0^-$, $g_1 \in L_1^-$, *then the mild solution belongs to* $L^1(]0,T[\times]0,1[\times\mathbb{R}_v)$.

Indeed, let us consider $\chi \in C^1(\mathbb{R})$, $0 \le \chi \le 1$, $\mathrm{supp}(\chi) \subset [-2,2]$, $\chi|_{[-1,1]} = 1$. By taking $\psi_R(t,x,v) = \chi(v/R) \in \mathcal{T}_m$, as a test function, we have

$$\begin{aligned}\int_0^T\int_0^1\int_{|v|<R} f(t,x,v)dtdxdv &\le \int_0^T\int_0^1\int_{\mathbb{R}_v} f(t,x,v)\psi_R(t,x,v)dtdxdv \\ &= \int_0^T\int_{v>0} vg_0(t,v)dtdv \int_t^{s_{out}^0} e^{-\alpha(s-t)}\chi\left(\frac{V^0(s)}{R}\right)ds \\ &- \int_0^T\int_{v<0} vg_1(t,v)dtdv \int_t^{s_{out}^1} e^{-\alpha(s-t)}\chi\left(\frac{V^1(s)}{R}\right)ds \\ &\le \frac{1}{\alpha}\left(\int_0^T\int_{v>0} vg_0(t,v)dtdv - \int_0^T\int_{v<0} vg_1(t,v)dtdv\right),\ R>0.\end{aligned}$$

Thus, by passing $R \to +\infty$, we deduce that f belongs to $L^1(]0,T[\times]0,1[\times\mathbb{R}_v)$ and that

$$\frac{1}{T}\|f\|_{L^1} \le \frac{G_1}{\alpha}, \quad \alpha > 0. \tag{2.7}$$

REMARK 2.10. *Moreover, under the same hypothesis as in the previous remark, if* $\psi \in L^\infty$ *is* T*-periodic with unbounded velocity support, then the mild formulation still holds.*

For this let us formulate a lemma concerning bounds for the velocity change along the characteristics. This result is the key point of our analysis, and it will be used several times throughout this paper.

LEMMA 2.11. *Assume that* $E \in L^\infty(\mathbb{R}_t; W^{1,\infty}(]0,1[))$ *is a regular electric field. Then, for all characteristics* $(X(s),V(s))$, $s_{in} \le s \le s_{out}$, *we have*

$$|V(s_1) - V(s_2)| \le 2\sqrt{2}\cdot\|E\|_{L^\infty}^{1/2}, \quad s_{in} \le s_1 \le s_2 \le s_{out}.$$

Proof. If $|V(s_{1,2})| \le \sqrt{2}\cdot\|E\|_{L^\infty}^{1/2}$ or $\|E\|_{L^\infty} = 0$, the conclusion follows trivially. Suppose that $\|E\|_{L^\infty} > 0$, $|V(s_1)| > \sqrt{2}\cdot\|E\|_{L^\infty}^{1/2}$; for the other case, the same argument applies. By integration along the characteristics curves we find

$$V(s) \ge V(s_1) - (s-s_1)\|E\|_{L^\infty},\ s \in [s_1,s_2],$$

$$V(s_1) \ge V(s) - (s-s_1)\|E\|_{L^\infty},\ s \in [s_1,s_2],$$

and also

$$1 \geq X(s) - X(s_1) \geq (s - s_1)V(s_1) - \frac{1}{2}(s - s_1)^2\|E\|_{L^\infty},\ s \in [s_1, s_2],$$

$$1 \geq X(s_1) - X(s) \geq -(s - s_1)V(s_1) - \frac{1}{2}(s - s_1)^2\|E\|_{L^\infty},\ s \in [s_1, s_2].$$

Therefore $F(s) := \frac{1}{2}(s - s_1)^2\|E\|_{L^\infty} - |V(s_1)|(s - s_1) + 1 \geq 0,\ s \in [s_1, s_2]$, and, since the discriminant $\Delta = |V(s_1)|^2 - 2\|E\|_{L^\infty}$ is positive, it follows that the quadratic function F has two real roots $s_1 < r_1 < r_2$ given by

$$r_{1,2} = s_1 + \frac{|V(s_1)| \mp \sqrt{|V(s_1)|^2 - 2\|E\|_{L^\infty}}}{\|E\|_{L^\infty}}.$$

On the other hand, we have

$$F(s_2) = \frac{\|E\|_{L^\infty}}{2}\left(s_2 - s_1 - \frac{|V(s_1)|}{\|E\|_{L^\infty}}\right)^2 + 1 - \frac{|V(s_1)|^2}{2\|E\|_{L^\infty}} \geq 0,$$

and therefore we deduce that

$$\left|s_2 - s_1 - \frac{|V(s_1)|}{\|E\|_{L^\infty}}\right| > \frac{\sqrt{\Delta}}{\|E\|_{L^\infty}}.$$

If $s_2 - s_1 - |V(s_1)|/\|E\|_{L^\infty} < -\sqrt{\Delta}/\|E\|_{L^\infty}$, by using the fact that $|V(s_1)| > \sqrt{2} \cdot \|E\|_{L^\infty}^{1/2}$ we have

$$|V(s_1) - V(s_2)| \leq (s_2 - s_1)\|E\|_{L^\infty} \leq |V(s_1)| - \sqrt{|V(s_1)|^2 - 2\|E\|_{L^\infty}} < \sqrt{2} \cdot \|E\|_{L^\infty}^{1/2}.$$

Now let us consider the case when $s_2 - s_1 - |V(s_1)|/\|E\|_{L^\infty} > \sqrt{\Delta}/\|E\|_{L^\infty}$, which implies that $s_2 > s_1 + (|V(s_1)| + \sqrt{\Delta})/\|E\|_{L^\infty} = r_2$. Therefore we have $s_1 < r_1 < r_2 < s_2$, which is in contradiction with $F(s) \geq 0,\ s \in [s_1, s_2]$, since $F(s) < 0$ for $s \in (r_1, r_2) \subset [s_1, s_2]$. ∎

Now let us consider the mild test function $\psi_R(t, x, v) = \psi(t, x, v) \cdot \chi(v/R) \in \mathcal{T}_m$. In order to simplify the calculation, we treat only the terms of the left boundary located in $x = 0$. Exactly the same calculus applies for the right boundary in $x = 1$. We have

$$\begin{aligned}
&\int_0^T\int_0^1\int_{|v|<R} f\psi dtdxdv + \int_0^T\int_0^1\int_{|v|>R} f\psi\chi\left(\frac{v}{R}\right)dtdxdv \\
&= \int_0^T\int_{0<v<R_1} vg_0(t,v)dtdv \int_t^{s_{out}^0} e^{-\alpha(s-t)}\psi\chi\left(\frac{V^0(s)}{R}\right)ds \\
&+ \int_0^T\int_{v>R_1} vg_0(t,v)dtdv \int_t^{s_{out}^0} e^{-\alpha(s-t)}\psi\chi\left(\frac{V^0(s)}{R}\right)ds + \{\text{right boundary terms}\} \\
&= \mathcal{I}_1(R) + \mathcal{I}_2(R) + \mathcal{I}_3(R) + \mathcal{I}_4(R), \qquad (2.8)
\end{aligned}$$

where $R_1 = R - 2\sqrt{2} \cdot \|E\|_{L^\infty}^{1/2}$. By the previous lemma, we deduce that for $0 < v < R_1$ we have $|V^0(s)| \leq 2\sqrt{2} \cdot \|E\|_{L^\infty}^{1/2} + v \leq R$, and therefore $\chi(V^0(s)/R) = 1$ for $s \in (t, s_{out}^0)$, which implies that

$$\lim_{R\to+\infty} \mathcal{I}_1(R) = \int_0^T\int_{v>0} vg_0(t,v)dtdv \int_t^{s_{out}^0} e^{-\alpha(s-t)}\psi(s, X^0(s), V^0(s))ds.$$

On the other hand, since $|\int_t^{s^0_{out}} e^{-\alpha(s-t)}\psi\chi(V^0(s)/R)ds| \le \|\psi\|_{L^\infty}/\alpha$ and $g_0 \in L^-_0$, we also have the convergence

$$\mathcal{I}_2(R) \le \frac{1}{\alpha}\|\psi\|_{L^\infty}\int_0^T\!\!\int_{v>R_1} vg_0(t,v)dtdv \to 0,$$

when $R \to +\infty$. Since f belongs to $L^1(]0,T[\times]0,1[\times\mathbb{R}_v)$, $\psi \in L^\infty$, $0 \le \chi \le 1$, we can pass to the limit in (2.8) for $R \to +\infty$ and the mild formulation holds. In particular, for $\psi_R = 1_{\{|v|>R\}}$, we have

$$\begin{aligned}\int_0^T\!\!\int_0^1\!\!\int_{|v|>R} f(t,x,v)dtdxdv &= \int_0^T\!\!\int_{v>0} vg_0(t,v)dtdv\int_t^{s^0_{out}} e^{-\alpha(s-t)}1_{\{|V^0(s)|>R\}}ds + \{\cdots\}\\ &= \int_0^T\!\!\int_{v>R_1} vg_0(t,v)dtdv\int_t^{s^0_{out}} e^{-\alpha(s-t)}1_{\{|V^0(s)|>R\}}ds + \{\cdots\}\\ &\le \frac{1}{\alpha}\left(\int_0^T\!\!\int_{v>R_1} vg_0(t,v)dtdv - \int_0^T\!\!\int_{v<-R_1} vg_1(t,v)dtdv\right).\end{aligned}$$

REMARK 2.12. *If $g_0 \in L^-_0$, $g_1 \in L^-_1$, and $E \in L^\infty$, then all T-periodic weak solutions to* (2.6), (2.3) *belong to $L^1(]0,T[\times]0,1[\times\mathbb{R}_v)$ and verify the same estimate* (2.7).

REMARK 2.13. *If $g_0 \in L^-_0$, $g_1 \in L^-_1$, and $E \in L^\infty$, then the weak formulation also holds for test function $\varphi \in W^{1,\infty}$ with unbounded support in velocity (take as a test function $\varphi_R = \varphi\cdot\chi(v/R)$ and pass $R \to +\infty$).*

3. The Vlasov–Poisson system. The electric field is due to the charge of the particles (it is a self-consistent field)

$$\partial_x E = -\partial_x^2 U = \rho(t,x) := \int_{\mathbb{R}_v} f(t,x,v)dv,\ (t,x) \in \mathbb{R}_t\times]0,1[, \tag{3.1}$$

and to the applied voltage on the boundary,

$$U(t,x=0) = \varphi_0(t),\ U(t,x=1) = \varphi_1(t),\ t \in \mathbb{R}_t. \tag{3.2}$$

As above, the electrostatic potentials φ_0, φ_1 are supposed T-periodic in time. The system formed by (2.1) and (3.1) and the boundary conditions (2.3) and (3.2) are called the Vlasov–Poisson problem (in one dimension). Obviously, in one dimension the Poisson electric field can be written as

$$E(t,x) = \int_0^x \rho(t,y)dy - \int_0^1 (1-y)\rho(t,y)dy - \varphi_1(t) + \varphi_0(t),\ (t,x) \in \mathbb{R}_t\times]0,1[,$$

and therefore we can give the following definition.

DEFINITION 3.1. *Assume that $g_0 \in L^-_{0,loc}$, $g_1 \in L^-_{1,loc}$, $\varphi_1 - \varphi_0 \in L^\infty(\mathbb{R}_t)$ are T-periodic functions. We say that $(f,E) \in L^1(]0,T[\times]0,1[\times\mathbb{R}_v)\times L^\infty(\mathbb{R}_t\times]0,1[)$ is a T-periodic weak solution for the Vlasov–Poisson problem iff f is a T-periodic weak solution for the Vlasov problem* (2.1), (2.3) *corresponding to the electric field E given by the Poisson problem*

$$E(t,x) = \int_0^x \rho(t,y)dy - \int_0^1 (1-y)\rho(t,y)dy - \varphi_1(t) + \varphi_0(t),\ (t,x) \in \mathbb{R}_t\times]0,1[,$$

with $\rho(t,x) := \int_{\mathbb{R}_v} f(t,x,v)dv,\ (t,x) \in \mathbb{R}_t\times]0,1[$.

As was explained in section 2.3, we need to also consider a perturbed system. Let us introduce the notion of T-periodic mild solution for the perturbed Vlasov–Poisson problem. For this we have to regularize the electric field; we consider mollifiers $\zeta_\varepsilon(\cdot) = \frac{1}{\varepsilon}\zeta(\frac{\cdot}{\varepsilon})$, $\varepsilon > 0$, where $\zeta \in C_0^\infty(\mathbb{R})$, $\zeta \geq 0$, $\mathrm{supp}(\zeta) \subset [-1,+1]$, $\int_{\mathbb{R}} \zeta(u)du = 1$.

DEFINITION 3.2. *Under the same hypothesis, we say that* $(f,E) \in L^1(]0,T[\times]0,1[\times\mathbb{R}_v)\times L^\infty(\mathbb{R}_t\times]0,1[)$ *is a* T*-periodic mild solution for the perturbed Vlasov–Poisson problem iff* f *is the* T*-periodic mild solution for the perturbed Vlasov problem* (2.6), (2.3) *corresponding to the regularized electric field* $E_\varepsilon(t,x) = \int_{\mathbb{R}} \zeta_\varepsilon(t-s)ds \int_0^1 \zeta_\varepsilon(x-y)E(s,y)dy$ *and* E *is given by the Poisson problem*

$$E(t,x) = \int_0^x \rho(t,y)dy - \int_0^1 (1-y)\rho(t,y)dy - \varphi_1(t) + \varphi_0(t),\ (t,x) \in \mathbb{R}_t\times]0,1[.$$

3.1. Existence for the perturbed Vlasov–Poisson problem. As a first step in the study of periodic weak solutions for the Vlasov–Poisson problem, we prove the existence for the perturbed problem. In this section the parameters $\alpha, \varepsilon > 0$ are fixed, and we use the Schauder fixed point theorem. For the moment consider that the electric field is given and let us deduce some bounds for f in the L^1 norm.

PROPOSITION 3.3. *Assume that* $E \in L^\infty(\mathbb{R}_t\times]0,1[)$, $g_0 \in L_0^-, g_1 \in L_1^-$ *are* T*-periodic and that* f *is the* T*-periodic weak solution for* (2.1), (2.3). *Then* $f \in L^\infty(\mathbb{R}_t; L^1(]0,1[\times\mathbb{R}_v))$ *and*

$$\|f\|_{L^\infty(\mathbb{R}_t;L^1(]0,1[\times\mathbb{R}_v))} \leq \left(\frac{1}{\alpha} + T\right) G_1.$$

Proof. As we saw in the previous section, $f \in L^1(]0,T[\times]0,1[\times\mathbb{R}_v)$ and $\|f\|_{L^1} \leq \frac{T}{\alpha}G_1$, $\alpha > 0$. Thus there is $t_1 \in]0,T[$ such that

$$\int_0^1\!\!\int_{\mathbb{R}_v} f(t_1,x,v)dxdv \leq \frac{G_1}{\alpha}.$$

Now by integration of the perturbated Vlasov equation on $]t_1,t[\times]0,1[\times\mathbb{R}_v$, where $t_1 \leq t \leq t_1 + T$, we find that

$$\begin{aligned}\|f(t)\|_{L^1} =& \int_0^1\!\!\int_{\mathbb{R}_v} f(t,x,v)dxdv \leq \int_0^1\!\!\int_{\mathbb{R}_v} f(t_1,x,v)dxdv + \int_{t_1}^t\!\!\int_{v>0} vg_0(s,v)dsdv\\ &- \int_{t_1}^t\!\!\int_{v<0} vg_1(s,v)dsdv \leq \left(\frac{1}{\alpha} + T\right) G_1,\end{aligned}$$

and therefore the conclusion follows by periodicity. ∎

THEOREM 3.4. *Assume that* $\varphi_1 - \varphi_0 \in L^\infty(\mathbb{R}_t), g_0, g_1$ *are* T*-periodic functions such that*

$$\begin{aligned}(H_1)\quad G_1 =& \frac{1}{T}\int_0^T\!\!\int_{v>0} vg_0(t,v)dtdv - \frac{1}{T}\int_0^T\!\!\int_{v<0} vg_1(t,v)dtdv < +\infty,\\ (H_\infty)\quad G_\infty =& \max\{\|g_0\|_{L^\infty(\mathbb{R}_t\times\Sigma_0^-)}, \|g_1\|_{L^\infty(\mathbb{R}_t\times\Sigma_1^-)}\} < +\infty.\end{aligned}$$

Then, for every $\alpha, \varepsilon > 0$, *there is a* T*-periodic mild solution for the perturbed Vlasov–Poisson problem.*

Proof. Let us define $R_\alpha = \left(T + \frac{1}{\alpha}\right) G_1 + \|\varphi_1 - \varphi_0\|_{L^\infty}$ and consider the set

$$X_{\alpha,\varepsilon} = \{E \in\ L^\infty(\mathbb{R}_t \times]0,1[) \mid E(t,x) = E(t+T,x),\ (t,x) \in \mathbb{R}_t \times]0,1[,\ \|E\|_{L^\infty} \leq R_\alpha\},$$

which is convex and compact with respect to the weak $\star$ topology of L^∞. As a fixed point application, we define $F_\alpha(E), E \in X_{\alpha,\varepsilon}$ as follows:

(3.3)
$$F_{\alpha,\varepsilon}(E)(t,x) = \int_0^x \rho(t,y)dy - \int_0^1 (1-y)\rho(t,y)dy - \varphi_1(t) + \varphi_0(t),\ (t,x) \in \mathbb{R}_t \times]0,1[,$$

where f is the mild T-periodic solution of (2.6), (2.3) corresponding to the regularized field $E_\varepsilon(t,x) = \int_{\mathbb{R}} \zeta_\varepsilon(t-s)ds \int_0^1 \zeta_\varepsilon(x-y)E(s,y)dy$. By Proposition 3.3, we deduce that ρ belongs to $L^\infty(\mathbb{R}_t; L^1(]0,1[))$, and from (3.3) it follows that $\|F_{\alpha,\varepsilon}(E)\|_{L^\infty} \leq R_\alpha,\ E \in X_{\alpha,\varepsilon}$. Since $F_{\alpha,\varepsilon}(E)$ is also T-periodic, it follows that $F_{\alpha,\varepsilon}(X_{\alpha,\varepsilon}) \subset X_{\alpha,\varepsilon}$. Now let us prove the continuity of the application $F_{\alpha,\varepsilon}$. For this, consider a sequence $(E_n)_n \subset X_{\alpha,\varepsilon}$ such that $E_n \rightharpoonup E$, weakly $\star$ in $L^\infty(]0,T[\times]0,1[)$, which implies pointwise convergence for $(t,x) \in \mathbb{R}_t \times]0,1[$:

$$\begin{aligned} E_{n,\varepsilon}(t,x) &= \int_{\mathbb{R}} \zeta_\varepsilon(t-s)ds \int_0^1 \zeta_\varepsilon(x-y)E_n(s,y)dy \\ &\to \int_{\mathbb{R}} \zeta_\varepsilon(t-s)ds \int_0^1 \zeta_\varepsilon(x-y)E(s,y)dy = E_\varepsilon(t,x). \end{aligned}$$

Thus, by the dominated convergence theorem, we obtain that $(E_{n,\varepsilon})_n$ converges strongly to E_ε in $L^2(]0,T[\times]0,1[)$ when $n \to +\infty$. Denote by $(f_n)_n$ the sequence of T-periodic mild solutions associated to $(E_{n,\varepsilon})_n$. Since $\|f_n\|_{L^\infty} \leq \|g\|_{L^\infty}$, we have, at least for a subsequence, that

$$f_n \rightharpoonup f,\ \text{weak}\ \star\ \text{in}\ \ L^\infty(]0,T[\times]0,1[\times\mathbb{R}_v).$$

As $(f_n)_n$ are mild solutions, they are also weak solutions, and therefore we have for all $\varphi \in \mathcal{T}_w,\ n$

$$\begin{aligned} &-\int_0^T\int_0^1\int_{\mathbb{R}_v} f_n(-\alpha\varphi + \partial_t\varphi + v\cdot\partial_x\varphi + E_{n,\varepsilon}(t,x)\cdot\partial_v\varphi)dtdxdv \\ &\quad = \int_0^T\int_{v>0} v g_0(t,v)\varphi(t,0,v)dtdv - \int_0^T\int_{v<0} v g_1(t,v)\varphi(t,1,v)dtdv. \end{aligned}$$

Obviously the following convergence holds:

$$\lim_{n\to+\infty} \int_0^T\int_0^1\int_{\mathbb{R}_v} f_n(-\alpha\varphi + \partial_t\varphi + v\cdot\partial_x\varphi)dtdxdv = \int_0^T\int_0^1\int_{\mathbb{R}_v} f(-\alpha\varphi + \partial_t\varphi + v\cdot\partial_x\varphi)dtdxdv.$$

In order to pass the other term to the limit, we remark that, since $(f_n)_n$ are uniformly bounded in L^∞ and φ has bounded support in velocity, we have that $\int_{\mathbb{R}_v} f_n(t,x,v)\partial_v\varphi dv$ converges to $\int_{\mathbb{R}_v} f(t,x,v)\partial_v\varphi dv$ weakly in $L^2(]0,T[\times]0,1[)$. Finally, by combining this with the strong convergence of $(E_{n,\varepsilon})_n$ in $L^2(]0,T[\times]0,1[)$ we deduce that

$$\lim_{n\to+\infty} \int_0^T\int_0^1 E_{n,\varepsilon}(t,x)\int_{\mathbb{R}_v} f_n(t,x,v)\partial_v\varphi dvdtdx = \int_0^T\int_0^1 E_\varepsilon(t,x)\int_{\mathbb{R}_v} f(t,x,v)\partial_v\varphi dvdtdx,$$

and thus f is a T-periodic weak solution for

$$\alpha f + \partial_t f + v \cdot \partial_x f + E_\varepsilon(t,x) \cdot \partial_v f = 0, (t,x,v) \in \mathbb{R}_t \times]0,1[\times \mathbb{R}_v,\ f|_{\mathbb{R}_t \times \Sigma^-} = g.$$

Since, for the perturbed Vlasov problem, we have uniqueness for the T-periodic weak solution, it follows that f is the T-periodic mild solution corresponding to E_ε. Note that it also follows from uniqueness that the whole sequence $(f_n)_n$ converges weakly $\star$ in L^∞. Let us analyze now the term $\int_0^x \rho_n(t,y)dy$. We have

$$\left| \int_0^x \rho_n(t,y)dy \right| \leq \int_0^1 \int_{\mathbb{R}_v} f_n(t,x,v)dxdv \leq \left(\frac{1}{\alpha} + T \right) G_1,\ (t,x) \in \mathbb{R}_t \times]0,1[,\ \forall n,$$

and thus $(\int_0^x \rho_n(t,y)dy)_n$ converges weakly $\star$ in $L^\infty(]0,T[\times]0,1[)$. In order to identify the weak $\star$ limit, let us calculate for $\theta \in C_c(]0,T[\times]0,1[)$

$$\int_0^T \int_0^1 \int_0^x dy \int_{\mathbb{R}_v} f_n(t,y,v)\theta(t,x)dvdtdx = \int_0^T \int_0^1 \int_{\mathbb{R}_v} f_n(t,x,v) \int_x^1 \theta(t,y)dydtdxdv$$
$$= \int_0^T \int_0^1 \int_{|v|<R} f_n \int_x^1 \theta(t,y)dydtdxdv + \int_0^T \int_0^1 \int_{|v|>R} f_n \int_x^1 \theta(t,y)dydtdxdv = \mathcal{I}_1^n(R) + \mathcal{I}_2^n(R).$$

Taking into account Remark 2.10, we deduce that $\lim_{R \to +\infty} \mathcal{I}_2^n(R) = 0$ uniformly in respect to n. On the other hand, since $\int_x^1 \theta(t,y)dy \cdot 1_{\{|v|<R\}}$ belongs to $L^1(]0,T[\times]0,1[\times\mathbb{R}_v)$, we also have the convergence

$$\lim_{n \to +\infty} \mathcal{I}_1^n(R) = \mathcal{I}_1(R) = \int_0^T \int_0^1 \int_{|v|<R} f(t,x,v) \int_x^1 \theta(t,y)dydtdxdv.$$

Finally, by combining the above convergences, one gets that

$$\int_0^x \rho_n(t,y)dy \rightharpoonup \int_0^x \rho(t,y)dy,\ \text{weak } \star \text{ in } L^\infty(]0,T[\times]0,1[),$$

where $\rho(t,x) = \int_{\mathbb{R}_v} f(t,x,v)dv$. In exactly the same manner, we find that

$$\int_0^1 (1-y)\rho_n(t,y)dy \rightharpoonup \int_0^1 (1-y)\rho(t,y)dy,\ \text{weak } \star \text{ in } L^\infty(]0,T[),$$

and therefore

$$\begin{aligned} F_{\alpha,\varepsilon}(E_n) &= \int_0^x \rho_n(t,y)dy - \int_0^1 (1-y)\rho_n(t,y)dy - \varphi_1(t) + \varphi_0(t) \\ &\rightharpoonup \int_0^x \rho(t,y)dy - \int_0^1 (1-y)\rho(t,y)dy - \varphi_1(t) + \varphi_0(t) \\ &= F_{\alpha,\varepsilon}(E),\ \text{weak } \star \text{ in } L^\infty(]0,T[\times]0,1[), \end{aligned}$$

which proves the continuity of the application $F_{\alpha,\varepsilon}$.

By using the Schauder fixed point theorem, we deduce that there is a T-periodic mild solution for the perturbed Vlasov–Poisson problem. □

4. Estimates for the perturbed T-periodic mild solutions. In order to simplify the formulas, in this section we shall systematically skip the indexes α, ε. Generally (f, E) stands for T-periodic mild solutions of the perturbed Vlasov–Poisson problem which can be written as

$$\alpha f(t,x,v) + \partial_t f + v \cdot \partial_x f + \widetilde{E}(t,x) \cdot \partial_v f = 0,\ (t,x,v) \in \mathbb{R}_t \times]0,1[\times \mathbb{R}_v,$$

$$f(t,x,v) = g(t,x,v),\ (t,x,v) \in \mathbb{R}_t \times \Sigma^-,$$

$$\widetilde{E}(t,x) = \int_{\mathbb{R}} \zeta_\varepsilon(t-s)ds \int_0^1 \zeta_\varepsilon(x-y)E(s,y)dy,\ (t,x) \in \mathbb{R}_t \times]0,1[,$$

$$E(t,x) = \int_0^x \rho(t,y)dy - \int_0^1 (1-y)\rho(t,y)dy - \varphi_1(t) + \varphi_0(t),\ (t,x) \in \mathbb{R}_t \times]0,1[.$$

As usual we use the notation

$$\rho(t,x) = \int_{\mathbb{R}_v} f(t,x,v)dv,\ \ j(t,x) = \int_{\mathbb{R}_v} v f(t,x,v)dv,\ (t,x) \in \mathbb{R}_t \times]0,1[.$$

In this section we are looking for uniform estimates of the charge, current, and electric field. It is convenient to introduce also

$$(M_\rho, M_{|j|}) := \sup_{\alpha,\varepsilon>0} \frac{1}{T} \int_0^T \int_0^1 \int_{\mathbb{R}_v} f_{\alpha,\varepsilon}(t,x,v)(1,|v|)dtdxdv,$$

$$(C_\rho, C_{|j|}) := \sup_{\alpha,\varepsilon>0, t\in\mathbb{R}_t} \int_0^1 \int_{\mathbb{R}_v} f_{\alpha,\varepsilon}(t,x,v)(1,|v|)dxdv,$$

and

$$C_E := \sup_{\alpha,\varepsilon>0} \|E_{\alpha,\varepsilon}\|_{L^\infty},$$

with $M_\rho, M_{|j|}, C_\rho, C_{|j|}, C_E \in [0, +\infty]$.

At the beginning we assume that

$$(H_0')\ \ G_0' := \int_{v>0} \sup_{t\in\mathbb{R}_t} \{g_0(t,v)\}dv + \int_{v<0} \sup_{t\in\mathbb{R}_t} \{g_1(t,v)\}dv < +\infty,$$

$$(H_1)\ \ G_1 := \frac{1}{T} \int_0^T \int_{v>0} v g_0(t,v)dtdv - \frac{1}{T} \int_0^T \int_{v<0} v g_1(t,v)dtdv < +\infty,$$

$$(H_\infty)\ \ G_\infty := \max\{\|g_0\|_{L^\infty(\mathbb{R}_t \times \Sigma_0^-)}, \|g_1\|_{L^\infty(\mathbb{R}_t \times \Sigma_1^-)}\} < +\infty,$$

which ensure the existence of the T-periodic mild solutions (see Theorem 3.4), but later on we shall see that only $(H_0'), (H_\infty)$ are sufficient.

Remember that the T-periodic mild solutions satisfy

$$\int_0^1 \int_{\mathbb{R}_v} f(t,x,v)dxdv \le \left(\frac{1}{\alpha} + T\right) G_1,\ t \in \mathbb{R}_t,$$

and

$$\|E\|_{L^\infty(\mathbb{R}_t \times]0,1[)} \le \left(\frac{1}{\alpha} + T\right) G_1 + \|\varphi_1 - \varphi_0\|_{L^\infty}.$$

4.1. Estimate of the total charge. Let us consider as a test function in the mild formulation $\psi(t,x,v) = 1_{\{|v|>R_2\}}$ with $R_2 = 6\sqrt{2}\cdot\|E\|_{L^\infty}^{1/2}$. We have

$$\begin{aligned}\int_0^T\!\!\int_0^1\!\!\int_{\mathbb{R}_v} f(t,x,v)dtdxdv &= \int_0^T\!\!\int_0^1\!\!\int_{|v|<R_2} f(t,x,v)dtdxdv + \int_0^T\!\!\int_0^1\!\!\int_{|v|>R_2} f(t,x,v)dtdxdv\\ &\le \int_0^T\!\!\int_{v>0} vg_0(t,v)dtdv \int_t^{s^0_{out}} e^{-\alpha(s-t)}1_{\{|V^0(s)|>R_2\}}ds\\ &\quad - \int_0^T\!\!\int_{v<0} vg_1(t,v)dtdv \int_t^{s^1_{out}} e^{-\alpha(s-t)}1_{\{|V^1(s)|>R_2\}}ds + 2TR_2G_\infty.\end{aligned}$$

Now, by using Lemma 2.11 we deduce that, if $R_3 = 4\sqrt{2}\cdot\|E\|_{L^\infty}^{1/2}$, we have

$$\begin{aligned}&\int_0^T\!\!\int_{v>0} vg_0(t,v)dtdv \int_t^{s^0_{out}} e^{-\alpha(s-t)}1_{\{|V^0(s)|>R_2\}}ds\\ &\quad= \int_0^T\!\!\int_{v>R_3} vg_0(t,v)dtdv \int_t^{s^0_{out}} e^{-\alpha(s-t)}1_{\{|V^0(s)|>R_2\}}ds\\ &\quad\le \int_0^T\!\!\int_{v>R_3} vg_0(t,v)\frac{1}{v-2\sqrt{2}\cdot\|E\|_{L^\infty}^{1/2}}dtdv\\ &\quad\le 2\int_0^T\!\!\int_{v>0} g_0(t,v)dtdv.\end{aligned}$$

Finally one gets that

$$\frac{1}{T}\int_0^T\!\!\int_0^1\!\!\int_{\mathbb{R}_v} f(t,x,v)dtdxdv \le 12\sqrt{2}\cdot\|E\|_{L^\infty}^{1/2}G_\infty + 2G_0 \le 12\sqrt{2}\cdot\|E\|_{L^\infty}^{1/2}G_\infty + 2G_0'.$$

We need also to estimate f in $L^\infty(\mathbb{R}_t; L^1(]0,1[\times\mathbb{R}_v))$. First, notice that from the previous estimate it follows that there is $t_1 \in]0,T[$ such that

$$\int_0^1\!\!\int_{\mathbb{R}_v} f(t_1,x,v)dxdv \le \frac{1}{T}\int_0^T\!\!\int_0^1\!\!\int_{\mathbb{R}_v} f(t,x,v)dtdxdv \le 12\sqrt{2}\cdot\|E\|_{L^\infty}^{1/2}G_\infty + 2G_0'.$$

On the other hand, by integration of the perturbed Vlasov equation on $]t_1,t[\times]0,1[\times\mathbb{R}_v, t_1 \le t \le t_1+T$, we have

$$\begin{aligned}e^{\alpha t}\int_0^1\!\!\int_{\mathbb{R}_v} f(t,x,v)dxdv &= e^{\alpha t_1}\int_0^1\!\!\int_{\mathbb{R}_v} f(t_1,x,v)dxdv + \int_{t_1}^t e^{\alpha\tau}\int_{\mathbb{R}_v} v(f(\tau,0,v)-f(\tau,1,v))d\tau dv\\ &\le e^{\alpha t_1}(12\sqrt{2}\cdot\|E\|_{L^\infty}^{1/2}G_\infty + 2G_0') + \int_{t_1}^t e^{\alpha\tau}\int_{|v|>R_2} v(f(\tau,0,v)-f(\tau,1,v))d\tau dv\\ &\quad + \int_{t_1}^t e^{\alpha\tau}\int_{0<v<R_2} vg_0(\tau,v)d\tau dv - \int_{t_1}^t e^{\alpha\tau}\int_{0>v>-R_2} vg_1(\tau,v)d\tau dv. \qquad (4.1)\end{aligned}$$

We need to estimate the integral $\mathcal{I}(t_1,t_2) = \int_{t_1}^{t_2} e^{\alpha\tau}\int_{|v|>R_2} v(f(\tau,0,v)-f(\tau,1,v))d\tau dv$ for $0 \le t_2 - t_1 \le T$. We shall consider the applications

$$F_0 : \mathbb{R}_t \times [R_3, +\infty[\to \mathbb{R}^2,\ F_0(t,v) = (s_{out}(t,0,v), V(s_{out}(t,0,v);t,0,v)),$$

and

$$F_1 : \mathbb{R}_t \times]-\infty, -R_3] \to \mathbb{R}^2,\ F_1(t,v) = (s_{out}(t,1,v), V(s_{out}(t,1,v);t,1,v)).$$

By again using Lemma 2.11, it is clear that F_0, F_1 are well defined, and we have

$$s^0_{out} \leq \frac{1}{v - 2\sqrt{2}\cdot \|E\|^{1/2}_{L^\infty}} \leq \frac{1}{2\sqrt{2}\cdot \|E\|^{1/2}_{L^\infty}},\ v \geq R_3,\ \ X(s^0_{out};t,0,v) = 1,$$

$$s^1_{out} \leq \frac{1}{-v - 2\sqrt{2}\cdot \|E\|^{1/2}_{L^\infty}} \leq \frac{1}{2\sqrt{2}\cdot \|E\|^{1/2}_{L^\infty}},\ v \leq -R_3,\ \ X(s^1_{out};t,1,v) = 0.$$

Moreover F_0, F_1 are one-to-one maps since $\|\widetilde{E}\|_{L^\infty(\mathbb{R}_t;W^{1,\infty}(]0,1[))} \leq \|E\|_{L^\infty}(1 + \int_{\mathbb{R}} |\zeta'(u)|du/\varepsilon)$ and therefore the uniqueness of the characteristics holds. By standard calculations we get

$$\left|\frac{\partial F_0}{\partial(t,v)}\right| = \frac{v}{V^0(s^0_{out})} \in \left[\frac{2}{3}, 2\right],\ (t,v) \in \mathbb{R}_t \times [R_3, +\infty[,$$

$$\left|\frac{\partial F_1}{\partial(t,v)}\right| = \frac{-v}{V^1(s^1_{out})} \in \left[\frac{2}{3}, 2\right],\ (t,v) \in \mathbb{R}_t \times]-\infty, -R_3].$$

We have

$$\begin{aligned}\mathcal{I}(t_1,t_2) =& \int_{t_1}^{t_2} e^{\alpha t}\int_{v>R_2} v g_0(t,v)dtdv - \int_{t_1}^{t_2} e^{\alpha\tau}\int_{u>R_2} u f(\tau,1,u)d\tau du\\ &-\int_{t_1}^{t_2} e^{\alpha t}\int_{v<-R_2} v g_1(t,v)dtdv + \int_{t_1}^{t_2} e^{\alpha\tau}\int_{u<-R_2} u f(\tau,0,u)d\tau du\\ &= \mathcal{I}^+(t_1,t_2) + \mathcal{I}^-(t_1,t_2).\end{aligned}$$

However, with the change of variables $(\tau,u) = (s_{out}(t,0,v), V(s_{out}(t,0,v);t,0,v)) = F_0(t,v)$, we have

$$\int_{t_1}^{t_2} e^{\alpha\tau}\int_{u>R_2} u f(\tau,1,u)d\tau du = \iint_{F_0^{-1}(]t_1,t_2[\times]R_2,+\infty[)} e^{\alpha t} v g_0(t,v)dtdv.$$

If we denote $R_4 = \max\{8\sqrt{2}\cdot\|E\|^{1/2}_{L^\infty}, 2\sqrt{2}\cdot\|E\|^{1/2}_{L^\infty} + 1/(t_2-t_1)\}$, we can easily check that $\cup_{v\geq R_4}([t_1, t_2 - \delta(v)] \times \{v\}) \subset F_0^{-1}([t_1,t_2]\times[R_2,+\infty[)$ with $\delta(v) = 1/(v - 2\sqrt{2}\cdot \|E\|^{1/2}_{L^\infty})$. Therefore, by taking into account that $v\cdot\delta(v) \leq 4/3$ for $v \geq R_4$, we get

$$\begin{aligned}\mathcal{I}^+(t_1,t_2) \leq& \int_{t_1}^{t_2} e^{\alpha t}\int_{R_2<v<R_4} v g_0(t,v)dtdv + \int_{v>R_4}\int_{t_2-\delta(v)}^{t_2} e^{\alpha t} v g_0(t,v)dvdt\\ \leq& e^{\alpha t_2}\left(\int_{t_1}^{t_2}\int_{R_2<v<R_4} v g_0(t,v)dtdv + \frac{4}{3}\int_{v>0}\sup_{t\in\mathbb{R}_t}\{g_0(t,v)\}dv\right).\end{aligned}$$

On the other hand, since $R_4 \leq 8\sqrt{2}\cdot\|E\|^{1/2}_{L^\infty} + 1/(t_2-t_1) = R_5 + 1/(t_2-t_1)$, we

also have

$$\begin{aligned}\int_{t_1}^{t_2}\int_{R_2<v<R_4} vg_0(t,v)dtdv &= \int_{t_1}^{t_2}\int_{R_2<v<R_5} vg_0(t,v)dtdv + \int_{t_1}^{t_2}\int_{R_5<v<R_5+1/(t_2-t_1)} vg_0(t,v)dtdv\\ &\le R_5\int_0^T\int_{v>0} g_0(t,v)dtdv + \left(R_5+\frac{1}{t_2-t_1}\right)\int_{t_1}^{t_2}\int_{v>0} g_0(t,v)dtdv\\ &\le 16\sqrt{2}\cdot\|E\|_{L^\infty}^{1/2}\int_0^T\int_{v>0} g_0(t,v)dtdv + \int_{v>0}\sup_{t\in\mathbb{R}_t}\{g_0(t,v)\}dv.\end{aligned}$$

The right boundary term $\mathcal{I}^-(t_1,t_2)$ can be estimated in the same manner, and finally one gets

$$\mathcal{I}(t_1,t_2)\le e^{\alpha t_2}\left(16\sqrt{2}\cdot\|E\|_{L^\infty}^{1/2}TG_0+\frac{7}{3}G_0'\right),$$

and therefore we deduce from (4.1) that

$$\int_0^1\int_{\mathbb{R}_v} f(t,x,v)dxdv \le (12\cdot G_\infty + 22\cdot TG_0)\sqrt{2}\cdot\|E\|_{L^\infty}^{1/2}+\frac{13}{3}G_0',\ t\in\mathbb{R}_t.$$

From the Poisson equation we deduce that

$$\|E(t)\|_{L^\infty(]0,1[)}\le|\varphi_1(t)-\varphi_0(t)|+\int_0^1\int_{\mathbb{R}_v} f(t,x,v)dxdv,\ t\in\mathbb{R}_t,$$

which combined with the previous inequality implies that $\|E\|_{L^\infty}\le A\cdot\|E\|_{L^\infty}^{1/2}+B$, with $A=12\cdot 2^{1/2}G_\infty+22\cdot 2^{1/2}TG_0,\ B=\|\varphi_1-\varphi_0\|_{L^\infty}+\frac{13}{3}G_0'$, and therefore

$$\begin{aligned}\|E\|_{L^\infty(\mathbb{R}_t\times]0,1[)}&\le A^2+2B,\\ \|\rho\|_{L^\infty(\mathbb{R}_t;L^1(]0,1[))}&\le A(A+B^{1/2})+\frac{13}{3}G_0',\\ \frac{1}{T}\|f\|_{L^1(]0,T[\times]0,1[\times\mathbb{R}_v)}&\le 12\sqrt{2}(A+B^{1/2})G_\infty+2G_0',\end{aligned}$$

which can be written

(4.2)
$$M_\rho\le 12\cdot 2^{1/2}(A+B^{1/2})G_\infty+2G_0',\ \ C_\rho\le A(A+B^{1/2})+\frac{13}{3}G_0',\ \ C_E\le A^2+2B.$$

4.2. Estimate of the rest of charge ($\int_{|v|>R} f dv$). We shall also need to estimate integrals like $\int_0^T\int_0^1\int_{|v|>R} f(t,x,v)dtdxdv$ or $\int_0^1\int_{|v|>R} f(t,x,v)dxdv,\ t\in\mathbb{R}_t$. In fact, since we know that $C_E<+\infty$, we have, by taking $\psi=1_{\{|v|>R\}}$ as a test

function in the mild formulation, with R large enough, $R_6 = R - 2\sqrt{2}\cdot C_E^{1/2}$,

$$\begin{aligned}
\int_0^T\!\!\int_0^1\!\!\int_{|v|>R} f dt dx dv &= \int_0^T\!\!\int_{v>0} v g_0(t,v) dt dv \int_t^{s_{out}^0} e^{-\alpha(s-t)} 1_{\{|V^0(s)|>R\}} ds \\
&\quad - \int_0^T\!\!\int_{v<0} v g_1(t,v) dt dv \int_t^{s_{out}^1} e^{-\alpha(s-t)} 1_{\{|V^1(s)|>R\}} ds \\
&\le \int_0^T\!\!\int_{v>R_6} v g_0(t,v)(s_{out}(t,0,v)-t) dt dv - \int_0^T\!\!\int_{v<-R_6} v g_1(t,v)(s_{out}(t,1,v)-t) dt dv \\
&\le \int_0^T\!\!\int_{v>R_6} v g_0(t,v) \frac{1}{v - 2\sqrt{2}\cdot C_E^{1/2}} dt dv - \int_0^T\!\!\int_{v<-R_6} v g_1(t,v) \frac{1}{-v - 2\sqrt{2}\cdot C_E^{1/2}} dt dv \\
&\le \frac{R - 2\sqrt{2}\cdot C_E^{1/2}}{R - 4\sqrt{2}\cdot C_E^{1/2}} \left(\int_0^T\!\!\int_{v>R_6} g_0(t,v) dt dv - \int_0^T\!\!\int_{v<-R_6} g_1(t,v) dt dv \right), \qquad (4.3)
\end{aligned}$$

and thus

$$\lim_{R\to+\infty} \int_0^T\!\!\int_0^1\!\!\int_{|v|>R} f(t,x,v) dt dx dv = 0,$$

uniformly in respect to $\alpha, \varepsilon > 0$. Moreover, in order to estimate $\int_0^1\!\int_{|v|>R} f(t,x,v) dx dv$, let us remark that there is $t_1 \in]0,T[$ such that

$$\begin{aligned}
\int_0^1\!\!\int_{|v|>R} f(t_1,x,v) dx dv &\le \frac{1}{T} \int_0^T\!\!\int_0^1\!\!\int_{|v|>R} f(t,x,v) dt dx dv \\
&\le \frac{R - 2\sqrt{2}\cdot C_E^{1/2}}{R - 4\sqrt{2}\cdot C_E^{1/2}} \cdot \frac{1}{T} \left(\int_0^T\!\!\int_{v>R_6} g_0(t,v) dt dv - \int_0^T\!\!\int_{v<-R_6} g_1(t,v) dt dv \right). \qquad (4.4)
\end{aligned}$$

After multiplication of the perturbed Vlasov equation by $1 - \chi_R(v) = 1 - \chi(v/R)$, we have

$$\begin{aligned}
&\partial_t(e^{\alpha t} f(1-\chi_R(v))) + v\cdot\partial_x(e^{\alpha t} f(1-\chi_R(v))) + \widetilde{E}(t,x)\cdot\partial_v(e^{\alpha t} f(1-\chi_R(v))) \\
&\quad = -e^{\alpha t}\widetilde{E} f \chi'(v/R)\frac{1}{R},
\end{aligned}$$

and after integration on $]t_1, t[\times]0,1[\times\mathbb{R}_v$ one gets

$$\begin{aligned}
e^{\alpha t}\int_0^1\!\!\int_{|v|>2R} f(t,x,v) dx dv &\le e^{\alpha t_1}\int_0^1\!\!\int_{|v|>R} f(t_1,x,v) dx dv \\
&\quad + \int_{t_1}^t e^{\alpha\tau}\int_{|v|>R} v(f(\tau,0,v) - f(\tau,1,v))(1-\chi_R(v)) d\tau dv \\
&\quad - \int_{t_1}^t e^{\alpha\tau}\int_0^1\int_{R<|v|<2R} \widetilde{E} f(\tau,x,v)\chi'(v/R)\frac{1}{R} d\tau dx dv,\ t\in[t_1,t_1+T]. \qquad (4.5)
\end{aligned}$$

The first term in the right-hand side of the previous inequality can be estimated by using (4.4). For the third one, we have

$$\left| \int_{t_1}^t e^{\alpha\tau}\int_0^1\int_{R<|v|<2R} \widetilde{E} f(\tau,x,v)\chi'(v/R)\frac{1}{R} d\tau dx dv \right| \le e^{\alpha t} C_E \|\chi'\|_{L^\infty} \frac{T M_\rho}{R} \to 0 \qquad (4.6)$$

when R goes to $+\infty$, uniformly in $\alpha, \varepsilon > 0$. In order to estimate integrals like $\mathcal{I}_R(t_1,t_2) = \int_{t_1}^{t_2} e^{\alpha\tau} \int_{|v|>R} v(f(\tau,0,v) - f(\tau,1,v))(1-\chi_R(v))d\tau dv$ as before, we remark that

$$\begin{aligned}\mathcal{I}_R(t_1,t_2) &= \int_{t_1}^{t_2} e^{\alpha t} \int_{v>R} v g_0(t,v)(1-\chi_R(v))dtdv - \int_{t_1}^{t_2} e^{\alpha\tau} \int_{u>R} u f(\tau,1,u)(1-\chi_R(u))d\tau du \\ &\quad - \int_{t_1}^{t_2} e^{\alpha t} \int_{v<-R} v g_1(t,v)(1-\chi_R(v))dtdv + \int_{t_1}^{t_2} e^{\alpha\tau} \int_{u<-R} u f(\tau,0,u)(1-\chi_R(u))d\tau du \\ &= \mathcal{I}_R^+(t_1,t_2) + \mathcal{I}_R^-(t_1,t_2).\end{aligned}$$

Taking into account that, for R large enough such that $\delta(R) \le t_2 - t_1$ and $\eta = C_E \cdot \delta(R)$, we have

$$\bigcup_{v \ge R+\eta} ([t_1, t_2 - \delta(v)] \times \{v\}) \subset F_0^{-1}([t_1,t_2] \times [R, +\infty[),$$

where $\delta(v) = 1/(v - 2\sqrt{2} \cdot C_E^{1/2})$; by the same change of variables it follows that

$$\begin{aligned}\mathcal{I}_R^+(t_1,t_2) &\le e^{\alpha t_2} \int_{t_1}^{t_2} \int_{R<v<R+\eta} v g_0(t,v)(1-\chi_R(v))dtdv \\ &\quad + e^{\alpha t_2} \int_{t_2-\delta(v)}^{t_2} \int_{v>R+\eta} v g_0(t,v)(\chi_R(V(s_{out}^0;t,0,v)) - \chi_R(v))dtdv.\end{aligned}$$

However, for $R < v < R+\eta$, we have $1 - \chi_R(v) = |\chi_R(R) - \chi_R(v)| \le \frac{\eta}{R}\|\chi'\|_{L^\infty}$. We also have

$$|\chi_R(V(s_{out}^0;t,0,v)) - \chi_R(v)| \le \frac{|V^0(s_{out}^0) - v|}{R}\|\chi'\|_{L^\infty} \le \frac{2\|\chi'\|_{L^\infty}\sqrt{2} \cdot C_E^{1/2}}{R},$$

and thus

$$\begin{aligned}\mathcal{I}_R^+(t_1,t_2) &\le e^{\alpha t_2} \frac{\eta(R+\eta)}{R}\|\chi'\|_{L^\infty} \int_{t_1}^{t_2} \int_{v>R} g_0(t,v)dtdv \\ &\quad + \frac{e^{\alpha t_2}}{R} \int_{v>R} v\delta(v) \sup_{t\in\mathbb{R}_t}\{g_0(t,v)\}\|\chi'\|_{L^\infty} 2\sqrt{2} \cdot C_E^{1/2} \\ &\le e^{\alpha t_2}\|\chi'\|_{L^\infty} const(C_E)\delta(R) \left(\int_0^T \int_{v>0} g_0(t,v)dtdv + \int_{v>0} \sup_{t\in\mathbb{R}_t}\{g_0(t,v)\}dv \right) \qquad (4.7)\end{aligned}$$

for $R \ge 2\sqrt{2} \cdot C_E^{1/2} + 1/(t_2 - t_1)$. The same arguments apply for the right boundary term $\mathcal{I}_R^-(t_1,t_2)$, and therefore we have

$$\mathcal{I}_R(t_1,t_2) = \mathcal{I}_R^+(t_1,t_2) + \mathcal{I}_R^-(t_1,t_2) \le e^{\alpha t_2}\|\chi'\|_{L^\infty} \mathrm{const}(C_E)\delta(R)(TG_0 + G_0'), \qquad (4.8)$$

where $R \ge 2\sqrt{2} \cdot C_E^{1/2} + 1/(t_2 - t_1)$, $0 \le t_2 - t_1 \le T$. Finally, by using (4.5), (4.4), (4.6), and (4.8), we find that

$$\begin{aligned}\int_0^1 \int_{|v|>2R} f(t,x,v)dxdv &\le \frac{R - 2\sqrt{2} \cdot C_E^{1/2}}{R - 4\sqrt{2} \cdot C_E^{1/2}} \cdot \frac{1}{T} \left(\int_0^T \int_{v>R_6} g_0(t,v)dtdv - \int_0^T \int_{v<-R_6} g_1(t,v)dtdv \right) \\ &\quad + \frac{1}{R} C_E\|\chi'\|_{L^\infty} T M_\rho + \delta(R)\|\chi'\|_{L^\infty} \mathrm{const}(C_E)(TG_0 + G_0'), \qquad (4.9)\end{aligned}$$

which implies that $\int_0^1 \int_{|v|>2R} f(t,x,v)dxdv \to 0$ when $R \to +\infty$, uniformly for $\alpha, \varepsilon > 0$ and $|t_2 - t_1| \geq \beta > 0$. By periodicity, we deduce that the convergence is uniform for $\alpha, \varepsilon > 0,\ t \in \mathbb{R}_t$. Notice that none of these estimates require any information about G_1. As we shall see, hypothesis (H_1) is not necessary for existence. In conclusion we prove the following proposition.

PROPOSITION 4.1. *Assume that $g_0, g_1, \varphi_0, \varphi_1$ are T-periodic functions satisfying $\varphi_1 - \varphi_0 \in L^\infty(\mathbb{R}_t)$ and hypotheses (H_0'), (H_1), and (H_∞). Denote by $(f_{\alpha,\varepsilon}, E_{\alpha,\varepsilon})$ T-periodic mild solutions of the perturbed Vlasov–Poisson problem with $\alpha > 0, \varepsilon > 0$ (see Theorem 3.4). Then the following estimates hold uniformly in respect to $\alpha > 0, \varepsilon > 0$:*

$$\|f_{\alpha,\varepsilon}\|_{L^1(]0,T[\times]0,1[\times\mathbb{R}_v)} \leq C,$$

$$\|f_{\alpha,\varepsilon}\|_{L^\infty(\mathbb{R}_t;L^1(]0,1[\times\mathbb{R}_v))} = \|\rho_{\alpha,\varepsilon}\|_{L^\infty(\mathbb{R}_t;L^1(]0,1[))} \leq C,$$

$$\|E_{\alpha,\varepsilon}\|_{L^\infty(\mathbb{R}_t\times]0,1[)} \leq C,$$

where the constant C depends only on $T, \|\varphi_1 - \varphi_0\|_{L^\infty(\mathbb{R}_t)}, G_0', G_\infty$ (and not on G_1). Moreover the following convergences hold:

$$\lim_{R\to+\infty} \int_0^T \int_0^1 \int_{|v|>R} f_{\alpha,\varepsilon}(t,x,v)dtdxdv = 0, \text{ uniformly with respect to } \alpha>0, \varepsilon>0, G_1,$$

$$\lim_{R\to+\infty} \int_0^1 \int_{|v|>R} f_{\alpha,\varepsilon}(t,x,v)dxdv = 0, \text{ uniformly with respect to } \alpha>0, \varepsilon>0, t\in\mathbb{R}_t, G_1.$$

5. Existence for the Vlasov–Poisson problem. Now we can prove the following existence result.

THEOREM 5.1. *Assume that $\varphi_1 - \varphi_0 \in L^\infty(\mathbb{R}_t)$ and g_0, g_1 are T-periodic functions such that*

$$(H_0')\ \ G_0' := \int_{v>0} \sup_{t\in\mathbb{R}_t}\{g_0(t,v)\}dv + \int_{v<0} \sup_{t\in\mathbb{R}_t}\{g_1(t,v)\}dv < +\infty,$$
$$(H_\infty)\ \ G_\infty := \max\{\|g_0\|_{L^\infty(\mathbb{R}_t\times\Sigma_0^-)}, \|g_1\|_{L^\infty(\mathbb{R}_t\times\Sigma_1^-)}\} < +\infty.$$

Then there is a T-periodic weak solution (f, E) for the Vlasov–Poisson problem such that

$$f \in L^1(]0,T[\times]0,1[\times\mathbb{R}_v),\ \rho \in L^\infty(\mathbb{R}_t; L^1(]0,1[)),\ E \in L^\infty(\mathbb{R}_t\times]0,1[).$$

Proof. For $\alpha > 0$ we consider the perturbed boundary data defined by

$$g_0^\alpha(t,v) = \frac{g_0(t,v)}{1+\alpha v},\ t \in \mathbb{R}_t, v > 0,$$

and

$$g_1^\alpha(t,v) = \frac{g_1(t,v)}{1-\alpha v},\ t \in \mathbb{R}_t, v < 0.$$

We have for $\alpha > 0$

$$G_0^{\alpha\prime} := \int_{v>0} \sup_{t\in\mathbb{R}_t}\{g_0^\alpha(t,v)\}dv + \int_{v<0} \sup_{t\in\mathbb{R}_t}\{g_1^\alpha(t,v)\}dv \leq G_0' < +\infty,$$

$$G_1^\alpha := \frac{1}{T}\int_0^T\int_{v>0} v g_0^\alpha(t,v)dtdv - \frac{1}{T}\int_0^T\int_{v<0} v g_1^\alpha(t,v)dtdv \leq \frac{1}{\alpha}G_0 \leq \frac{1}{\alpha}G_0' < +\infty,$$

and

$$G_\infty^\alpha := \max\{\|g_0^\alpha\|_{L^\infty(\mathbb{R}_t\times\Sigma_0^-)}, \|g_1^\alpha\|_{L^\infty(\mathbb{R}_t\times\Sigma_1^-)}\} \le G_\infty,$$

and therefore there is a T-periodic mild solution for the perturbed Vlasov–Poisson problem with $\alpha = \varepsilon > 0$. Moreover, since $G_0^{\alpha\prime} \le G_0'$, $G_\infty^\alpha \le G_\infty$, we have the following estimates for $\alpha > 0$:

$$\frac{1}{T}\|f_\alpha\|_{L^1(]0,T[\times]0,1[\times\mathbb{R}_v)} = \frac{1}{T}\int_0^T\!\!\int_0^1\!\!\int_{\mathbb{R}_v} f_\alpha(t,x,v)dtdxdv \le M_\rho,$$

$$\|\rho_\alpha\|_{L^\infty(\mathbb{R}_t;L^1(]0,1[))} = \sup_{t\in\mathbb{R}_t}\int_0^1\!\!\int_{\mathbb{R}_v} f_\alpha(t,x,v)dxdv \le C_\rho,$$

$$\|f_\alpha\|_{L^\infty} \le G_\infty, \quad \|E_\alpha\|_{L^\infty(\mathbb{R}_t\times]0,1[)} \le C_E,$$

where M_ρ, C_ρ, C_E verify the inequalities (4.2). Therefore there are $f \in L^\infty(\mathbb{R}_t\times]0, 1[\times\mathbb{R}_v), E \in L^\infty(\mathbb{R}_t\times]0,1[)$ T-periodic functions such that

$$\begin{aligned} &f_{\alpha_n} \rightharpoonup f, \text{ weak } \star \text{ in } L^\infty(\mathbb{R}_t\times]0,1[\times\mathbb{R}_v),\\ &E_{\alpha_n} \rightharpoonup E, \text{ weak } \star \text{ in } L^\infty(\mathbb{R}_t\times]0,1[), \end{aligned}$$

where $\alpha_n \to 0$ when $n \to +\infty$. Moreover we can easily check that we also have the convergence $\widetilde{E}_{\alpha_n} \rightharpoonup E$ weakly $\star$ in $L^\infty(\mathbb{R}_t\times]0,1[)$ when $n \to +\infty$. Since f_{α_n} are mild solutions, they are also weak solutions and thus

$$\begin{aligned} &-\int_0^T\!\!\int_0^1\!\!\int_{\mathbb{R}_v} f_{\alpha_n}(t,x,v)(-\alpha_n\varphi(t,x,v) + \partial_t\varphi + v\cdot\partial_x\varphi + \widetilde{E}_{\alpha_n}(t,x)\cdot\partial_v\varphi)dtdxdv\\ &= \int_0^T\!\!\int_{v>0} vg_0^{\alpha_n}(t,v)\varphi(t,0,v)dtdv - \int_0^T\!\!\int_{v<0} vg_1^{\alpha_n}(t,v)\varphi(t,1,v)dtdv \end{aligned}$$

$\forall\varphi \in \mathcal{T}_w$. Obviously we have

$$\begin{aligned} &\lim_{n\to+\infty}\int_0^T\!\!\int_0^1\!\!\int_{\mathbb{R}_v} f_{\alpha_n}(t,x,v)(-\alpha_n\varphi + \partial_t\varphi + v\cdot\partial_x\varphi)dtdxdv\\ &\qquad = \int_0^T\!\!\int_0^1\!\!\int_{\mathbb{R}_v} f(t,x,v)(\partial_t\varphi + v\cdot\partial_x\varphi)dtdxdv. \end{aligned}$$

On the other hand, since φ has bounded support in velocity, by the dominated convergence theorem, we deduce that

$$\begin{aligned} &\lim_{n\to+\infty}\ \int_0^T\!\!\int_{v>0} vg_0^{\alpha_n}(t,v)\varphi(t,0,v)dtdv - \int_0^T\!\!\int_{v<0} vg_1^{\alpha_n}(t,v)\varphi(t,1,v)dtdv\\ &\qquad = \int_0^T\!\!\int_{v>0} vg_0(t,v)\varphi(t,0,v)dtdv - \int_0^T\!\!\int_{v<0} vg_1(t,v)\varphi(t,1,v)dtdv. \end{aligned}$$

In order to pass the other term to the limit, we shall prove that $(E_{\alpha_n}(t))_n$ is relatively compact in $L^1(]0,1[)$, $t \in \mathbb{R}_t$ (see [9, p. 73] for compactness results in L^1). Indeed,

first, $(E_{\alpha_n}(t))_n$ is bounded in $L^\infty(]0,1[)$ and thus in $L^1(]0,1[)$. Moreover it is clear that $\forall \varepsilon > 0$ there is ω open set such that $\overline{\omega} \subset]0,1[$ and $\int_{]0,1[-\omega} |E_{\alpha_n}(t,x)|dx < \varepsilon \ \forall n$. Let us consider now $\varepsilon > 0, \omega =]x_1, x_2[\subset]0,1[$, and $|h| < \min\{x_1, 1-x_2\}$. We have

$$\begin{aligned}\int_{x_1}^{x_2} |E_{\alpha_n}(t,x+h) - E_{\alpha_n}(t,x)|dx &\leq \int_{x_1}^{x_2} \left| \int_x^{x+h} \rho_{\alpha_n}(t,y)dy \right| dx \\ &\leq |h| \int_0^1 \rho_{\alpha_n}(t,x)dx \\ &\leq C_\rho |h| \to 0, \ h \to 0,\end{aligned}$$

which implies that $(E_{\alpha_n}(t))_n$ is relatively compact in $L^1(]0,1[)$. Since $E_{\alpha_n}(t) \rightharpoonup E(t)$ weakly $\star$ in $L^\infty(]0,1[)$, we deduce that all of the sequence $(E_{\alpha_n}(t))_n$ converges to $E(t)$ in $L^1(]0,1[)$, $t \in \mathbb{R}_t$, and by the dominated convergence theorem it follows that $E_{\alpha_n} \to E$ strongly in $L^1(]0,T[\times]0,1[)$. Now, since φ has bounded support in velocity, we can write

$$\begin{aligned}&\left| \int_0^T \int_0^1 E_{\alpha_n}(t,x) \int_{\mathbb{R}_v} f_{\alpha_n}(t,x,v)\partial_v \varphi dv dt dx - \int_0^T \int_0^1 E(t,x) \int_{\mathbb{R}_v} f(t,x,v)\partial_v \varphi dv dt dx \right| \\ &\leq \left| \int_0^T \int_0^1 (E_{\alpha_n} - E) \int_{\mathbb{R}_v} f_{\alpha_n}(t,x,v)\partial_v \varphi dv dt dx \right| \\ &+ \left| \int_0^T \int_0^1 \int_{\mathbb{R}_v} (f_{\alpha_n}(t,x,v) - f(t,x,v)) E \partial_v \varphi dt dx dv \right| \to 0,\end{aligned}$$

and thus f is a T-periodic weak solution for the Vlasov problem corresponding to the field E. Moreover, since $f_{\alpha_n} \rightharpoonup f$ weakly $\star$ in $L^\infty(\mathbb{R}_t \times]0,1[\times \mathbb{R}_v)$, we have that $f_{\alpha_n} \rightharpoonup f$ weakly in $L^1(]0,T[\times]0,1[\times B_R), \ R > 0$, and therefore

$$\begin{aligned}\frac{1}{T} \int_0^T \int_0^1 \int_{|v|<R} f(t,x,v)dt dx dv &\leq \frac{1}{T} \liminf_{n \to +\infty} \int_0^T \int_0^1 \int_{|v|<R} f_{\alpha_n}(t,x,v)dt dx dv \\ &\leq \frac{1}{T} \liminf_{n \to +\infty} \int_0^T \int_0^1 \int_{\mathbb{R}_v} f_{\alpha_n} dt dx dv \leq M_\rho, \ R > 0,\end{aligned}$$

which implies that $f \in L^1(]0,T[\times]0,1[\times\mathbb{R}_v)$ and $\frac{1}{T}\int_0^T \int_0^1 \int_{\mathbb{R}_v} f(t,x,v)dt dx dv \leq M_\rho$. We can prove that $f_{\alpha_n} \rightharpoonup f$ weakly in $L^1(]0,T[\times]0,1[\times\mathbb{R}_v)$. Indeed, for $\theta \in L^\infty(]0,T[\times]0,1[\times\mathbb{R}_v)$, we can write

$$\begin{aligned}\left| \int_0^T \int_0^1 \int_{\mathbb{R}_v} (f_{\alpha_n} - f)\theta(t,x,v)dt dx dv \right| &\leq \left| \int_0^T \int_0^1 \int_{|v|<R} (f_{\alpha_n} - f)\theta(t,x,v)dt dx dv \right| \\ &+ \|\theta\|_{L^\infty} \left(\int_0^T \int_0^1 \int_{|v|>R} f_{\alpha_n} dt dx dv + \int_0^T \int_0^1 \int_{|v|>R} f dt dx dv \right).\end{aligned}$$

From (4.3) it follows that we can take $R = R(\varepsilon)$ large enough such that

$$\int_0^T \int_0^1 \int_{|v|>R} f_{\alpha_n}(t,x,v)dt dx dv \leq \frac{\varepsilon}{4\|\theta\|_{L^\infty}}, \ n > 0,$$

$$\int_0^T \int_0^1 \int_{|v|>R} f(t,x,v)dt dx dv \leq \frac{\varepsilon}{4\|\theta\|_{L^\infty}},$$

and, since $f_{\alpha_n} \rightharpoonup f$ weakly $\star$ in $L^\infty(]0,T[\times]0,1[\times\mathbb{R}_v)$, we also have

$$\left|\int_0^T\int_0^1\int_{|v|<R}(f_{\alpha_n}-f)\theta(t,x,v)dtdxdv\right|<\frac{\varepsilon}{2},\ n\geq n_\varepsilon,$$

which ensures the weak convergence of (f_{α_n}) in L^1. In particular $\rho_{\alpha_n}\rightharpoonup\rho$ weakly in $L^1(]0,T[\times]0,1[)$. Now, for all $t\in\mathbb{R}_t$, we also have the convergence $f_{\alpha_n}(t)\rightharpoonup f(t)$ weakly $\star$ in $L^\infty(]0,1[\times\mathbb{R}_v)$. In particular we have $f_{\alpha_n}(t)\rightharpoonup f(t)$ weakly in $L^1(]0,1[\times B_R),\ R>0$, and therefore

$$\begin{aligned}\int_0^1\int_{|v|<R}f(t,x,v)dxdv&\leq\liminf_{n\to+\infty}\int_0^1\int_{|v|<R}f_{\alpha_n}(t,x,v)dxdv\\&\leq\liminf_{n\to+\infty}\int_0^1\int_{\mathbb{R}_v}f_{\alpha_n}(t,x,v)dxdv\leq C_\rho,\ t\in\mathbb{R}_t,\end{aligned}$$

which implies that $f(t)\in L^1(]0,1[\times\mathbb{R}_v)$ and $\|f\|_{L^\infty(\mathbb{R}_t;L^1(]0,1[\times\mathbb{R}_v))}=\|\rho\|_{L^\infty(\mathbb{R}_t;L^1(]0,1[))}\leq C_\rho$. By using (4.9), we can prove that $f_{\alpha_n}(t)\rightharpoonup f(t)$ weakly in $L^1(]0,1[\times\mathbb{R}_v)$, $t\in\mathbb{R}_t$. We also have the convergence $\rho_{\alpha_n}(t)\rightharpoonup\rho(t)$ weakly in $L^1(]0,1[)$ for all $t\in\mathbb{R}_t$ and therefore

$$\int_0^x\rho_{\alpha_n}(t,y)dy\to\int_0^x\rho(t,y)dy,\ (t,x)\in\mathbb{R}_t\times[0,1],$$

and

$$\int_0^1(1-y)\rho_{\alpha_n}(t,y)dy\to\int_0^1(1-y)\rho(t,y)dy,\ t\in\mathbb{R}_t.$$

Now, by using the Poisson equation, we deduce that there is E_1 such that

$$\begin{aligned}E_1(t,x)&=\lim_{n\to+\infty}E_{\alpha_n}(t,x)\\&=\lim_{n\to+\infty}\left(\int_0^x\rho_{\alpha_n}(t,y)dy-\int_0^1(1-y)\rho_{\alpha_n}(t,y)dy-\varphi_1(t)+\varphi_0(t)\right)\\&=\int_0^x\rho(t,y)dy-\int_0^1(1-y)\rho(t,y)dy-\varphi_1(t)+\varphi_0(t),\ (t,x)\in\mathbb{R}_t\times[0,1],\end{aligned}$$

with $\|E_1\|_{L^\infty}\leq C_E$, which also implies that $E_{\alpha_n}\to E_1$ in $L^1(]0,T[\times]0,1[)$, and therefore the field $E=E_1$ also verifies the Poisson equation:

$$E(t,x)=\int_0^x\rho(t,y)dy-\int_0^1(1-y)\rho(t,y)dy-\varphi_1(t)+\varphi_0(t),\ (t,x)\in\mathbb{R}_t\times]0,1[.\qquad\square$$

Let us now state another existence result. This time we suppose that (H_1) and (H_∞) hold, but not (H_0'); and we shall prove that the solution has more regularity.

THEOREM 5.2. *Assume that $\varphi_1-\varphi_0\in L^\infty(\mathbb{R}_t)$ and g_0,g_1 are T-periodic functions such that*

$$\begin{aligned}&(H_1)\ \ G_1:=\frac{1}{T}\int_0^T\int_{v>0}vg_0(t,v)dtdv-\frac{1}{T}\int_0^T\int_{v<0}vg_1(t,v)dtdv<+\infty,\\&(H_\infty)\ \ G_\infty:=\max\{\|g_0\|_{L^\infty(\mathbb{R}_t\times\Sigma_0^-)},\|g_1\|_{L^\infty(\mathbb{R}_t\times\Sigma_1^-)}\}<+\infty.\end{aligned}$$

Then there is a T-periodic weak solution (f, E) for the Vlasov–Poisson problem which verifies

$$f \in L^1(]0,T[\times]0,1[\times\mathbb{R}_v), \rho \in L^\infty(\mathbb{R}_t; L^1(]0,1[)),$$
$$|v|f \in L^1(]0,T[\times]0,1[\times\mathbb{R}_v), E \in L^\infty(\mathbb{R}_t\times]0,1[).$$

Moreover, if (H_1') holds,

$$(H_1')\ \ G_1' := \int_{v>0} v \cdot \sup_{t\in\mathbb{R}_t}\{g_0(t,v)\}dv - \int_{v<0} v \cdot \sup_{t\in\mathbb{R}_t}\{g_1(t,v)\}dv < +\infty,$$

then $|v|f$ belongs to $L^\infty(\mathbb{R}_t; L^1(]0,1[\times\mathbb{R}_v))$; in particular $j = \int_{\mathbb{R}_v} vf(t,x,v)dv \in L^\infty(\mathbb{R}_t; L^1(]0,1[))$.

The proof is quite similar to the previous one. We do not go into detail, but we sketch the different arguments below. This time, since $(H_1), (H_\infty)$ are verified, we can apply Theorem 3.4 with $\alpha = \varepsilon > 0$ for the boundary data g_0, g_1. Exactly as in section 4.1, we have

$$\frac{1}{T}\int_0^T\int_0^1\int_{\mathbb{R}_v} f_{\alpha_n}(t,x,v)dtdxdv \le 12\sqrt{2}\cdot\|E_{\alpha_n}\|_{L^\infty}^{1/2}G_\infty + 2G_0,$$

and there is $t_1 = t_1^\alpha \in]0,T[$ such that

$$\int_0^1\int_{\mathbb{R}_v} f_{\alpha_n}(t_1,x,v)dxdv \le 12\sqrt{2}\cdot\|E_{\alpha_n}\|_{L^\infty}^{1/2}G_\infty + 2G_0.$$

By integration of the perturbed Vlasov equation on $]t_1, t[\times]0,1[\times\mathbb{R}_v,\ t \in [t_1, t_1+T]$, we have

$$\begin{aligned}\int_0^1\int_{\mathbb{R}_v} f_{\alpha_n}(t,x,v)dxdv \le& \int_0^1\int_{\mathbb{R}_v} f_{\alpha_n}(t_1,x,v)dxdv\\ &+\int_{t_1}^t\int_{v>0} vg_0(\tau,v)d\tau dv - \int_{t_1}^t\int_{v<0} vg_1(\tau,v)d\tau dv\\ &\le 12\sqrt{2}\cdot\|E_{\alpha_n}\|_{L^\infty}^{1/2}G_\infty + 2G_0 + TG_1.\end{aligned}$$

From the Poisson equation, we have

$$\|E_{\alpha_n}\|_{L^\infty} \le \|\varphi_1 - \varphi_0\|_{L^\infty} + \|\rho_{\alpha_n}\|_{L^\infty(\mathbb{R}_t; L^1(]0,1[))},$$

and therefore we obtain that

$$\|E_{\alpha_n}\|_{L^\infty} \le C\cdot\|E_{\alpha_n}\|_{L^\infty} + D,\ \alpha > 0,$$

with $C = 12\cdot 2^{1/2}G_\infty$, $D = \|\varphi_1 - \varphi_0\|_{L^\infty} + 2G_0 + TG_1$. Finally, one gets for $\alpha > 0$

$$\begin{aligned}\|E_{\alpha_n}\|_{L^\infty} &\le C_E \le C^2 + 2D,\\ \|\rho_{\alpha_n}\|_{L^\infty(\mathbb{R}_t;L^1(]0,1[))} &\le C_\rho \le C(C + D^{1/2}) + 2G_0 + TG_1,\\ \frac{1}{T}\|f\|_{L^1(]0,T[\times]0,1[\times\mathbb{R}_v)} &\le M_\rho \le C(C + D^{1/2}) + 2G_0.\end{aligned}$$

In order to estimate the charge outside a ball B_R in $\mathbb{R}_v$, this time it is easy to calculate integrals like $\mathcal{I}_R(t_1,t_2)$ with

$$\begin{aligned}\mathcal{I}_R(t_1,t_2)&=\int_{t_1}^{t_2}e^{\alpha t}\int_{|v|>R}v(f(t,0,v)-f(t,1,v))(1-\chi_R(v))dtdv\\&\le e^{\alpha t_2}\left(\int_{t_1}^{t_2}\int_{v>R}vg_0(t,v)dtdv-\int_{t_1}^{t_2}\int_{v<-R}vg_1(t,v)dtdv\right)\\&\le e^{\alpha t_2}\left(\int_0^{T}\int_{v>R}vg_0(t,v)dtdv-\int_0^{T}\int_{v<-R}vg_1(t,v)dtdv\right),\ 0\le t_2-t_1\le T,\end{aligned}$$

and the proof follows exactly as before.

Now, in order to prove that $|v|f_{\alpha_n}\in L^1(]0,T[\times]0,1[\times\mathbb{R}_v)$, let us multiply the perturbed Vlasov equation by $|v|$:

$$\begin{aligned}&\alpha(|v|f_{\alpha_n})+\partial_t(|v|f_{\alpha_n})+v\cdot\partial_x(|v|f_{\alpha_n})+\widetilde{E}_{\alpha_n}(t,x)\cdot\partial_v(|v|f_{\alpha_n})\\&\quad=\widetilde{E}_{\alpha_n}\frac{v}{|v|}f_{\alpha_n},(t,x,v)\in\mathbb{R}_t\times]0,1[\times\mathbb{R}_v.\end{aligned}$$

The mild formulation can be written this time as

$$\begin{aligned}\int_0^T\int_0^1\int_{\mathbb{R}_v}|v|f_{\alpha_n}\psi dtdxdv=&\int_0^T\int_0^1\int_{\mathbb{R}_v}\widetilde{E}_{\alpha_n}\frac{v}{|v|}f_{\alpha_n}\int_t^{s_{out}}e^{-\alpha(s-t)}\psi(s,X(s),V(s))dsdtdxdv\\&+\int_0^T\int_{v>0}|v|^2g_0(t,v)\int_t^{s^0_{out}}e^{-\alpha(s-t)}\psi(s,X^0(s),V^0(s))dsdtdv\\&+\int_0^T\int_{v<0}|v|^2g_1(t,v)\int_t^{s^1_{out}}e^{-\alpha(s-t)}\psi(s,X^1(s),V^1(s))dsdtdv\end{aligned}$$

for all $\psi\in\mathcal{T}_m$, and thus, for $\psi=1_{\{|v|>R\}}$ (in fact take $\psi=\chi_{R'}-\chi_R\in\mathcal{T}_m$ with $R'>2R$ and pass $R'\to+\infty$) and R large enough such that $R_1=R-2\sqrt{2}\cdot C_E^{1/2}>0$, we get

$$\begin{aligned}&\int_0^T\int_0^1\int_{|v|>R}|v|f_{\alpha_n}dtdxdv\le C_E\int_0^T\int_0^1\int_{|v|>R_1}f_{\alpha_n}(t,x,v)\frac{1}{|v|-2\sqrt{2}\cdot C_E^{1/2}}dtdxdv\\&+\int_0^T\int_{v>R_1}vg_0(t,v)\frac{v}{v-2\sqrt{2}\cdot C_E^{1/2}}dtdv+\int_0^T\int_{v<-R_1}vg_1(t,v)\frac{v}{-v-2\sqrt{2}\cdot C_E^{1/2}}dtdv\to0,\end{aligned}$$

when $R\to+\infty$, uniformly in respect to $\alpha>0$. By taking for example $R=6\sqrt{2}\cdot C_E^{1/2}$, one gets

$$\begin{aligned}\frac{1}{T}\int_0^T\int_0^1\int_{\mathbb{R}_v}|v|f_{\alpha_n}dtdxdv=&\frac{1}{T}\int_0^T\int_0^1\int_{|v|<R}|v|f_{\alpha_n}dtdxdv+\frac{1}{T}\int_0^T\int_0^1\int_{|v|>R}|v|f_{\alpha_n}dtdxdv\\&\le M_{|j|}\le\left(6\cdot2^{1/2}+\frac{1}{2\cdot2^{1/2}}\right)\cdot C_E^{1/2}M_\rho+2G_1,\end{aligned}$$

and thus $|v|f_{\alpha_n}\in L^1(]0,T[\times]0,1[\times\mathbb{R}_v)$. Now, if $f_{\alpha_n}\rightharpoonup f$ weakly $\star$ in $L^\infty(]0,T[\times]0,1[\times\mathbb{R}_v)$

we also have that $|v|f_{\alpha_n} \rightharpoonup |v|f$ weakly in $L^1(]0,T[\times]0,1[\times B_R)$, $R>0$, and thus

$$\frac{1}{T}\int_0^T\!\!\int_0^1\!\!\int_{|v|<R}|v|f(t,x,v)dtdxdv \le \frac{1}{T}\liminf_{n\to+\infty}\int_0^T\!\!\int_0^1\!\!\int_{|v|<R}|v|f_{\alpha_n}(t,x,v)dtdxdv$$
$$\le \frac{1}{T}\liminf_{n\to+\infty}\int_0^T\!\!\int_0^1\!\!\int_{\mathbb{R}_v}|v|f_{\alpha_n}(t,x,v)dtdxdv \le M_{|j|},\ R>0.$$

It follows that $|v|f \in L^1(]0,T[\times]0,1[\times\mathbb{R}_v)$ and $\frac{1}{T}\int_0^T\int_0^1\int_{\mathbb{R}_v}|v|f(t,x,v)dtdxdv \le M_{|j|}$. In fact we can prove that $|v|f_{\alpha_n} \rightharpoonup |v|f$ weakly in $L^1(]0,T[\times]0,1[\times\mathbb{R}_v)$ and $j_{\alpha_n} \rightharpoonup j$ weakly in $L^1(]0,T[\times]0,1[)$. Assume now that (H_1') holds. In order to estimate j_{α_n} and j in $L^\infty(\mathbb{R}_t;L^1(]0,1[))$, we can apply the same arguments as for the estimates of ρ_{α_n},ρ in $L^\infty(\mathbb{R}_t;L^1(]0,1[))$. This time we have an extra term which can be written as

$$\left|\int_{t_1}^{t_2}e^{\alpha t}\int_0^1\!\!\int_{\mathbb{R}_v}\widetilde{E}_{\alpha_n}\frac{v}{|v|}f_{\alpha_n}dtdxdv\right| \le e^{\alpha t_2}\|E_{\alpha_n}\|_{L^\infty}\int_0^T\!\!\int_0^1\!\!\int_{\mathbb{R}_v}f_{\alpha_n}(t,x,v)dtdxdv,\ 0\le t_2-t_1\le T.$$

In order to estimate the rest of the current, we remark that we have

$$\left|\int_{t_1}^{t_2}e^{\alpha t}\int_0^1\!\!\int_{|v|>R}\widetilde{E}_{\alpha_n}\frac{v}{|v|}f_{\alpha_n}(1-\chi_R(v))dtdxdv\right| \le e^{\alpha t_2}\|E_{\alpha_n}\|_{L^\infty}\int_0^T\!\!\int_0^1\!\!\int_{|v|>R}f_{\alpha_n}(t,x,v)dtdxdv$$

for $0\le t_2-t_1\le T$, and therefore $\int_0^1\int_{|v|>R}|v|f_{\alpha_n}(t,x,v)dxdv \to 0$ when $R\to+\infty$, uniformly for $\alpha>0,\ t\in\mathbb{R}_t$. Finally, we prove that there is $C_{|j|}<+\infty$ (which depends on G_1') such that

$$\int_0^1\!\!\int_{\mathbb{R}_v}|v|f_{\alpha_n}(t,x,v)dxdv \le C_{|j|},\ t\in\mathbb{R}_t,\ \alpha>0,$$
$$\int_0^1\!\!\int_{\mathbb{R}_v}|v|f(t,x,v)dxdv \le C_{|j|},\ t\in\mathbb{R}_t,$$

$$|v|f_{\alpha_n}(t) \rightharpoonup |v|f(t), \text{weak in } L^1(]0,1[\times\mathbb{R}_v),$$
$$j_{\alpha_n}(t) \rightharpoonup j(t), \text{weak in } L^1(]0,1[).$$

Obviously this result can be generalized as follows.

THEOREM 5.3. *Assume that $\varphi_1-\varphi_0\in L^\infty(\mathbb{R}_t), g_0, g_1$ are T-periodic functions such that*

$$(H_p)\ \ G_p := \frac{1}{T}\int_0^T\!\!\int_{v>0}|v|^p g_0(t,v)dtdv + \frac{1}{T}\int_0^T\!\!\int_{v<0}|v|^p g_1(t,v)dtdv < +\infty,$$
$$(H_\infty)\ \ G_\infty := \max\{\|g_0\|_{L^\infty(\mathbb{R}_t\times\Sigma_0^-)}, \|g_1\|_{L^\infty(\mathbb{R}_t\times\Sigma_1^-)}\} < +\infty$$

for some integer $p\ge 1$. Then there is a T-periodic weak solution (f,E) for the Vlasov–Poisson problem which verifies

$$|v|^p f\in L^1(]0,T[\times]0,1[\times\mathbb{R}_v), |v|^{p-1}f\in L^\infty(\mathbb{R}_t;L^1(]0,1[\times\mathbb{R}_v)), E\in L^\infty(\mathbb{R}_t\times]0,1[).$$

Moreover, if (H_p') holds,

$$(H_p')\ \ G_p' := \int_{v>0}|v|^p\sup_{t\in\mathbb{R}_t}\{g_0(t,v)\}dv + \int_{v<0}|v|^p\sup_{t\in\mathbb{R}_t}\{g_1(t,v)\}dv < +\infty,$$

then $|v|^p f$ belongs to $L^\infty(\mathbb{R}_t;L^1(]0,1[\times\mathbb{R}_v))$.

6. Remarks. First, notice that the estimates of f on the outgoing boundary follow immediately. For example, under the hypothesis of the last theorem, after multiplication by $|v|^{p-1}, p \geq 1$ (in fact $|v|^{p-1}\chi_R(v)$, with $R \to +\infty$), and integration of the Vlasov equation, we deduce that

$$\begin{aligned}\frac{1}{T}\int_0^T\!\!\int_{v>0}|v|^p f(t,1,v)dtdv + \frac{1}{T}\int_0^T\!\!\int_{v<0}|v|^p f(t,0,v)dtdv \\ \leq G_p + \frac{1}{T}\int_0^T\!\!\int_0^1\!\!\int_{\mathbb{R}_v}|E|f(p-1)|v|^{p-2}dtdxdv \\ \leq G_p + (p-1)\cdot\|E\|_{L^\infty}\frac{1}{T}\int_0^T\!\!\int_0^1\!\!\int_{\mathbb{R}_v}f(t,x,v)|v|^{p-2}dtdxdv.\end{aligned}$$

On the other hand, it is possible to pass the nonlinear term $E_{\alpha_n}\cdot\partial_v f_{\alpha_n}$ in the perturbed Vlasov equation to the limit by using the average velocity lemma of Diperna and Lions (see [11]). In fact once we have proved that $(f_{\alpha_n})_n, (E_{\alpha_n})_n$ are uniformly bounded in $L^1(]0,T[\times]0,1[\times\mathbb{R}_v)$ and $L^\infty(\mathbb{R}_t\times]0,1[)$, respectively, we deduce that $\partial_t f_{\alpha_n} + v\cdot\partial_x f_{\alpha_n} = -\alpha_n f_{\alpha_n} - \widetilde{E}_{\alpha_n}(t,x)\cdot\partial_v f_{\alpha_n}$ are uniformly bounded in $L^2(]0,T[\times]0,1[\times H^{-1}(\mathbb{R}_v))$. This implies that $(\int_{\mathbb{R}_v} f_{\alpha_n}(t,x,v)\partial_v\varphi dv)_n$ is uniformly bounded in $H^{1/4}(]0,T[\times]0,1[)$ and therefore converges to $\int_{\mathbb{R}_v} f(t,x,v)\partial_v\varphi dv$ strongly in $L^2(]0,T[\times]0,1[)$. The conclusion follows by combining this with the weak convergence of $(E_{\alpha_n})_n$.

All these results can be easily adapted to the Vlasov–Poisson problem (in one dimension) involving several densities f_e, f_i, where, for example, f_e represents the density of electrons and f_i the density of ions.

Let us remark that changing the sign of the right-hand side of the Poisson equation $-\partial^2 U/\partial x^2 = -\rho(t,x)$, which corresponds to an attractive (gravitational) potential obviously does not affect any argument, so all the previous results still hold in this case.

It would be interesting to see if the same kind of arguments apply for studying the multidimensional case. This analysis will be the topic of future related works [8]. We point out that Lemma 2.11 can be easily generalized for a bounded domain $\Omega \subset \mathbb{R}^N$. Indeed, if $(X(s),V(s))$, $s_{in} \leq s \leq s_{out}$ is an arbitrary characteristic associated to a regular field and $u \in \mathbb{R}^N$ with $\|u\| = 1$, then we have

$$\frac{d}{ds}x(s) = v(s), \quad \frac{d}{ds}v(s) = e(s), \ s_{in} \leq s \leq s_{out},$$

where $x(s) = (X(s),u)$, $v(s) = (V(s),u)$, $e(s) = (E(s,X(s)),u)$ for $s_{in} \leq s \leq s_{out}$. Obviously, $x(s)$ belongs to a bounded interval $\omega \subset \mathbb{R}$ of length $\mathrm{diam}(\omega) \leq \mathrm{diam}(\Omega)$ and $\|e\|_{L^\infty} \leq \|E\|_{L^\infty}$. After performing the same computations as in section 2, we get

$$|\, v(s_1) - v(s_2)\,| \leq 2\cdot(2\cdot\mathrm{diam}(\omega))^{1/2}\cdot\|e\|_{L^\infty}^{1/2}, \ s_{in} \leq s_1 \leq s_2 \leq s_{out},$$

which can also be written as

$$|(V(s_1)-V(s_2),u)| \leq 2\cdot(2\cdot\mathrm{diam}(\Omega))^{1/2}\cdot\|E\|_{L^\infty}^{1/2}, \ s_{in}\leq s_1\leq s_2\leq s_{out}, \forall\, u\in\mathbb{R}^N, \|u\|=1,$$

and the conclusion follows.

REFERENCES

[1] N. Abdallah, *Weak solutions of the initial-boundary value problem for the Vlasov-Poisson system*, Math. Methods Appl. Sci., 17 (1994), pp. 451–476.

[2] A. Arseneev, *Global existence of a weak solution of the Vlasov system of equations*, U.R.S.S. Comp. and Math. Phys., 15 (1975), pp. 131–143.

[3] C. Bardos, *Problèmes aux limites pour les équations aux dérivées partielles du premier ordre*, Ann. Sci. École Norm. Sup., 3 (1969), pp. 185–233.

[4] B. Bodin, *Modélisation et simulation numérique du régime de Child-Langmuir*, Thèse de l'Ecole Polytechnique, Palaiseau, France, 1995.

[5] M. Bostan and F. Poupaud, *Periodic solutions of the Vlasov-Poisson system with boundary conditions*, C. R. Acad. Sci. Paris Sér. I Math., 325 (1997), pp. 1333–1336.

[6] M. Bostan and F. Poupaud, *Periodic solutions of the Vlasov-Poisson system with boundary conditions*, Math. Models Methods Appl. Sci., 10 (2000), pp. 651–672.

[7] M. Bostan and F. Poupaud, *Periodic solutions of the* $1D$ *Vlasov-Maxwell system with boundary conditions*, Math. Methods Appl. Sci., 23 (2000), pp. 1195–1221.

[8] M. Bostan, *Permanent Regimes for the Multidimensional Vlasov–Poisson Problem with Boundary Conditions*, manuscript.

[9] H. Brezis, *Analyse fonctionnelle*, Masson, Paris, 1998.

[10] P. Degond and P.-A. Raviart, *An asymptotic analysis of the one-dimensional Vlasov-Poisson system: The Child-Langmuir law*, Asymptot. Anal., 4 (1991), pp. 187–214.

[11] R. J. Diperna and P. L. Lions, *Global weak solutions of the Vlasov-Maxwell system*, Comm. Pure Appl. Math., 42 (1989), pp. 729–757.

[12] R. J. Diperna and P. L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[13] C. Greengard and P.-A. Raviart, *A boundary value problem for the stationary Vlasov-Poisson equations: The plane diode*, Comm. Pure Appl. Math., 43 (1990), pp. 473–507.

[14] Y. Guo, *Singular solutions to the Vlasov-Maxwell system in a half line*, Arch. Ration. Mech. Anal., 131 (1995), pp. 241–304.

[15] E. Horst and R. Hunze, *Weak solutions of the initial value problem for the unmodified nonlinear Vlasov equation*, Math. Methods Appl. Sci., 6 (1984), pp. 262–279.

[16] R. Illner and H. Neunzert, *An existence theorem for the unmodified Vlasov equation*, Math. Methods Appl. Sci., 1 (1979), pp. 530–544.

[17] F. Poupaud, *Boundary value problems for the stationary Vlasov-Maxwell system*, Forum Math., 4 (1992), pp. 499–527.

SIAM J. MATH. ANAL.
Vol. 35, No. 4, pp. 949–973

A MULTICLASS HOMOGENIZED HYPERBOLIC MODEL OF TRAFFIC FLOW*

P. BAGNERINI† AND M. RASCLE‡

Abstract. We introduce a new homogenized hyperbolic (multiclass) traffic flow model, which allows us to take into account the behaviors of different type of vehicles (cars, trucks, buses, etc.) and drivers. We discretize the starting Lagrangian system introduced below with a Godunov scheme, and we let the mesh size h in (x,t) go to 0: the typical length (of a vehicle) and time vanish. Therefore, the variables—here (w,a)—which describe the heterogeneity of the reactions of the different car-driver pairs in the traffic, develop large oscillations when $h \to 0$. These (known) oscillations in (w,a) persist in time, and we describe the *homogenized* relations between velocity and density. We show that the velocity is the *unique* solution "à la Kružkov" of a scalar conservation law, with variable coefficients, discontinuous in x. Finally, we prove that the same macroscopic homogenized model is also the *hydrodynamic limit* of the corresponding multiclass Follow-the-Leader model.

Key words. hyperbolic systems of conservation laws, traffic flow, multiclass models, homogenization, discontinuous flux, uniqueness, hydrodynamic limit

AMS subject classifications. 35L, 35L65

DOI. 10.1137/S0036141002411490

1. Introduction.

In [1], Aw and Rascle developed a new macroscopic model of traffic flow which allows us to avoid some inconsistencies of the models inspired by the gas dynamic system. In [2], a connection is established between the microscopic "Follow-the-Leader model," and (a semidiscretization of) the macroscopic model introduced in [1], which is its hydrodynamic limit.

In this paper, we introduce, still for a single lane traffic, a new macroscopic homogenized hyperbolic model for multiclass traffic flow, described by a nonlinear hyperbolic system of three conservation laws. For references regarding multiclass (or multipopulation) models see, e.g., [6, 13, 17, 48]; see also [43] for inhomogeneous road conditions and uniqueness results.

The starting system, written in Lagrangian mass coordinates, is the following:

$$\begin{cases} \partial_t \tau - \partial_x v = 0, \\ \partial_t w = 0, \\ \partial_t a = 0, \\ (\tau, w, a)(x,0) = (\tau_0, w_0, a_0)(x), \end{cases} \tag{1.1}$$

where $\tau = 1/\rho$ is the specific volume, i.e., the inverse of the density of vehicles (i.e., of the fraction of space occupied by the cars), v is the velocity, and $a \in [0,1]$ is a dimensionless coefficient, which allows us to take into account the behaviors of different types of vehicles and/or of drivers. Here, w describes the difference (up to a

*Received by the editors July 17, 2002; accepted for publication (in revised form) January 31, 2003; published electronically November 4, 2003. This work was partially supported by EU financed network HPRN-CT-2002-00282.
http://www.siam.org/journals/sima/35-4/41149.html
†Instituto di Analisi Numerica del CNR, Universita di Pavia, Pavia, Italy (pbagneri@dimat.unipv.it).
‡Laboratoire J.A. Dieudonné, UMR CNRS n. 6621, Université de Nice, Parc Valrose B.P. 71, 06108 Nice Cedex 2, France (rascle@math.unice.fr).

constant) between the velocity and some equilibrium velocity

$$w = v - V(\tau, a), \tag{1.2}$$

where (up to a constant) $V(\tau, a)$ is an "equilibrium" velocity for a given τ and for a given class a. For simplicity, *all the results are stated with* $w = v - V(\tau, a) := v + aP(\tau)$, *but they remain valid* in the more general case described below.

The last two equations express the fact that w and a are characteristic of each vehicle and therefore do not depend on time in Lagrangian coordinates.

In some sense, the model introduced in [1] was already multiclass, with each class being described by its distance from the *same* equilibrium curve $v = V(\tau)$. Here we add (at least) a second parameter a to let each equilibrium curve $v = V(\tau, a)$ depend on each class. Of course, we could add more parameters, with the same results. We could also take into account the length of cars by considering a nonuniform mesh in x.

We discretize the model (1.1) with the Godunov scheme: let $U_h(\cdot, \cdot)$ be the approximate solution for a discretization $(\Delta x, \Delta t)$ and initial piecewise-constant data U_h^0. On the one hand, h is the step-size of the discretization, but, on the other hand, h can be viewed as a scaling parameter $(x, t) \to (x', t') = (hx, ht)$. Assuming typically that in each cell there is a unique vehicle, we thus consider a large number of vehicles on a long stretch of the road, and the length of the vehicles vanishes as $h \to 0$.

Practically, the distribution of the different types of car-driver pairs can be highly oscillatory. We are thus led to studying a *homogenized* system: we consider a *sequence* of initial data $(v_h^0, w_h^0, a_h^0, \tau_h^0)$, with $w_h^0 = v_h^0 - V(\tau_h^0, a_h^0)$, and $h \to 0$. We assume that the sequence (v_h^0) converges boundedly almost everywhere to some v_0^*, whereas (w_h^0, a_h^0) and therefore τ_h^0 only converge *weakly* to (w_0^*, a_0^*) and τ_0^*.

Let us also emphasize the fact that this model, like the Follow-the-Leader models, is in principle a *single lane* model: no car can pass another car, and therefore the velocities cannot be wildly oscillating, although some differences and even discontinuities (braking, etc.) are permitted. Therefore it is natural to assume that v_0 is a function BV, i.e., with finite total variation.

By contrast, w_0 and a_0 can definitely oscillate: for instance, in rescaled coordinates (x', t'), a good prototype would be $w_h^0(x') := W_0(x', \frac{x'}{h})$ for some given function W_0 (or A^0) of (x', θ) oscillating (but not necessarily periodic) in θ. In what follows we will drop the primes and write (x, t) instead of (x', t').

We then study the weak-star limit (v^*, w^*, a^*, τ^*) of the solution (v_h, w_h, a_h, τ_h) of *the discretization of* system (1.1)–(1.2) as $h \to 0$. Using the notion a Young measure [41] and compensated compactness theory, we study the propagation of these initial oscillations: we show that the Young measure $\nu_{x,t}$ associated with the variables (v, w, a) is a.e. a tensor product

$$\nu_{x,t} = \delta(v - v^*(x, t)) \otimes \mu_x(w, a),$$

where δ is the Dirac mass at the origin and $\mu_x(w, a)$ a probability measure defined a.e. in $\mathbb{R}_x$.

We prove that the approximate solutions U_h constructed by the Godunov scheme satisfy a discrete Lax entropy inequality for any entropy convex *with respect to* τ, and, under the CFL condition, converge as $h = (\Delta t, \Delta x) \to 0$ to an entropy solution of the homogenized system described below, in fact to some measure-valued solution [18]. The existence of such an mv-solution is thus trivial.

In the case where the above-mentioned functions W_0 and A_0 are periodic in the second variable (and more regular in x) we recover the results on the homogenization

of the corresponding Hamilton–Jacobi equation; see, e.g., [29], [32], [21] and the references cited therein. Still in the periodic case, see also [19] for the construction of corrector terms to the next order.

We then study the *uniqueness* of the solution. The main difficulty is that w_h and a_h only converge weakly as $h \to 0$, so that the Young measure $\nu_{x,t}$ is not a δ-function in (w, a). Due to its special form, the system can be rewritten as the two trivial equations $\partial_t w^* = \partial_t a^* = 0$, coupled with a scalar equation where the flux explicitly depends on x, with a low regularity in x, so that we cannot use directly the uniqueness result of Kružkov.

However, since the flux is strictly increasing in τ, we can *exchange* the roles of x and t (and τ and v), so as to obtain an entropy inequality in conservative form, without any additional term involving the x-derivative of the flux. Therefore, we do not need stronger assumptions on the regularity of the flux with respect to x, and we show uniqueness by a variant of the Kružkov "doubling of variables" argument, in which we *first* "let y tend to x", and *then* "let s tend to t", as in [4].

Finally, *last but not least*, we consider the corresponding microscopic multiclass "Follow-the-Leader model":

$$\begin{cases} \dot{\tau}_j = \frac{v_{j+1} - v_j}{\Delta x}, \\ \dot{w}_j = \dot{v}_j - \frac{\partial V}{\partial \tau_j}(\tau_j, a_j) = 0, \ \dot{a}_j = 0, \\ \tau_j(0) = \tau_j^0, \ w_j(0) = w_j^0, \ a_j(0) = a_j^0, \end{cases} \tag{1.3}$$

where v_j is the speed of the vehicles at time t, Δx the length of the vehicle, $\tau_j = 1/\rho_j$ the local "specific volume around vehicle j," and ρ_j the local density, whereas a_j is a coefficient which depends on the type of vehicle.

The function V is the same as in (1.2) and w has the same meaning. The last two equations are due to the assumption that the coefficients (w_j, a_j) characterize each vehicle and therefore do not change in time.

We can easily see, at least formally, that system (1.3) is a semidiscretization in the space of the continuum model (1.1) in Lagrangian coordinates. In fact (see also [2]), we establish *rigorously* in section 6 that the solution of (1.3) converges as $\Delta x \to 0$ to the unique entropy solution of the homogenized macroscopic model.

The outline of the paper is as follows. In section 2, we describe the model and the Riemann problem. We also describe the scaling, and show in Remark 2.1 a *prototype* of measure μ_x for "practical" applications. In section 3, we study the Godunov scheme and the corresponding a priori estimates. In section 4, we first show in Theorem 4.2—at least for a subsequence—the convergence to a (variant of) measure-valued solution when $(\Delta x, \Delta t) \to (0, 0)$, with a fixed ratio and the CFL condition. In Theorem 4.3, we then reformulate the limit system. The homogenized relation between τ and v is now given by (4.6). The limit v is a solution "à la Kružkov" of the first equation of the limit system, i.e., of the scalar equation (4.14) (with variable nonsmooth coefficients) combined with (4.6) and the two trivial equations (4.5) for w and a. Finally, in section 5 we prove the *uniqueness* of this solution and in section 6 we show that we recover (1.3) when $\Delta t \to 0$ and Δx is fixed, and then we show that the solution of (1.3) converges to the *unique* solution of the *same* homogenized model (4.14), (4.5), (4.6), which is therefore the *hydrodynamic limit* of (1.3).

2. Description of the model. In this section, we describe the properties of system (1.1), which we first write in Eulerian coordinates. We then study the Riemann problem, before describing more precisely in section 3 the approximate solutions

constructed by the Godunov scheme and the corresponding a priori estimates for all h. We recall that h is a scaling parameter which tends to 0 (see section 2.4 below) so the sequence (w_h^0, a_h^0, τ_h^0) only converges weak-star in the L^∞ space when $h \to 0$.

2.1. The macroscopic model. Aw and Rascle [1] have introduced a macroscopic model of traffic flow which allows us to avoid the severe inconsistencies of the so-called "second order" models, whose prototype is the Payne–Whitham model [33, 47]. Then, in [2] (see also [23]) a *rigorous* connection is established between the microscopic "Follow-the-Leader model" and a semidiscretization of the macroscopic model introduced in [1] (see also [49]): namely, the macroscopic system can be viewed as the limit of the microscopic Follow-the-Leader (ODE) system (resp. of its explicit first order Euler (time) discretization) when $\Delta x \to 0$ (resp. when $(\Delta x, \Delta t) \to 0$ with a fixed ratio $\Delta t/\Delta x$ satisfying the CFL condition).

In this section, we extend this macroscopic model to describe a multiclass traffic flow, in order to take into account the behaviors of different types of vehicles (cars, trucks, buses, etc.) and drivers (slow, aggressive, etc.). In conservative form, the model is written in Eulerian coordinates as

$$\begin{cases} \partial_t \rho + \partial_x(v\rho) = 0, \\ \partial_t(\rho w) + \partial_x(v\rho w) = 0, \\ \partial_t(\rho a) + \partial_x(v\rho a) = 0, \end{cases} \tag{2.1}$$

where ρ denotes the (normalized) dimensionless density of vehicles, $a \in [0,1]$ a dimensionless coefficient which characterizes each type of vehicle-driver pair, and w is defined by (1.2) with $\tau := 1/\rho$: in other words, up to some constant, w describes the difference between the actual velocity and the equilibrium one. In Lagrangian "mass" coordinates (X,T) [15], the system (2.1)–(1.2) can be rewritten using the form of (1.1)–(1.2). We recall that

$$\partial_x X = \rho, \qquad \partial_t X = -\rho v, \qquad T = t, \qquad \tau := 1/\rho,$$

and that—even for a weak (entropy) solution (see [46])—systems (2.1) and (1.1) are equivalent.

In our case (see [2]), since ρ is dimensionless, $X = \int^x \rho(y,t)dy$ is not a mass, but it describes the total length occupied by cars up to point x.

Now, when $h \to 0$, system (1.1) is again preserved in the rescaled variables $(X', T') := (hX, hT)$ (see section 2.4 below). We then drop the primes and even rewrite (x,t) instead of (X', T').

Finally, in these rescaled Lagrangian variables, we consider system (1.1), with $\tau := 1/\rho$ and w given by (1.2).

We assume that $\forall a$, $V(\cdot, a)$ is strictly increasing and strictly concave. A good prototype of function V could be, up to a constant,

$$V(\tau, a) = -[(1-a)P_1(\tau) + aP_2(\tau)], \ 0 \le a \le 1,$$

or even, more simply,

$$V(\tau, a) = -\,aP(\tau), \ 0 < a_{min} \le a \le a_{max} \le 1. \tag{2.2}$$

Here, P (or P_1, P_2) satisfies the same assumptions as $-V(\cdot, a)$, i.e.,

$$P'(\tau) \le c_1 < 0, \ P''(\tau) \ge c_2 > 0. \tag{2.3}$$

Again, up to a constant, w describes a distance to equilibrium. Practically, $a = a_{min}$ (resp. a_{max}) would correspond to the slowest (resp. fastest) car-driver pairs and P_1 and P_2 to minimum and maximum equilibrium velocities for a given τ. Since V is invertible in τ, we write

$$\tau := \mathcal{T}(v, w, a) := (V(\cdot, a))^{-1}(v - w) = P^{-1}((w - v)/a). \tag{2.4}$$

As we already said, *all results below will be stated* in the particular case (2.2), but they remain valid in the more general situation (1.2). For concreteness, see [2]. A practical example of function $P(\tau)$ is, up to a constant,

$$P(\tau) := \begin{cases} \frac{v_{ref}}{\gamma} \frac{1}{\tau^\gamma}, & \gamma > 0, \\ -v_{ref} \ln(\tau), & \gamma = 0, \end{cases} \tag{2.5}$$

where v_{ref} is a given reference velocity for all classes of vehicles (for instance 90 km/h). Here the parameter γ has no physical meaning, but similar power laws appear, e.g., in the (strongly related) microscopic models [22, 24]; see section 6.

In the remainder of this article, unless explicitly stated, we work in Lagrangian coordinates.

2.2. The Riemann problem. In order to introduce the Godunov method, we first describe the solution of the Riemann problem for (1.1), i.e., of the initial value problem (IVP), with particular data $(\tau_0, w_0, a_0)(x) := (\tau_\pm, w_\pm, a_\pm)$ for $\pm x > 0$.

The eigenvalues of the system (1.1) are

$$\lambda_1 = -\frac{\partial V}{\partial \tau}(\tau, a) = aP'(\tau) < 0 = \lambda_2 = \lambda_3.$$

The system is strictly nonhyperbolic: the associated matrix is diagonalizable, and its diagonal form is

$$\begin{cases} \partial_t v + aP'(\tau)\partial_x v = 0, \\ \partial_t w = 0, \\ \partial_t a = 0, \end{cases} \tag{2.6}$$

where v and (w, a) are the strict Riemann invariants, respectively, associated with λ_1 and $\lambda_2 = \lambda_3 = 0$. The eigenvalue λ_1 is genuinely nonlinear (GNL), i.e., $\forall U$, $\nabla_U \lambda_1 \cdot r^1(U) = -a\, P''(\tau) < 0$, and $\lambda_2 = \lambda_3$ is linearly degenerate (LD), i.e., $\forall U$, $\nabla_U \lambda_k \cdot r^k(U) \equiv 0$, $k = 2, 3$.

We show that the solution of the Riemann problem associated to the system (1.1), consists of two waves: these are a rarefaction or a shock wave associated with λ_1, followed by a contact discontinuity associated with $\lambda_2 = \lambda_3$ (see Figure 2.1).

The equivalent systems (1.1) and (2.1) are called Temple systems (in the extended sense) (see [42]): their shock curves and rarefaction curves coincide. Therefore, in the space of the Riemann invariants v, w, a, the wave curves are straight lines.

PROPOSITION 2.1. *We consider the Riemann problem*

$$\begin{cases} \partial_t \begin{pmatrix} \tau \\ w \\ a \end{pmatrix} - \partial_x \begin{pmatrix} v \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \\ (\tau, w, a)(x, 0) = (\tau_\pm, w_\pm, a_\pm) := U_\pm \text{ for } \pm x \geq 0. \end{cases} \tag{2.7}$$

The solution $U(x, t) := (\tau, w, a)(x, t)$ is as follows:

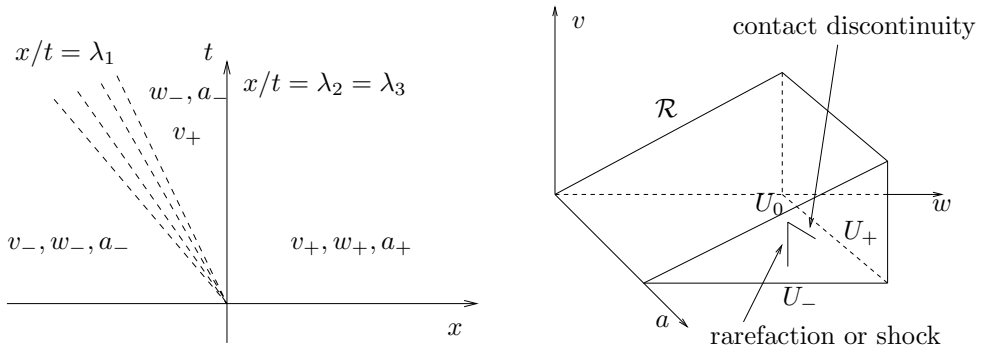


FIG. 2.1. *Riemann problem (here in the case of a* 1*-rarefaction wave) and invariant region.*

(i) *we connect* U_- *to an intermediate constant state* $U_0 = (\tau_0, w_0, a_0)$ *such that* $v_0 = v_+$, $w_0 = w_-$, $a_0 = a_-$, *by a* 1*-rarefaction if* $v_+ > v_-$ *or by a* 1*-shock if* $v_+ < v_-$. *We then connect* U_0 *through* U_+ *by a* 2*-*3 *contact discontinuity of velocity* 0, *with* $v_0 = v_+$.

(ii) *Moreover,* w *and* a *only take the values* (w_-, a_-) *and* (w_+, a_+), *and remain constant in time for each* x. *Now,* v *is a monotone function of* x/t, *with* $\min(v_-, v_+) \le v(x,t) \le \max(v_-, v_+)$.

(iii) *In a* 1*-wave, the specific volume* τ *varies in the same direction as* v; *i.e., it is a monotone function of* x/t. *Finally,* $\forall x$, τ *is monotone with respect to* t, *and* $\forall t \ge 0$,

$$\min(\tau_-, \tau_0) \le \tau(x,t) \le \max(\tau_0, \tau_+).$$

(iv) *Therefore,* $U(x,t)$ *and* $v(x,t)$ *remain in an invariant bounded region* $\mathcal{R}$, *away from the vacuum:*

(2.8)
$$\mathcal{R} := \{(\tau, w, a); (v, w, a) \in [v_{min}, v_{max}] \times [w_{min}, w_{max}] \times [a_{min}, a_{max}]\},$$

with v_{min} *(e.g.* $v_{min} = 0$*),* $v_{max}, w_{min}, w_{max}, a_{min}, a_{max} \ge 0$ *some constants and* $\max\{\tau, (\tau, w, a) \in \mathcal{R}\} < \infty$ *(in the case* (2.2) *we assume* $a_{min}, a_{max} > 0$*).*

Proof. The proof is classical. The monotonicity of v is due to the fact that v only takes the values v_-, v_+ and the wave curves are straight lines in the (v, w, a) space. Finally, since (w, a) is constant in a 1-wave and P is monotone, $\tau = P^{-1}((w - v)/a)$ varies in the same sense as v. ∎

The qualitative properties of the solution are as expected: braking corresponds to a shock and accelerating to a rarefaction; no information travels faster than the velocity of cars; the velocity and the density remain nonnegative and bounded, etc.

2.3. Entropies of the system. In what follows, since $v = w + V(\tau, a)$, we will sometimes denote the entropies,

$$\eta(v, w, a) = \eta(w + V(\tau, a), w, a) \overset{e.g.}{=} \eta(w - aP(\tau), w, a), \tag{2.9}$$

by $\tilde{\eta}(U) = \tilde{\eta}(\tau, w, a)$ or even—incorrectly—by $\eta(\tau, w, a)$, when this notation is not ambiguous. In such a case $\partial_\tau \eta$ means $\partial_\tau \tilde{\eta}$, so that, e.g., $\partial_\tau \eta(\tau,w,a) = \partial_v \eta(v,w,a)(-aP'(\tau))$. There is no such problem for the entropy fluxes q, which only depend on v (see below). Now we study the entropy-flux pairs (η, q) of the system (1.1). In terms of the

variables (v, w, a), formally multiplying the left-hand side of (2.6) by $(\partial_v\eta, \partial_w\eta, \partial_a\eta)$, we obtain

$$\begin{cases} aP'(\tau)\partial_v\eta = \partial_v q, \\ 0 = \partial_w q, \\ 0 = \partial_a q. \end{cases} \tag{2.10}$$

Therefore, the flux q associated with η only depends on the variable v, i.e., $q \equiv q(v)$, and *all* entropy-flux pairs are given by

$$\eta(v, w, a) = \int_0^v \frac{q'(s)}{aP'(\tau(s, w, a))} ds + \eta_0(w, a),\ q = q(v) \tag{2.11}$$

for any function $\eta_0 = \eta_0(w, a) := \eta(w, a, 0)$ and $q = q(v)$. By (2.9) and the first equation (2.10), we obtain

$$\partial_\tau\eta(\tau, w, a) = \partial_v\eta(v, w, a)\, \partial_\tau v(\tau, w, a) = \frac{1}{aP'(\tau)} q'(v)(-aP'(\tau)) = -q'(v). \tag{2.12}$$

PROPOSITION 2.2. *For all entropy $\eta = \eta(\tau, w, a)$, associated with the flux $q \equiv q(v)$, we have the following:*

(i) *η is convex* with respect to τ *if and only if $q \equiv q(v)$ is concave: $q''(v) \le 0$.*

(ii) *Let U be the solution to the Riemann Problem* (2.7)*. Then, for all entropy η satisfying (i), we have*

$$\partial_t\eta + \partial_x q \le 0 \text{ in } \mathcal{M}(\mathbb{R} \times (0, \infty)). \tag{2.13}$$

Proof. (i) Differentiating η with respect to τ twice (with fixed (w, a)), and using (2.12), we obtain

$$\partial_\tau^2\eta(\tau, w, a) = \partial_\tau(-q'(v)) = \partial_v(-q'(v))\, \partial_\tau v = q''(v)\, a\, P'(\tau).$$

Therefore, η is convex if and only if q is concave, since $a > 0$ and $P' < 0$.

(ii) Through a 1-rarefaction wave and a 2-3 contact discontinuity, for $x > 0$, (2.13) is an equality. Therefore, through a 1-shock wave, for $x < 0$, we have $(w, a) = (w_-, a_-)$. Using the Rankine–Hugoniot relations between U_- and $U = U_0$, the entropy condition is equivalent to proving that

$$S(v) := S_{v_-}(v) = \frac{v_- - v}{\tau - \tau_-}(\eta(\tau, w, a) - \eta(\tau_-, w, a)) - (q(v) - q(v_-)) \ge 0,$$

i.e., that $S(\cdot)$ is decreasing, since $v < v_-$ in a 1-shock and $S_{v_-}(v_-) = 0$. Since $\tau = \mathcal{T}(v, w, a) := P^{-1}((w - v)/a)$, differentiating S with respect to v, we obtain

$$S'(v) = \frac{-(\tau - \tau_-) - (v_- - v)\frac{\partial\mathcal{T}}{\partial v}}{(\tau - \tau_-)^2}(\eta(\tau, w, a) - \eta(\tau_-, w, a)) + \frac{v_- - v}{\tau - \tau_-}\frac{\partial\eta}{\partial\tau}\frac{\partial\mathcal{T}}{\partial v} - q'(v).$$

Using (2.12) and recombining the terms, we obtain

$$S'(v) = -\frac{\eta(\tau_-, w, a) - \eta(\tau, w, a) - (\tau_- - \tau)\frac{\partial\eta}{\partial\tau}}{(\tau - \tau_-)^2}\left(\mathcal{T}(v_-, w, a) - \mathcal{T}(v, w, a) - (v_- - v)\frac{\partial\mathcal{T}}{\partial v}\right),$$

which is nonpositive since η is convex *with respect to* τ and $\mathcal{T}$ strictly convex with respect to v.

We emphasize that we *only* need the convexity of η in τ, since (w, a) is constant through a 1-shock, i.e., here for $x < 0$. ∎

2.4. Scaling. An example of μ_x. The macroscopic models are only valid if we consider a large number of vehicles on a long stretch of the road. Therefore, we introduce a scaling (zoom) such that the size of the domain and the number of vehicles is going to infinity, whereas the length of the vehicles tends to 0 (see also [2]). Given the Lagrangian coordinates x, t, we consider the new rescaled coordinates

$$x' := hx, \quad t' := ht,$$

where h is a small parameter, which is proportional to the inverse of the maximal possible number of cars per new unit length. In the new coordinates, x', t', the variables ρ, τ and the Riemann invariants v, w, a are unchanged, whereas the length of a vehicle becomes $\Delta x' = h\Delta x$, which tends to 0 as $h \to 0$: in the new coordinates, the convergence of the Godunov scheme to the entropy solution of (4.4) can be viewed as the convergence of the microscopic system to the macroscopic model, when the size of the road and the number of vehicles are going to infinity.

For instance, assume that the initial units are meters and seconds and the new units are, 1500 m and 60 seconds, with $\Delta x = 5m$. In the rescaled coordinates, $x' := x/1500$, $t' := t/60$, so that a reference velocity of 90 km/h, i.e., 25 m/s, becomes 1 in the new units and the length of the car $\Delta x' = 1/300$, which is a "reasonable" step-size.

In particular, a typical sequence of oscillating initial data $(w_0^h, a_0^h)(x) := (W_0, A_0)(x, hx)$ (and its limit in L^∞ weak*) could be in the new coordinates

$$(w_h^0, a_h^0)(x') = (W_0, A_0)\left(\frac{x'}{h}, x'\right) \underset{h\to 0}{\rightharpoonup} (w_0^*, a_0^*)(x') = (\langle \mu_x, w\rangle, \langle \mu_x, a\rangle)\ L^\infty \text{ weak*},$$

where the Young measure μ_x will be introduced in section 4.1.

REMARK 2.1. *Typically, given a finite number N of classes a_i, associated with $w_i = v_i - V(\tau_i, a_i)$, a possible choice of μ_x is given by*

$$\mu_x(w, a) = \sum_{i=1}^{N} \mu_i(x)\delta_{w_i, a_i}(w, a), \ \textit{with} \ \sum_{i=1}^{N} \mu_i(x) \equiv 1,$$

where the nonnegative coefficients $\mu_i(x)$ are the local proportion (possibly 0) of each class of car-driver pairs and δ the Dirac measure. When we then compute an approximation of the average *velocity v^* by the Godunov scheme, not only do we know the average specific volume $\tau^*(x,t) = \sum_{i=1}^{N} \mu_i(x)\mathcal{T}(v^*(x,t), w_i, a_i)$ introduced in equation* (4.6)*, but also the specific volume τ_i (and then the density) of* every class a_i.

3. The Godunov scheme. Now we discretize system (1.1), which can be written in general form as

$$\partial_t U + \partial_x G(U) = 0,$$

with the Godunov scheme. We introduce a grid in time and space, with step-size Δx and Δt (related to a parameter h as indicated below) and grid point $x_{j-1/2}$ and t_n. Let I_j be the *open* interval $(x_{j-1/2}, x_{j+1/2})$ and, $\forall(\Delta x, \Delta t)$, with $\Delta t/\Delta x = C$, the sequence $U_h := (\tau_h, w_h, a_h)$ is the approximation by the Godunov scheme, given $\forall n \geq 0$ by

$$\begin{cases} \partial_t \tau_h - \partial_x v_h = 0, & \text{in } \mathbb{R} \times (t_n, t_{n+1}), \\ \partial_t w_h = 0, \ \partial_t a_h = 0, \end{cases} \tag{3.1}$$

with $v_h := w_h - a_h P(\tau_h)$ and with piecewise-constant initial data

$$U_h(x, t_n^+) := (\tau_h, w_h, a_h)(x, t_n^+) := \sum_{j\in\mathbb{Z}} (\tau_j^n, w_j^n, a_j^n)\chi_j(x),$$

where $\chi_j(y) = 1$ on $(x_{j-1/2}, x_{j+1/2})$ and 0 elsewhere. Let $U_j^n := (\tau_j^n, w_j^n, a_j^n)$ denote the average value of the function $U_h(x,t)$ in the interval I_j, i.e.,

$$U_j^n := (1/\Delta x) \int_{I_j} U_h(x, t_n^-)dx.$$

In every cell $I_j \times]t_n, t_{n+1}[$, we compute the solution to the local Riemann problems centered on the grid points $x_{j\pm1/2}$, with initial data (U_j^n, U_{j+1}^n). Let $G_{j+1/2}^n := G(U_j^n, U_{j+1}^n) = (-v_{j+1}, 0, 0)$ be the flux at point $x_{j+1/2}$. The solution U_j^{n+1} to the Godunov scheme is given by

$$U_j^{n+1} := U_j^n - \frac{\Delta t}{\Delta x}(G_{j+1/2}^n - G_{j-1/2}^n) = U_j^n + \frac{\Delta t}{\Delta x}(v_{j+1}^n - v_j^n, 0, 0). \tag{3.2}$$

with $v_j^n := w_j^n - a_j^n P(\tau_j^n)$. Now, we assume that $(\forall h)$, $\forall x \in \mathbb{R}$, $U_h(x,0) = U_h^0(x) = (\tau_h^0, w_h^0, a_h^0) \in \mathcal{R}$, where $\mathcal{R}$ is defined in (2.8), and for simplicity (see (4.1)), we also assume that the sequence v_h^0 is bounded in $BV(\mathbb{R})$, i.e., $\sum_{j\in\mathbb{Z}} |v_{j+1}^0 - v_j^0| \le C_0 < +\infty$. This assumption is sufficient in practical life (just think of a road where the velocity of cars is *not* in $BV(\mathbb{R})$!), but it is not necessary since λ_1 is "GNL." We do *not* assume that the sequences w_h^0 and a_h^0 are bounded in BV.

The following theorem shows that the region $\mathcal{R}$ defined in (2.8) is also invariant for the Godunov scheme and gives estimates on the total variation of some components of the solution. Under the above assumptions, we have the following.

THEOREM 3.1. *Let Δx and Δt satisfy the CFL condition. Then the following hold:*

(i) *The region $\mathcal{R}$ remains invariant for the Godunov scheme.*

(ii) *$\forall n \ge 0$, the total variation (in x) of $v_h(\cdot, t)$ is nonincreasing in time, and the total variation in t of $\tilde{v}_h(\cdot,\cdot)$ is bounded on $\mathbb{R}\times[0,T]$:*

$$TV_x(v_h^n; \mathbb{R}) := \sum_{j\in\mathbb{Z}} |v_j^n - v_{j+1}^n| \le TV(v_h^0; \mathbb{R}) , \tag{3.3}$$

$$\sup_h \sup_{t\ge0} TV_x(v_h(\cdot,t); \mathbb{R}) \le \sup_h TV(v_h^0(\cdot); \mathbb{R}) := C_0, \tag{3.4}$$

$$\sup_h TV_t(\tilde{v}_h(\cdot,\cdot); \mathbb{R}\times[0,T]) \le C \max(T, \Delta t)\, C_0 , \tag{3.5}$$

where $v_h^0 := v_h(\cdot, 0^+)$ is the piecewise-constant approximation of the initial datum $v_0^*(\cdot)$, and

$$\tilde{v}_h(x,t) \equiv v_j^n + (t - t_n)(v_j^{n+1} - v_j^n)/\Delta t \text{ on } I_j \times (t_n, t_{n+1}), \tag{3.6}$$

and similarly for $\tilde{\tau}_h$.

(iii) *The total variation (in x) of $\tau_h(\cdot,t)$ on $\cup_{j\in\mathbb{Z}} I_j$ and the total variation in t of $\tilde{\tau}_h(\cdot,\cdot)$ on $\mathbb{R}\times[0,\infty[$ are bounded, uniformly in h:*

$$\sup_h \sup_{t\ge0} \sum_{j\in\mathbb{Z}} TV_x(\tau_h(\cdot,t); \cup_{j\in\mathbb{Z}} I_j) \le C'\, C_0, \tag{3.7}$$

$$\forall 0 \le t \le t',\ \sup_h TV_t(\tilde{\tau}_h(\cdot,\cdot); \mathbb{R}\times[t,t']) \le C' \max(|t'-t|, \Delta t)\, C_0. \tag{3.8}$$

(iv) *$\forall x \in \mathbb{R}$, $(w_h, a_h)(x,t) \equiv (w_h^0, a_h^0)(x)$.*

Proof. (i) This part of the proof follows easily from the solution of the Riemann problem in Proposition 2.1.

(ii) Adding and subtracting v_{j+1}^n and recombining the terms in the sum, we obtain

$$\begin{aligned} TV_x(v_h^{n+1};\mathbb{R}) &= \sum_{j\in\mathbb{Z}} |v_j^{n+1} - v_{j+1}^{n+1}| \le \sum_{j\in\mathbb{Z}} (|v_j^{n+1} - v_{j+1}^n| + |v_{j+1}^n - v_{j+1}^{n+1}|) \\ &= \sum_{j\in\mathbb{Z}} (|v_j^{n+1} - v_{j+1}^n| + |v_j^n - v_j^{n+1}|). \end{aligned}$$

Now, using the monotonicity property of v in the Riemann problem (see Proposition 2.1), we see that the average value v_j^n of $v(\cdot, t_{n+1})$ on I_j in the projection step belongs to the interval $I(v_j^n, v_{j+1}^n)$ (see Figure 2.1), and thus

$$|v_j^{n+1} - v_{j+1}^n| + |v_j^n - v_j^{n+1}| = |v_{j+1}^n - v_j^n|.$$

Therefore, we obtain

$$TV_x(v_h^{n+1};\mathbb{R}) \le \sum_{j\in\mathbb{Z}} |v_{j+1}^n - v_j^n| \le \cdots \le \sum_{j\in\mathbb{Z}} |v_{j+1}^0 - v_j^0| = TV(v_h^0;\mathbb{R}),$$

and for the same reason we obtain (3.4).

Concerning (3.5), since $\tilde{v}_h$ is piecewise-linear in time, $v_j^{n+1} \in I(v_j^n, v_{j+1}^n)$ for $t \in [t_n, t_{n+1})$, with $I(a,b) = [\min(a,b), \max(a,b)]$. Therefore,

$$\begin{aligned} TV_t(\tilde{v}_h(x,\cdot);[0,T]) &= \sum_{j\in\mathbb{Z}} \sum_{n\le \frac{T}{\Delta t}} |v_j^{n+1} - v_j^n| \Delta x \le \sum_{j\in\mathbb{Z}} \sum_{n\le \frac{T}{\Delta t}} |v_{j+1}^n - v_j^n| \Delta x \\ &\le \frac{T}{\Delta t} \Delta x TV(v_h^0;\mathbb{R}) \le C \ \max(T, \Delta t) \ C_0, \end{aligned} \tag{3.9}$$

where $C = \Delta x/\Delta t$. See an alternate proof below with a uniform constant C.

(iii) Since in each cell I_j, $\tau_h(x,t) = \mathcal{T}(v_h(x,t), w_j, a_j) = P^{-1}\left(\frac{w_j - v_h(x,t)}{a_j}\right)$, with P^{-1} Lipschitz-continuous, we have

$$TV_x(\tau_h(\cdot,t), \cup_{j\in\mathbb{Z}} I_j) = \sum_{j\in\mathbb{Z}} TV_x(\tau_h(\cdot,t), I_j) \le \|(P^{-1})'\|_{L^\infty} (a_{min})^{-1} \, C_0$$

and similarly for (3.8).

(iv) This part of the proof is obvious. ∎

REMARK 3.1. *We first note that, in general, the Godunov approximate solutions do* not *satisfy such a BV estimate. The result is true here since in each cell* (w,a) *is* constant. *We also note that the total variation in space of* τ_h *on* $\cup_{j\in\mathbb{Z}} I_j$ *is bounded, whereas its total variation on* $\mathbb{R}$ *can be infinite, due to the jumps in* $x = x_{j+1/2}\ \forall j \in \mathbb{Z}$. *Finally, as to the total variation in time, we could proceed differently and use* (3.4) *and the first equation in* (3.2) *to show first* (3.8) *and then* (3.9) *with in each case a constant* C *independent of* $\Delta x/\Delta t$*; see section* 6.

PROPOSITION 3.2. *On any time interval* $[t_n, t_{n+1})$, *the solution* $U_h = (\tau_h, w_h, a_h)$ *satisfies the* discrete entropy inequality *in the sense of Lax: for any entropy* $\eta(U_h)$ *convex with respect to* τ_h *and associated to the entropy flux* $q(U_h)$, *for any* n *and* j,

$$\eta(U_j^{n+1}) \le \eta(U_j^n) - (\Delta t/\Delta x)(q(U_{j+1}^n) - q(U_j^n)). \tag{3.10}$$

Proof. Classically, on any time interval $[t_n, t_{n+1})$, U_h is the solution of the Riemann problem (2.7) and satisfies the entropy inequality (2.13) in every cell I_j. The Jensen inequality allows us to conclude this. We remark that, since (w, a) is *constant* in each cell, we only need the convexity of η *with respect to* τ. ∎

Finally, the sequence (U_h) is a sequence of *approximate* solutions to (1.1), associated for each h with the initial data U_h^0. More precisely, we have the following.

PROPOSITION 3.3.

(i) $\forall h > 0$, $\forall \varphi \in \mathcal{D}(\mathbb{R}_x \times \mathbb{R}_+) = C_0^\infty((\mathbb{R}_x \times \mathbb{R}_+)$,

$$\begin{aligned} &\int_0^\infty \int_{\mathbb{R}} (\tau_h \partial_t \varphi - v_h \partial_x \varphi)(x,t) dx dt + \int_{\mathbb{R}} \tau_h(x,0)\varphi(x,0) dx \\ &= -\sum_{n\geq 1} \sum_{j\in\mathbb{Z}} \int_{I_j} (\tau_j^n - \tau_h(x, t_n^-))\varphi(x, t_n) dx := \langle L_h, \varphi \rangle. \end{aligned} \tag{3.11}$$

(ii) *If the support of* φ *is compact in* $\{0 \leq t \leq T\}$, *then*

$$|\langle L_h, \varphi \rangle| \leq (\Delta x)^2 \|\partial_x \varphi\|_{L^\infty} \sum_{n\geq 0} \sum_{j\in\mathbb{Z}} TV_x(\tau_h(\cdot, t); I_j), \tag{3.12}$$

$$|\langle L_h, \varphi \rangle| \leq C\, T\, \|\varphi\|_{L^\infty} \sup_{n \leq \frac{T}{\Delta t}} \sum_j TV_x(\tau_h(\cdot, t_n^-); I_j). \tag{3.13}$$

(iii) *Therefore,* $L_h \to 0$ *as* $h \to 0$ *in* $\mathcal{D}'(\mathbb{R} \times \mathbb{R})$ *and* $\forall T > 0$, *the sequence* (L_h) *is bounded in* $\mathcal{M}(\mathbb{R} \times [0, T])$.

(iv) *Concerning the entropy production,* $\forall \eta$ convex with respect to τ, *associated to the flux* q, $\forall \varphi \in \mathcal{D}(\mathbb{R} \times \mathbb{R}_+)$, $\varphi \geq 0$, *we have*

$$\begin{aligned} &\int_0^\infty \int_{\mathbb{R}} (\eta(U_h)\partial_t \varphi + q(U_h)\partial_x \varphi)(x,t) dx dt + \int_{\mathbb{R}} \eta(U_h(x,0))\varphi(x,0) dx \\ &\geq -\sum_{n\geq 1} \sum_{j\in\mathbb{Z}} \int_{I_j} (\eta(U_j^n) - \eta(U_h(x, t_n^-))\varphi(x, t_n) dx \geq 0. \end{aligned} \tag{3.14}$$

Consequently, $\forall C^2$ *the nonnegative entropy* η *associated with* q *such that* $x \to \eta(U_h^0(x))$ *is integrable on* $\mathbb{R}$, $\partial_t \eta + \partial_x q$ *is a bounded measure on* $\mathbb{R} \times \mathbb{R}_+$, *which is nonpositive if* η *is convex with respect to* τ.

Proof. The proof of (i), (ii), (iii) uses the classical arguments provided by the BV estimates in the Theorem 3.1. In particular, the BV property of τ_h on $\cup_{j\in\mathbb{Z}} I_j$ (and not on $\mathbb{R}$) is sufficient for (3.12). On the other hand (3.13) shows that the functional L_h is a bounded measure. The proof of (iv) classically combines the entropy inequality for the Riemann problem (Proposition 2.2) and a discrete integration by parts (see for instance [28]), as well as the obvious remark that any C^2 function is the difference of two C^2 convex functions. ∎

We are now ready to pass to the limit as $h \to 0$.

4. The homogenized model: Existence of a solution.

4.1. The Young measure. In order to introduce into the model oscillations describing the heterogeneity of the reaction of each car-driver pair in the traffic, we consider a sequence of oscillating initial data (w_h^0, a_h^0, τ_h^0) bounded in $L^\infty(\mathbb{R})$. We study the evolution in time of these initial oscillations for the approximate solution (w_h, a_h, τ_h) constructed by the Godunov method.

For a few references in the study of large amplitude oscillations in nonlinear hyperbolic systems of conservation laws, we refer, e.g., to [18, 40, 35, 38, 10, 14, 34].

Here, we simply consider the measure-valued solution [18] associated with this sequence of approximate solutions U_h. We consider that there is typically a unique vehicle in every cell I_j, and we pass to the limit in the system (3.1) as $h \to 0$, in order to obtain a homogenized model. To describe the limit as $h \to 0$ of τ_h, which is a nonlinear function of the three variables v_h, w_h, a_h, we need the concept of Young measures.

Let us briefly recall the concept of Young measures associated with a sequence u^n; see, e.g., [5, 41, 16, 30]. For any sequence $u^n : \mathbb{R}^N \to \mathbb{R}^p$ of measurable functions, with values in a fixed compact set $K \subset \mathbb{R}^p$, and such that $u^n \overset{*}{\rightharpoonup} u^*$ in $L^\infty(\mathbb{R}^N)^p$, there exists a subsequence of u^n, still denoted by u^n, and a family of probability measures $\{\nu_x\}_{x\in\mathbb{R}^N}$, called *Young measures*, uniformly supported in K such that for *any* $f \in C(\mathbb{R}^p;\mathbb{R}^q)$, and for almost all x in $\mathbb{R}^N$,

$$f(u^n) \overset{*}{\rightharpoonup} f^* \neq f(u^*) \text{ in } L^\infty(\mathbb{R}^N)^q, \text{ with } f^*(x) = \int_{\mathbb{R}^N} f(s)d\nu_x(s) = \langle \nu_x(\cdot), f(\cdot)\rangle.$$

From now on, we assume that the sequence of initial data, with values in $\mathcal{R}$, satisfies the following:

(4.1)

The *whole* sequence (w_h^0, a_h^0, τ_h^0) converges in L^∞ weak* to a unique limit (w_0^*, a_0^*, τ_0^*) whereas $v_h^0 \to v_0^*$ boundedly a.e. as $h \to 0$, with $\sup_h TV(v_h^0) \le C_0$.

Thus, we assume that $\forall h$, the total variation of the sequence v_h^0 is bounded from above. As we already stated, this assumption is not necessary, since λ_1 is GNL, but it is more than sufficient for practical applications.

Here by Theorem 3.1, the sequence (v_h, w_h, a_h) remains in the invariant region $\mathcal{R}$. Consequently, for any continuous function f, at least for a subsequence, the associated Young measure $\nu_{x,t}$ satisfies

$$f(v_h, w_h, a_h)(x,t) \overset{*}{\rightharpoonup} f^* := \langle \nu_{x,t}(v,w,a), f(v,w,a)\rangle,$$

where (v,w,a) are (dummy) integration variables and $\overset{*}{\rightharpoonup}$ denotes the convergence in $L^\infty(\mathbb{R}\times\mathbb{R}_+)$ weak*. In particular, since $\tau_h = \mathcal{T}(v_h, w_h, a_h) = P^{-1}((w_h - v_h)/a_h)$,

$$\tau_h \overset{*}{\rightharpoonup} \tau^*(x,t) = \langle \nu_{x,t}(w,a), \mathcal{T}(v,w,a)\rangle = \langle \nu_{x,t}(w,a), P^{-1}((w-v)/a)\rangle. \tag{4.2}$$

In the following theorem, we show that $\nu_{x,t}$ is a tensor product of a Dirac measure associated with v and a probability measure μ_x (depending only on x) associated with (w,a).

THEOREM 4.1. *Let $U_h = (v_h, w_h, a_h)$ be constructed by the Godunov scheme. We assume that U_h^0 is bounded in $L^\infty(\mathbb{R})$ and (v_h^0) bounded in $BV(\mathbb{R})$. Then, by assumptions* (2.3) *and* (4.1)*, the following hold:*

(i) *Under the CFL condition, with $\Delta t/\Delta x = $ constant, as $h \to 0$, there exists a subsequence, still denoted by U_h, such that*

$$(w_h, a_h)(x,t) \equiv (w_h^0, a_h^0)(x) \overset{*}{\rightharpoonup} (w_0^*, a_0^*), \ \tau_h(\cdot,\cdot) \overset{*}{\rightharpoonup} \tau^*(\cdot,\cdot),$$
$$v_h(\cdot,\cdot) \to v^*(\cdot,\cdot) \text{ in } L^1_{loc}(\mathbb{R}\times\mathbb{R}_+) \text{ strong and in } L^\infty(\mathbb{R}\times\mathbb{R}_+) \text{ weak*.}$$

(ii) *In the variables* (v, w, a), $\nu_{x,t}$ *is a tensor product. More precisely, since* λ_1 *is genuinely nonlinear (GNL), even if* v *were not initially in* $BV(\mathbb{R})$, *we have*

$$\nu_{x,t} = \gamma_{x,t}(v) \otimes \beta_{x,t}(w,a) := \delta(v - v^*(x,t)) \otimes \mu_x(w,a), \tag{4.3}$$

where δ *is the Dirac measure at* 0 *and* μ_x *does not depend on time.*

(iii) *In particular,*

$$\tau_h \overset{*}{\rightharpoonup} \tau^*(x,t) = \langle \mu_x(w,a), P^{-1}((w - v^*(x,t))/a) \rangle.$$

Proof. (i) This part of the proof follows directly from the L^∞ and BV estimates on U_h and v_h.

Here is a first proof of (ii). The sequence $(\tilde{v}_h)$ has a uniformly bounded total variation and is equicontinuous in time in the L^1 space. Therefore it is strongly convergent (at least for a subsequence) to some $v^*(x,t)$, and therefore $\nu_{x,t}$ satisfies (4.3).

A second proof, with no BV assumptions, is the following. Using Murat's lemma [31] and Proposition 3.3, we apply the div-curl lemma [41] to entropy-flux pairs ($\eta_1 = \eta_1(w,a), q_1 \equiv 0$) and ($\eta_2 = \int^v q'(s)/(aP'(\tau(s,w,a)))ds, q_2$), with arbitrary η_1, q_2. Therefore, $\forall \eta_1, q_2$,

$$\langle \nu_{x,t}, \eta_1(w,a) q_2(v) \rangle = \langle \nu_{x,t}, \eta_1(w,a) \rangle \langle \nu_{x,t}, q_2(v) \rangle.$$

Therefore $\nu_{x,t} = \gamma_{x,t}(v) \otimes \mu_x(w,a)$. The associated measure μ_x depends only on x, since w and a do not depend on t. Now, let $[v_1, v_2]$ be the convex hull of the support of $\gamma_{x,t}$, with $v_1 < v_2$. We want to prove that $\gamma_{x,t}$ is a Dirac measure, i.e., that $v_1 = v_2$.

We again apply the div-curl theorem to entropy-flux pairs of "east-west type" $(\eta_1^\epsilon, q_1^\epsilon)$ and $(\eta_2^\epsilon, q_2^\epsilon)$ (see [38]): e.g., we choose a smooth q_1^ϵ such that its support is contained in $(-\infty, \bar{v}_1]$, where $\bar{v}_1 = v_1 + \epsilon(v_2 - v_1)$, and such that $q_1^\epsilon(v) \to -H(v_1 - v)$ as $\epsilon \to 0$, with H the Heaviside function. Similarly, $q_2^\epsilon(v) \to H(v - v_2)$.

Let η_1^ϵ,η_2^ϵ be the associated entropies. Applying the div-curl theorem and passing to the limit as $\epsilon \to 0$, we obtain

$$\left\langle \mu_x(w,a), \frac{1}{aP'(\mathcal{T}(v_1,w,a))} \right\rangle = \left\langle \mu_x(w,a), \frac{1}{aP'(\mathcal{T}(v_2,w,a))} \right\rangle,$$

which implies $v_1 = v_2$, since $P'(\mathcal{T}(v,w,a))$ is strictly monotone with respect to v.

(iii) This part of the proof is then obvious. ∎

4.2. Existence of a weak entropy solution. In this section, we prove the convergence of the Godunov scheme to a weak entropy solution to the initial value problem for the homogenized system introduced below, as $(\Delta t, \Delta x) \to (0,0)$ with a fixed ratio satisfying the CFL condition.

THEOREM 4.2.

Under the same assumptions as in Theorem 4.1, *the following hold:*

(i) *At least for a subsequence, in fact for the whole sequence, the sequence* (U_h) *constructed by the Godunov scheme converges to a weak solution of the system*

$$\begin{cases} \partial_t \tau^* - \partial_x v^* = 0, \\ \partial_t w^* = 0, \ \partial_t a^* = 0, \end{cases} \tag{4.4}$$

with initial data

$$\begin{cases} \tau_0^* = \langle \mu_x(w,a), \mathcal{T}(v_0^*(x), w, a)\rangle, \\ w_0^* = \langle \mu_x(w,a), w\rangle,\ a_0^* = \langle \mu_x(w,a), a\rangle, \end{cases} \tag{4.5}$$

and

$$\tau^*(x,t) = \langle \mu_x(w,a), \mathcal{T}(v^*(x,t), w, a)\rangle := \mathcal{T}^*(x, v^*(x,t)), \tag{4.6}$$

where $\mathcal{T}$ is defined by (2.4).

(ii) *The family of probability measures $\nu_{x,t} = \delta(v - v^*(x,t)) \otimes \mu_x(w,a)$ is a measure-valued (mv) solution in the following sense: for any entropy-flux pair (η, q), defined by* (2.11)*, with*

$$\eta(v,w,a) := \tilde{\eta}(\tau, w, a)\big|_{\tau = \mathcal{T}(v,w,a)} \tag{4.7}$$

and $\tilde{\eta}$ convex with respect to τ*, i.e., q concave with respect to v (by Proposition* 2.2 (i)*); and for any test-function $\varphi \geq 0$, we have*

$$\begin{aligned} &\int_0^\infty \int_{\mathbb{R}} (\langle \nu_{x,t}(v,w,a), \eta(v,w,a)\rangle \partial_t \varphi + \langle \nu_{x,t}, q(v)\rangle \partial_x \varphi)\ dxdt \\ &+ \int_{\mathbb{R}} \langle \nu_{x,0}(v,w,a), \eta(v,w,a)\rangle\ \varphi(x,0)\, dx \geq 0. \end{aligned} \tag{4.8}$$

(iii) *Applying* (4.8) *to arbitrary entropy-flux pairs $(\eta(w,a), 0)$, we recall that $\mu_{x,t}(w,a) \equiv \mu_x(w,a)$ satisfies* (4.5)*. In addition, $\forall \varphi \geq 0$, $\forall k \in \mathbb{R}$, we have*

$$\begin{aligned} &\int_0^\infty \int_{\mathbb{R}} (\langle \mu_x(w,a), |\mathcal{T}(v^*(x,t), w, a) - \mathcal{T}(k,w,a)|\rangle \partial_t \varphi - |v^*(x,t) - k| \partial_x \varphi) dxdt \\ &+ \int_{\mathbb{R}} \langle \mu_x(w,a), |\mathcal{T}(v_0^*(x), w, a) - \mathcal{T}(k,w,a)|\rangle \varphi(x,0) dx \geq 0. \end{aligned} \tag{4.9}$$

Proof. We first prove this theorem for a subsequence. The uniqueness Theorem 5.2 will then imply the same results for the whole sequence.

(i) We write the first equation of (4.4) in the weak form as

$$\langle L, \varphi\rangle := \int_0^\infty \int_{\mathbb{R}} (\tau^* \partial_t \varphi - v^* \partial_x \varphi) dxdt + \int_{\mathbb{R}} \tau_0^*(x) \varphi(x,0) dx = 0. \tag{4.10}$$

We add and subtract in (4.10) the term $\langle L_h, \varphi\rangle$ of (3.11), to obtain

$$\begin{aligned} \langle L, \varphi\rangle = &\int_0^\infty \int_{\mathbb{R}} (\tau^* - \tau_h) \partial_t \varphi\, dxdt - \int_0^\infty \int_{\mathbb{R}} (v^* - v_h) \partial_t \varphi\, dxdt \\ &+ \int_{\mathbb{R}} (\tau^*(x,0) - \tau_h(x,0)) \varphi(x,0) dx + \langle L_h, \varphi\rangle. \end{aligned}$$

Of course the first three integrals converge to 0 in $L^\infty(\mathbb{R}\times\mathbb{R}_+)$ weak*, due to Theorem 4.1, and, by (3.12), the term $\langle L_h, \varphi\rangle$ tends to 0 as $h \to 0$.

(ii) Inequality (4.8) is obtained by passing to the weak* limit in (3.14), using the L^1 equicontinuity in time of $\tilde{v}_h$ or $\tilde{\tau}_h$, see Proposition 5.1.

(iii) Now, we choose in (4.8) the entropy $\eta(v,w,a) = |\tau - P^{-1}((w-k)/a)| = |\mathcal{T}(v,w,a) - \mathcal{T}(k,w,a)|$, associated with the concave flux $(-|v-k|)$. Using the information (4.3) on the structure of $\nu_{x,t}$, we obtain (4.9). □

The above theorem describes the point of view of *measure-valued solutions*; see again [18, 40, 35], etc. We note that the uniqueness results of Di Perna and Szepessy are not directly applicable here (see Remark 5.2).

Now, the structure of $\mu_{x,t} \equiv \mu_x$ is obvious, whereas the evolution of $\delta(v - v^*(x,t))$ turns out to be governed by a scalar conservation law, whose flux depends on x in a nonsmooth way through μ_x. The striking fact is that, due to the strict monotonicity of function $\mathcal{T}(\cdot,w,a)$, we can do two things:

(i) the map $\{v^* \to \mathcal{T}^*(x,v^*) := \tau^* = \langle \mu_x, \mathcal{T}(v^*,w,a)\rangle$ is strictly increasing, and (therefore) invertible. We define $\mathcal{V}(x,\tau^*) = v^*$.

(ii) On the other hand, in (4.8), we can "take the absolute value out of the integral with respect to μ_x." Therefore (4.8) looks like the Kružkov entropy inequality for the scalar equation (4.14) below. We consider *all* the entropy-flux pairs of the form $H(x,v^*) = \tilde{H}(x,\tau^*)$, $Q(v^*)$, "conservative" in the sense that, for any smooth solution of (4.14),

$$\partial_t H(x,v^*) + \partial_x Q(v^*) = 0,$$

without any additional term. Moreover, we can show as in section 2.3 that

$$\frac{\partial \tilde{H}}{\partial \tau}(x,\tau^*) = \frac{\partial H}{\partial v}(x,v^*)\frac{1}{\frac{\partial \mathcal{T}^*}{\partial v}(x,v^*)} = -Q'(v^*), \tag{4.11}$$

and that $\tilde{H}(x,\tau^*)$ is convex in τ^* if and only if $Q(v^*)$ is concave. Therefore, *all* these entropy-flux pairs are given by

$$H(x,v^*) = H_0(x) - \int_0^{v^*} Q'(s)\frac{\partial \mathcal{T}^*}{\partial v}(x,s)ds = \tilde{H}(x,\tau^*),\ \ Q \equiv Q(v^*), \tag{4.12}$$

with arbitrary functions $Q(v)$ and $H_0(x) = H(x,0)$.

THEOREM 4.3.

(i) *The first equation of system* (4.4) *can be rewritten as a scalar equation with a flux depending explicitly on* x *as*

$$\partial_t \tau^* + \partial_x \mathcal{V}(x,\tau^*) = 0, \tag{4.13}$$

or in the more convenient form as

$$\partial_t \mathcal{T}^*(x,v^*) - \partial_x v^* = 0. \tag{4.14}$$

Therefore, the weak solution of system (4.4) *satisfies* (4.14), (4.6).

(ii) *Using the monotonicity in* v^* *of* $\mathcal{T}^*(x,v^*)$, v^* *is also a weak solution "à la Kružkov" of* (4.14)*:* $\forall k$ *in* $\mathbb{R}$, (4.9) *is equivalent to*

$$\begin{aligned}\iint &(|\mathcal{T}^*(x,v^*) - \mathcal{T}^*(x,k)|\partial_t\varphi - |v^*-k|\partial_x\varphi)dxdt \\ &+ \int_{\mathbb{R}} |\mathcal{T}^*(x,v_0^*) - \mathcal{T}^*(x,k)|\varphi(x,0)dx \geq 0.\end{aligned} \tag{4.15}$$

(iii) *For any such entropy* $H(x,v^*)=\tilde H(x,\tau^*)$ *convex in* τ^**, associated to a concave* $Q(v^*)$ *by* (4.12)*, let* η *be given by* (2.11)*, with the* same $q\equiv Q$*. Then, for any test-function* $\varphi\ge 0$*, we have*

$$H(x,v^*)=\langle\nu_{x,t}(v,w,a),\eta(v,w,a)\rangle=\langle\mu_x(w,a),\eta(v^*(x,t),w,a)\rangle, \tag{4.16}$$

$$\int_0^\infty\int_{\mathbb{R}}(H(x,v^*)\,\partial_t\varphi-Q(v^*)\,\partial_x\varphi)dxdt+\int_{\mathbb{R}}H(x,v_0^*)\varphi(x,0)dx\ge 0. \tag{4.17}$$

Proof. (i) This part of the proof is obvious.

(ii) Using the fact that $\mathcal{T}$ is increasing in v^*, we see that $\forall(w,a)$,

$$\operatorname{sign}(\mathcal{T}(v^*,w,a)-\mathcal{T}(k,w,a))=\operatorname{sign}(v^*-k)$$

does not depend on (w,a), so that

$$\begin{aligned}&\langle\mu_x(w,a),|\mathcal{T}(v^*,w,a)-\mathcal{T}(k,w,a)|\rangle\\&=\operatorname{sign}(v^*-k)\langle\mu_x(w,a),\mathcal{T}(v^*,w,a)-\mathcal{T}(k,w,a)\rangle=|\mathcal{T}^*(x,v^*)-\mathcal{T}^*(x,k)|.\end{aligned}$$

(iii) Let η be defined by (2.11). Using (4.11) and differentiating the integral with respect to v, one can show that necessarily

$$\begin{aligned}&H(x,v^*)-H_0(x)=\int_0^{v^*}\frac{\partial H}{\partial v}(x,s)ds=\int_0^{v^*}-Q'(s)\frac{\partial\mathcal{T}^*}{\partial v}(x,s)ds\\&=\int_0^{v^*}-Q'(s)\frac{\partial}{\partial v}\,\langle\mu_x(w,a),\mathcal{T}(s,w,a)\rangle\,ds=\left\langle\mu_x(w,a),\int_0^{v^*}-Q'(s)\frac{\partial\mathcal{T}}{\partial v}(s,w,a)ds\right\rangle\\&=\langle\mu_x(w,a),\eta(v^*,w,a)\rangle.\end{aligned}$$

Substituting the expression of (4.16) into (4.8), we obtain (4.17). ∎

REMARK 4.1. *Therefore we have proved* (4.8) *and* (4.17)*, respectively, for any entropy* $\eta(v,w,a)$ *and for the homogenized entropy* $H(x,v^*)$ convex *in* τ*. Thus these two quantities satisfy the corresponding entropy inequalities, but there is a priori no obvious relation between them, since* $\eta(v,w,a)=\tilde\eta(\tau,w,a)$ *is only convex* in τ*, with* no convexity assumption in (w,a)*:* (4.16) *holds true, but in general*

$$H(x,v^*)=\langle\mu_x(w,a),\eta(v^*,w,a)\rangle\ \neq\eta(\langle\nu_{x,t},(v,w,a)\rangle)=\eta(v^*,w^*,a^*).$$

5. Uniqueness of the solution. We are interested in the homogenized model (4.4)–(4.6). The existence of an entropy solution of (4.4) follows directly from the convergence of the Godunov method in Theorem 4.2.

Concerning uniqueness, knowing the measure μ_x, we have written the first equation of system (4.4)–(4.6) as the scalar equation (4.13) with a flux depending explicitly on x. The low regularity in x of the flux does not allow us to directly use the uniqueness result of Kružkov [27].

For references on the uniqueness of the solution for scalar conservation laws with a flux discontinuous in x, see, e.g., [4, 26, 25, 44, 37], and concerning Temple systems, e.g., [12, 3, 11, 8, 9], but these references are not applicable in this homogenized case.

Since in (4.14) the function $\mathcal{T}^*(x, v^*)$ (explicitly depending on x) appears in the derivative with respect to t, we have exchanged the roles of x and t so that inequality (4.15) is in conservative form, even though the flux depends on x. Therefore, assumptions on the x-regularity of the flux of (4.13) are not required.

The following proposition gives the L^1-continuity in time of τ^*, which will be useful for the uniqueness of the solution.

PROPOSITION 5.1. *Let $(\tilde{\tau}_h, \tilde{v}_h)$ be the approximate solution defined in* (3.1), (3.6) *of system* (1.1), *with $\tau_0^* \in L^\infty(\mathbb{R})$ and $v_0^* \in L^\infty(\mathbb{R}) \cap BV(\mathbb{R})$. Then, at least for a subsequence,*

$$\tilde{v}_h \to v^* \text{ in } C^0([0,T]; L^1_{loc}(\mathbb{R})), \tag{5.1}$$

$$\|\tau^*(\cdot,t) - \tau_0^*(\cdot)\|_{L^1_{loc}(\mathbb{R})} \le C \max(t, \Delta t)\, VT_x(v_0^*(\cdot); \mathbb{R}), \tag{5.2}$$

and then

$$\|\tau^*(\cdot,t) - \tau_0^*(\cdot)\|_{L^1_{loc}(\mathbb{R})} \to 0 \text{ as } t \to 0.$$

Proof. Let $X = BV(\mathbb{R})$ and $B = L^1_{loc}(\mathbb{R})$. The sequence $(\tilde{v}_h)$ is bounded in $L^\infty(0,T;X)$ by (3.4). On the other hand, $\forall$ compact subsets K of $\mathbb{R}$, $v_h \in C([0,T]; L^1(K))$, and by (3.8),

$$\begin{aligned} &\|\tilde{v}_h(\cdot,t') - \tilde{v}_h(\cdot,t)\|_{L^1(K)} \le \int_K \int_t^{t'} |\partial_s \tilde{v}_h(x,s)| ds dx = \|\partial_t \tilde{v}_h\|_{L^1(t,t';L^1_{loc}(\mathbb{R}))} \\ &= TV_t(\tilde{v}_h(x,\cdot); [t,t']) \le C' \max(|t'-t|, \Delta t), \end{aligned} \tag{5.3}$$

where C' depends on $TV(v_0^*(\cdot); \mathbb{R})$. By a theorem of Simon ([39], Thm. 3), the sequence $\tilde{v}_h(\cdot,t)$ is relatively compact in $C([0,T]; L^1_{loc}(\mathbb{R}))$, which implies (5.1). Passing to the limit as $h \to 0$ in (5.3), with $t' = 0_+$, since $\mathcal{T}$ (or P^{-1}) is L-Lipschitz continuous and $\tau^*(\cdot, 0_+) = \tau_0^*(\cdot)$, we obtain

$$\|\tau^*(\cdot,t) - \tau_0^*(\cdot)\|_{L^1_{loc}(\mathbb{R})} \le L \|v^*(x,t) - v_0^*(\cdot)\|_{L^1_{loc}(\mathbb{R})} \le L\, C'\, t \to 0 \text{ as } t \to 0. \qquad \square$$

THEOREM 5.2. *We consider the scalar equation* (4.14), (4.6) *(or the equivalent equation* (4.13), (4.6), *due to the monotonicity in v^* of $\mathcal{T}^*$), with initial data* (4.5) *and under the assumption that V, defined by* (1.2), *is strictly increasing and strictly concave. We also assume that $(\tau_0^*, w_0^*, a_0^*) \in L^\infty(\mathbb{R})$, $v_0^* \in L^\infty(\mathbb{R}) \cap BV(\mathbb{R})$, and* (4.1). *Then, there is a unique weak entropy solution "à la Kružkov" v^* of problem* (4.14), (4.6), (4.5).

Therefore the whole sequence (τ_h, w_h, a_h, v_h) *converges to the unique limit of the system described in Theorem* 4.2.

In the proof, we will need the following lemma, *in which we write v instead of v^*.*

LEMMA 5.3. *Let $\mathcal{T}^*(X, v(Y,t)) := \langle \mu_X(w,a), \mathcal{T}(v(Y,t), w, a) \rangle$ with $\mathcal{T}$ Lipschitz with respect to $v \in L^1_{loc}(\mathbb{R} \times \mathbb{R}_+)$. Then we have, for almost all X in $\mathbb{R}$ and t in $(0, +\infty)$,*

$$\lim_{h \to 0} \frac{1}{h} \int_{|Y| \le h} |\mathcal{T}^*(X \pm Y, v(X \pm Y, t)) - \mathcal{T}^*(X, v(X,t))|\, dY = 0,$$

where the two signs $\pm$ *can be chosen independently.*

Proof. The result would be *wrong* for a general function of two variables. It holds true here since, roughly speaking, $\mathcal{T}^*(X, v(Y,t))$ "looks like" a product $f(X)g(Y,t)$.

Here, we write λ instead of (w,a). For almost all (fixed) t, we first choose the *precise representing function* of $v(\cdot,t)$ ([20] p. 46), which is precisely defined at every Lebesgue point for $\{y \mapsto v(y,t)\}$, and identically equal to 0 on the null set $N_1 := N_1(t)$ of points y which are not Lebesgue points for this function. Similarly, (see, e.g., [5]), we "remove" a null set $N_2 := N_2(t)$ of points x such that either μ_x is *not* a probability measure or x is *not* a Lebesgue point for the L^1_{loc} function

$$\{x \mapsto F_v(x,t) := \mathcal{T}^*(x, v(x,t)) := \langle \mu_x(\lambda), \mathcal{T}(v(x,t),\lambda)\rangle\}, \tag{5.4}$$

with the same (fixed) t. We see that the set $\{(x,x), x \in N_1 \cup N_2\}$ is a (one dimensional) null set. Therefore, each $x \notin N_1(t) \cup N_2(t)$ is *simultaneously* a Lebesgue point for the two functions of *one* variable (5.4) and $\{y \mapsto G_v(x,y,t) := \langle \mu_x(\lambda), \mathcal{T}(v(y,t),\lambda)\rangle\}$. In particular, $\forall x \notin N_1(t) \cup N_2(t)$, $G_v(x,x,t) = F_v(x,t)$. The same result would be true if $\mathcal{T}$ were only *continuous*; see, e.g., [7] for the related notion of Caratheodory functions.

Therefore, we have for instance

$$\begin{aligned}
&\lim_{h\to 0} \frac{1}{h} \quad \int_{|Y|\le h} |\mathcal{T}^*(X+Y, v(X-Y,t)) - \mathcal{T}^*(X, v(X,t))|\, dY \\
&= \lim_{h\to 0} \frac{1}{h} \quad \int |\langle \mu_{X+Y}(\lambda), \mathcal{T}(v(X-Y,t),\lambda)\rangle - \langle \mu_X(\lambda), \mathcal{T}(v(X,t),\lambda)\rangle\, dY \\
&\le \lim_{h\to 0} \frac{1}{h} \quad \int \{ |\langle \mu_{X+Y}(\lambda), \mathcal{T}(v(X-Y,t),\lambda)\rangle - \langle \mu_{X+Y}(\lambda), \mathcal{T}(v(X,t),\lambda)\rangle| \\
&\qquad\qquad + |\langle \mu_{X+Y}(\lambda), \mathcal{T}(v(X,t),\lambda)\rangle - \langle \mu_{X+Y}(\lambda), \mathcal{T}(v(X+Y,t),\lambda)\rangle| \\
&\qquad\qquad + |\langle \mu_{X+Y}(\lambda), \mathcal{T}(v(X+Y,t),\lambda)\rangle - \langle \mu_X(\lambda), \mathcal{T}(v(X,t),\lambda)\rangle| \}\, dY \\
&\le \lim_{h\to 0} \frac{1}{h} \quad \int \{ \langle \mu_{X+Y}(\lambda), L|v(X-Y,t) - v(X,t)| + L|v(X,t) - v(X+Y,t)|\rangle \\
&\qquad\qquad + |\langle \mu_{X+Y}(\lambda), \mathcal{T}(v(X+Y,t),\lambda)\rangle - \langle \mu_X(\lambda), \mathcal{T}(v(X,t),\lambda)\rangle| \}\, dY \\
&\le \lim_{h\to 0} \frac{1}{h} \quad \int \{ L|v(X-Y,t) - v(X,t)| + L|v(X,t) - v(X+Y,t)| \\
&\qquad\qquad + |\langle \mu_{X+Y}(\lambda), \mathcal{T}(v(X+Y,t),\lambda)\rangle - \langle \mu_X(\lambda), \mathcal{T}(v(X,t),\lambda)\rangle| \}\, dY.
\end{aligned} \tag{5.5}$$

Therefore, almost everywhere in t, for any $X \notin N_1(t) \cup N_2(t)$, the integrals in (5.5) converge to 0 as $h \to 0$. ∎

REMARK 5.1. *Integrating the result of Lemma* 5.3 *with respect to* X *and applying the Lebesgue theorem, the corresponding double integrals in* (X,Y) *converge to* 0 *as* h *tends to* 0. *This new result does not explicitly involve the above Lebesgue points and could be also proved* [45] *as follows: approximate the function* $\{v : y \mapsto v(y,t)\}$ *by a sequence of smooth functions* (v_n) *for which there is no ambiguity: a.e. in* x, $\forall n, F_{v_n}(x,t) = G_{v_n}(x,y,t)|_{y=x}$. *Then, justify the result for each* v_n, *and pass it to the limit as* $n \to \infty$.

Proof of Theorem 5.2. We consider two weak entropy solutions, σ and τ, of (4.13) (or two solutions u and v of (4.14)) which satisfy the entropy inequality (4.15). In this proof, we write $\sigma, \tau, u, v, \mathcal{T}, \dots$ instead of $\sigma^*, \tau^*, u^*, v^*, \mathcal{T}^*, \dots$. We have

$$\int_{\mathbb{R}}\int_0^\infty |\mathcal{T}(x,u(x,t)) - \mathcal{T}(x,k)|\phi_t(x,t) - |u(x,t)-k|\phi_x(x,t)\,dx\,dt \geq 0, \tag{5.6}$$

$$\int_{\mathbb{R}}\int_0^\infty |\mathcal{T}(y,v(y,s)) - \mathcal{T}(y,l)|\phi_s(y,s) - |v(y,s)-l|\phi_y(y,s)\,dy\,ds \geq 0, \tag{5.7}$$

where $\phi \geq 0$ is a test-function $\in C_0^\infty(\mathbb{R}\times(0,+\infty))$. Following Kružkov [27], we obtain classically

$$\int_{\mathbb{R}}\int_0^\infty \int_{\mathbb{R}}\int_0^\infty \{|\mathcal{T}(x,u(x,t)) - \mathcal{T}(x,v(y,s))|\phi_t + |\mathcal{T}(y,v(y,s)) - \mathcal{T}(y,u(x,t))|\phi_s$$
$$-|u(x,t)-v(y,s)|(\phi_x+\phi_y)\}\,dx\,dt\,dy\,ds \geq 0. \tag{5.8}$$

We choose

$$\phi(x,t,y,s) = \psi\left(\frac{x+y}{2}, \frac{t+s}{2}\right)\delta_h\left(\frac{x-y}{2}\right)\delta_k\left(\frac{t-s}{2}\right)$$

for any function ψ and where $\{\delta_h\}_{h\geq 0}$ and $\{\delta_k\}_{k\geq 0}$ are the usual regularizing sequences, with bounded support in $(-h,h)$, with $0 \leq \delta_1(\cdot) \leq 1$. Denoting by

$$X = \frac{x+y}{2}, \quad T = \frac{t+s}{2}, \quad Y = \frac{x-y}{2}, \quad S = \frac{t-s}{2}, \tag{5.9}$$

we rewrite (5.8) as

$$\frac{1}{2}\int \{|\mathcal{T}(x,u(x,t)) - \mathcal{T}(x,v(y,s))| + |\mathcal{T}(y,v(y,s)) - \mathcal{T}(y,u(x,t))|\}\psi_T\,\delta_h(Y)\,\delta_k(S)$$
$$+\{|\mathcal{T}(x,u(x,t)) - \mathcal{T}(x,v(y,s))| - |\mathcal{T}(y,v(y,s)) - \mathcal{T}(y,u(x,t))|\}\,\psi\,\delta_h(Y)\,\delta_k'(S)$$
$$-|u(x,t)-v(y,s)|\,\psi_X(X,T)\,\delta_h(Y)\,\delta_k(S)\,dx\,dt\,dy\,ds := (I_1+I_2)+(I_3-I_4)-I_5 \geq 0.$$

We now proceed in the same spirit as [4]: we let h and k tend to 0 separately. We *only* show the convergence of $(I_3 - I_4) := (I_3^{k,h} - I_4^{k,h})$, for which *it is crucial to first let h tend to* 0, since this term involves $\delta_k'(S)$. The proof for the other terms is similar. Writing $(I_3 - I_4)$ in variables (X,Y,t,s), adding and subtracting the term

$$I_6 := \int |\mathcal{T}(X,u(X,t)) - \mathcal{T}(X,v(X,s))|\psi(X,(t+s)/2)\delta_h(Y)\delta_k'((t-s)/2)dXdYdtds,$$

we first obtain

$$|I_3 - I_4| \leq |I_3 - I_6| + |I_6 - I_4|.$$

Then, using the relation $||a-b|-|c-d|| \leq |a-b-(c-d)| \leq |a-c|+|b-d|$, we have

$$|I_3 - I_6| \leq \frac{1}{h}\int_{|Y|\leq h} \{|\mathcal{T}(X+Y,u(X+Y,t)) - \mathcal{T}(X,u(X,t))|$$
$$+ |\mathcal{T}(X+Y,v(X-Y,s)) - \mathcal{T}(X,v(X,s))|\}|\psi(X,(t+s)/2)|\delta_k'((t-s)/2)\,dY\,dX\,dt\,ds, \tag{5.10}$$

and similarly,

(5.11)
$$|I_6 - I_4| \le \frac{1}{h}\int_{|Y|\le h} \{|\mathcal{T}(X-Y, v(X-Y,s)) - \mathcal{T}(X, v(X,s))| \\ + |\mathcal{T}(X-Y, u(X+Y,t)) - \mathcal{T}(X, u(X,t))|\}|\psi(X,(t+s)/2)|\delta_k'((t-s)/2)\,dY\,dX\,dt\,ds.$$

By Lemma 5.3, for any fixed $k > 0$, a.e. in (s,t,X), the integrals in Y corresponding to the above integrals tend to 0 as $h \to 0$. Applying then the Lebesgue dominated convergence theorem, the corresponding integrals in (X,Y), and finally in (X,Y,t,s), also tend to 0 as $h \to 0$ for any fixed k. A fortiori,

$$\lim_{k\to 0}\left(\lim_{h\to 0}(I_3^{k,h} - I_4^{k,h})\right) = \lim_{k\to 0}(0) = 0.$$

Therefore, this singular term $(I_3 - I_4)$ vanishes at the limit, contrary to the other terms for which we again apply the Lebesgue theorem when k tends to 0.

Then, we have shown that, for any $\psi \ge 0$ (we now write (x,t) instead of (X,T) and $\mathcal{T}^*$ instead of $\mathcal{T}$),

(5.12)
$$\int_{\mathbb{R}}\int_0^\infty |\mathcal{T}^*(x,u(x,t)) - \mathcal{T}^*(x,v(x,t)|\psi_t(x,t) - |u(x,t) - v(x,t)|\psi_x(x,t)\,dx\,dt \ge 0.$$

Now, we classically choose the test-function $\psi(x,t)$ in (5.12) as a regularization of the characteristic function of the set $\Omega = \{(T,X); t_1 \le T \le t_2, |X| \le R - NT\}$ for any $R > 0$. Using the L^1-continuity in time of the solution at $t = 0$ (see Proposition 5.1), we obtain the L^1 contraction property

$$\int_{\mathbb{R}} |\mathcal{T}^*(x,u(x,t)) - \mathcal{T}^*(x,v(x,t)|\,dx \le \int_{\mathbb{R}} |\mathcal{T}^*(x,u(x,0)) - \mathcal{T}^*(x,v(x,0)|\,dx; \text{ i.e.,}$$

$$\int_{\mathbb{R}} |\sigma(x,t) - \tau(x,t)|\,dx \le \int_{\mathbb{R}} |\sigma(x,0) - \tau(x,0)|\,dx,$$

and therefore we have shown the uniqueness of the solution for a given $\tau(\cdot,0)$. ∎

Remark 5.2.

(i) *First, we have established here the L^1 contraction for equations* (4.14), (4.6), *with only L^∞ assumptions in μ_x and $v^*(x,t)$. We have essentially used the strict monotonicity of $\mathcal{T}(\cdot,w,a)$, which allows us to exchange the role of x and therefore those of τ and v. After this exchange, the stationary solutions $u^\zeta(x)$ introduced in* [4] *to solve the equation*

$$\partial_t u + \partial_x F(x,u) := \partial_t u + \partial_x f(v(x),u) = 0$$

arise much more naturally: compare (4.15) *with formula* (60) *in* [4].

(ii) *On the other hand, the uniqueness results on measure-valued solutions of Di Perna* [18] *and Szepessy* [40] *are (at least) not directly applicable here, since we deal with a system: the Young measure $\nu_{x,t} = \delta(\cdot - v^*(x,t)) \otimes \mu_x(\cdot,\cdot)$ involves several variables, which do not play the same role. Here, we wanted to prove the uniqueness of v^* for a given μ_x. Note that for instance, convolving in (x,t) such a measure $\nu_{x,t}$ does* not *preserve its tensor product structure.*

6. The microscopic model.

6.1. Introduction of the model. Microscopic models of vehicular traffic are usually based on so-called *Follow-the-Leader models* [22, 24], which usually consist, in Eulerian coordinates, of a system of second order ordinary differential equations. The basic idea is that the acceleration at time t depends on the relative speeds of the vehicle and its leading vehicle at time t and the distance between the cars.

The system in (1.3) is a Follow-the-Leader type of model in which the function V also depends on the coefficients a_j, characteristic of the different types of vehicles and their heterogeneous response to their leading vehicle.

At least formally, the system (1.3) is clearly a semidiscretization in space of the macroscopic model in Lagrangian coordinates (1.1). In the following sections, we will give a rigorous justification that (1.3) converges to the entropy solution of (4.4)–(4.6) as $\Delta x \to 0$.

6.2. First order Euler time approximation. We consider the infinite system of ordinary differential equations (1.3), which is written in general form as

$$\begin{cases} \frac{dU(t)}{dt} = F(U(t)), \\ U(0) = U_0, \end{cases} \tag{6.1}$$

where $U := (U_1, U_2, \dots, U_j, \dots)$ and $F(U) = (F_1(U), F_2(U), \dots, F_j(U), \dots)$, with $U_j := (\tau_j, w_j, a_j)$ and

$$F_j(U) := \left(\frac{v_{j+1} - v_j}{\Delta x}, 0, 0 \right) = \left(\frac{w_{j+1} - w_j}{\Delta x} - \frac{a_{j+1} P(\tau_{j+1}) - a_j P(\tau_j)}{\Delta x}, 0, 0 \right).$$

We introduce the first order explicit Euler discretization in time:

$$\begin{cases} \tau_j^{n+1} = \tau_j^n + \frac{\Delta t}{\Delta x} (v_{j+1}^n - v_j^n), \\ w_j^{n+1} = w_j^n, \\ a_j^{n+1} = a_j^n. \end{cases} \tag{6.2}$$

In the following theorem, we prove that, for any fixed Δx, the previous discretization is stable and consistent and therefore convergent as $\Delta t \to 0$ to $U_{\Delta x}(x,t) := \sum_{j \in \mathbb{Z}} U_j(t) \chi_j(x)$, where χ_j is the characteristic function of the interval I_j. We also show that the microscopic multiclass model (1.3) is the semidiscretization of the Lagrangian system (1.1).

THEOREM 6.1. *We consider the system* (6.2)*, with initial data* $U_j^0 = (\tau_j^0, w_j^0, a_j^0)$ *in the invariant region* $\mathcal{R}$ *defined by* (2.8)*, away from vacuum. We assume that the initial data are constant for* x *large enough so that there is a "first" vehicle. Then, the following hold:*

(i) *The operator* F *is Lipschitz-continuous in the* l^∞ *space. Therefore the initial value problem* (1.3) *has a unique solution* $U(t)$*, globally defined in time.*

(ii) *The first order approximation* (6.2) *is stable and consistent in* l^∞*. Therefore the sequences* $U_j^n := (\tau_j^n, w_j^n, a_j^n)$ *and* v_j^n *converge as* $\Delta t \to 0$*, for any fixed* Δx*. Moreover, their limits* $U_{\Delta x}(x,t)$ *and* $v_{\Delta x}(x,t)$ *stay in the region* $\mathcal{R}$ *and satisfy the uniform* L^∞ *and* BV *estimates inherited from the Godunov scheme. The microscopic model* (1.3) *is then the semidiscretization of the macroscopic system* (1.1), (1.2).

(iii) *$U_j(t)$ satisfies a semidiscrete entropy inequality, i.e., for any entropy η* convex with respect to *τ_j, and associated to the entropy flux q and for any j,*

$$\frac{d\eta(U_j(t))}{dt} + (1/\Delta x)(q(U_{j+1}(t))) - q(U_j(t)))) \leq 0. \tag{6.3}$$

Proof. (i) The proof is obvious, by the Cauchy–Lipschitz theorem.

(ii) Since F is Lipschitz continuous, by adapting classical results, (see, e.g., [36]), we can show that the Euler approximation (6.2) is stable and consistent and therefore convergent when $\Delta t \to 0$.

On the other hand, since the first order approximation (6.2) coincides with the Godunov scheme (3.2), U_j^n and v_j^n stay in the same bounded invariant region $\mathcal{R}$ and satisfy the L^∞ and BV estimates in x for v (resp. in t for τ and v) as in Theorem 3.1. In the latter case, we slightly modify the proofs of (3.4) and (3.8) to obtain constants C independent of the ratio $\Delta x/\Delta t$, as indicated in Remark 3.1.

(iii) Since for each j the sequence v_j^n converges as $\Delta t \to 0$ to $v_j(t)$ for $t = n\Delta t$, $q(v_j^n)$ is also convergent. On the other hand, since w_j, a_j are constant in the cell I_j, we have for each j

$$\eta(U_j^n) = \eta(\tau_j^n, w_j^n, a_j^n) = \eta(\mathcal{T}(v_j^n, w_j^n, a_j^n), w_j^n, a_j^n) \to \eta(\tau_j(t), w_j(t), a_j(t)).$$

Finally, (iii) is obtained by passing to the limit in the fully discrete entropy inequality (3.10). ∎

6.3. Hydrodynamic limit of the microscopic multiclass model. We rewrite (6.1) in the form

$$\begin{cases} \frac{dU_j(t)}{dt} + \frac{G(U_{j+1}(t)) - G(U_j(t))}{\Delta x} = 0, \\ U_j(0) = U_j^0, \end{cases} \tag{6.4}$$

with $G(U_j) := (-v_j, 0, 0) = (-(w_j - a_j P(\tau_j)), 0, 0)$. We now show that the entropy solution of the system (4.4) is the limit as $\Delta x \to 0$ (i.e., when the number of vehicles goes to infinity) of the solution of the infinite-dimensional system of ordinary differential equations (6.4).

THEOREM 6.2. *Under the same assumptions as in of Theorem* 6.1*, when $\Delta x \to 0$, the* whole *sequence $U_{\Delta x}$ converges in L^∞ weak* (and almost everywhere for the velocity) to the* unique *entropy weak solution of the macroscopic system* (4.4)–(4.6).

Proof. Multiplying (6.4) by an arbitrary test-function $\varphi(x,t)$ and performing a discrete integration by parts (see, e.g., [28]), we have

$$I_{\Delta x} := \int_0^\infty \sum_j \int_{I_j} U_j(t)\partial_t\varphi dx dt - \int_0^\infty \sum_j G(U_j(t)) \left(\int_{I_j} \frac{\varphi(x,t) - \varphi(x-\Delta x, t)}{\Delta x} dx \right) dt + \sum_j \int_{I_j} U_j^0 \varphi(x,0)\, dx = 0.$$

Due to the uniform L^∞ estimates (see Theorem 6.1) at least for a subsequence, $U_{\Delta x}(x,t)$ converges in $L^\infty(\mathbb{R}\times\mathbb{R}_+)$ weak* to some function $U^{**}(x,t)$ when $\Delta x \to 0$. Moreover, by compactness, $G(U_{\Delta x}(.,.)) = (-v_{\Delta x}(.,.), 0, 0) \to (-v^{**}, 0, 0)$ in L^1_{loc}

strongly when $\Delta x \to 0$. Therefore we obtain at the limit

$$\int_0^\infty \int_{\mathbb{R}} U^{**}(x,t)\partial_t\varphi \, dx \, dt + \int_0^\infty \int_{\mathbb{R}} G(U^{**}(x,t))\partial_x\varphi \, dx \, dt + \int_{\mathbb{R}} U_0^*(x)\varphi(x,0) \, dx = 0,$$

which shows that U^{**} is a weak solution of (4.4).

Due to the uniform BV estimates on $v_{\Delta x}$, the Young measure associated with the sequence $(v_{\Delta x}, w_{\Delta x}, a_{\Delta x})$ is still a tensor product $\gamma_{x,t}(v) \otimes \beta_{x,t}(w,a) = \delta(v - v^{**}(x,t)) \otimes \beta_{x,t}(w,a)$ as in (4.3), with the *same* initial data (v_0^*, w_0^*, a_0^*) as in section 4.2, since we have assumed in (4.1) that *all* the sequence (v_h^0, w_h^0, a_h^0) converges to $(v_0^*, \langle \mu_x, w \rangle, \langle \mu_x, a \rangle)$ as $h \to 0$. Therefore $\beta_{x,t}(w,a) \equiv \beta_x(w,a) \equiv \mu_x(w,a)$. Now integrate by parts in the semidiscrete entropy inequality (6.3), $\forall \varphi \geq 0$:

$$\int_0^\infty \sum_j \int_{I_j} \eta(U_j(t))\partial_t\varphi \, dx \, dt + \int_0^\infty \sum_j q(U_j(t)) \left(\int_{I_j} \frac{\varphi(x,t) - \varphi(x - \Delta x, t)}{\Delta x} \, dx \right) dt$$
$$+ \sum_j \int_{I_j} \eta(U_j^0)\varphi(x,0) \, dx \geq 0,$$

and pass to the limit as $\Delta x \to 0$. Note that $q(U_{\Delta x}(\cdot,\cdot)) \to q(U * (\cdot,\cdot))$ strongly, whereas, as in section 4, $\eta(U_{\Delta x}(\cdot,\cdot))$ converges weakly to $\langle \mu_x, \eta(\mathcal{T}(v * *(x,t), w, a) \rangle$. Finally, as in Proposition 5.1, the L^1 equicontinuity in time is preserved for the sequence $\eta(U_{\Delta x})$ when Δx tends to 0. Therefore, the limit v^{**} satisfies (4.8) and the (Kružkov) entropy condition (4.15). Consequently, by the uniqueness result of Theorem 5.2, $v^{**} = v^*$ almost everywhere in (x,t), which also implies that the *whole* sequence converges to the same limit. ∎

In conclusion, starting from the fully discrete system (3.1), we obtain the *same* limit (i.e., the macroscopic system (4.4)), either by letting $(\Delta x, \Delta t) \to 0$ with a fixed ratio and the CFL condition, or by first letting $\Delta t \to 0$ with a fixed Δx, and then letting $\Delta x \to 0$. This last limit process says that the homogenized model (4.4) is the hydrodynamic limit of the microscopic Follow-the-Leader system.

REFERENCES

[1] A. Aw and M. Rascle, *Resurrection of "second order" models of traffic flow*, SIAM J. Appl. Math., 60 (2000), pp. 916–938.

[2] A. Aw, A. Klar, T. Materne, and M. Rascle, *Derivation of continuum traffic flow models from microscopic follow-the-leader models* SIAM J. Appl. Math., 63 (2002), pp. 259–278.

[3] P. Baiti and A. Bressan, *The semigroup generated by a temple class system with large data*, Differential Integral Equations, 10 (1997), pp. 401–418.

[4] P. Baiti and H. K. Jenssen, *Well-posedness for a class of* 2×2 *conservation laws with* L^∞ *data*, J. Differential Equations, 140 (1997), pp. 161–185.

[5] J. M. Ball, *A version of the fundamental theorem for Young measures*, in PDEs and Continuum Models of Phase Transitions, Lecture Notes in Phys. 344, M. Rascle, D. Serre, and M. Slemrod, eds., Springer-Verlag, Berlin, 1989, pp. 207–215.

[6] S. Benzoni-Gavage and R. M. Colombo, *An n-Populations Model for Traffic Flow*, preprint, http://www.umpa.ens-lyon.fr/~benzonil/ or http://dm.ing.unibs.it/rinaldo/, 2002; European J. Appl. Math., to appear.

[7] H. Berliocchi and J.-M. Lasry, *Intégrandes normales et mesures paramétrées en calcul des variations*, Bull. Soc. Math. France, 101 (1973), pp. 129–184.

[8] S. Bianchini, *The semigroup generated by a Temple class system with non-convex flux function*, Differential Integral Equations, 13 (2000), pp. 1529–1550.

[9] S. Bianchini, *Stability of L^∞ solutions for hyperbolic systems with coinciding shocks and rarefactions*, SIAM J. Math. Anal., 33 (2001), pp. 959–981.
[10] M. Bonnefille, *Propagation des oscillations dans deux classes de systèmes hyperboliques (*2×2 *et* 3×3*)*, Comm. Partial Differential Equations, 13 (1988), pp. 905–925.
[11] A. Bressan, *Hyperbolic Systems of Conservation Laws: The One Dimensional Cauchy Problem*, Oxford University Press, Oxford, UK, 2000.
[12] A. Bressan and P. Goatin, *Stability of L^∞ solutions of Temple class systems*, Differential Integral Equations, 13 (2000), pp. 1503–1528.
[13] S. Chanut, *Vers une modélisation macroscopique d'un écoulement bi-fluide poids lourds et véhicules légers*, Ph.D. thesis, ENTPE, Vaulx en Velin, France, in preparation.
[14] G.-Q. Chen, *Propagation and cancellation of oscillations for hyperbolic systems of conservation laws*, Comm. Pure Appl. Math., 44 (1991), pp. 121–140.
[15] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves*, Springer-Verlag, New York, 1976.
[16] C. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, Berlin, 2000.
[17] C. F. Daganzo, *A Behavioral Theory of Multi-lane Traffic Flow. Part* I: *Long Homogeneous Freeway Sections*, Research report UCB-ITS-RR 99-5, Univeristy of California Berkeley, Berkeley, CA, 2002.
[18] R. J. DiPerna, *Measure-valued solutions to conservation laws*, Arch. Ration. Mech. Anal., 88 (1985), pp. 223–270.
[19] E. Weinan and D. Serre, *Correctors for the homogenization of conservation laws with oscillatory forcing terms*, Asymptot. Anal., 5 (1992), pp. 311–316.
[20] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.
[21] L. C. Evans and D. Gomes, *Effective Hamiltonians and averaging for Hamiltonian dynamics* I, Arch. Ration. Mech. Anal., 157 (2001), pp. 1–33.
[22] D. C. Gazis, R. Herman, and R. W. Rothery, *Nonlinear follow-the-leader models of traffic flow*, Oper. Res., 9 (1961), pp. 545–567.
[23] J. M. Greenberg, *Extensions and amplifications of a traffic model of Aw and Rascle*, SIAM J. Appl. Math., 62 (2001), pp. 729–745.
[24] R. Herman and I. Prigogine, *Kinetic Theory of Vehicular Traffic*, American Elsevier, New York, 1971.
[25] R. A. Klausen and N. H. Risebro, *Stability of conservation laws with discontinuous coefficients*, J. Differential Equations, 157 (1999), pp. 41–60.
[26] C. Klingenberg and N. Risebro, *Convex conservation laws with discontinuous coefficients. Existence, uniqueness and asymptotic behavior*, Comm. Partial Differential Equations, 20 (1995), pp. 1959–1990.
[27] S. N. Kruzkov, *First order quasilinear equations with several independent variables*, Mat. USSR Sb., 10 (1970), pp. 217–243.
[28] R. J. LeVeque, *Numerical Methods for Conservation Laws*, 2nd ed., Birkhäuser Verlag, Basel, 1992.
[29] P. L. Lions, G. Papanicolaou, and S. R. S. Varadhan, *Homogenisation of Hamilton-Jacobi Equations*, manuscript.
[30] J. Málek, J. Nečas, M. Rokyta, and M. Růžička, *Weak and Measure-Valued Solutions to Evolutionary PDEs*, Appl. Math. Math. Comput. 13, Chapman and Hall, London, 1996.
[31] F. Murat, *L'injection du cône positif de H^{-1} dans W^{-1}, q est compacte pour tout $q<2$*, J. Math. Pures Appl. (9), 60 (1981), pp. 309–322.
[32] G. Namah and J. M. Roquejoffre, *Remarks on the long time behaviour of the solutions of Hamilton-Jacobi equations*, Comm. Partial Differential Equations, 24 (1999), pp. 883–893.
[33] H. J. Payne, *Models of Freeway Traffic and Control*, Simulation Council, 1971.
[34] M. Rascle, *On the static and dynamic study of oscillations for some nonlinear hyperbolic systems of conservation laws*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 8 (1991), pp. 333–350.
[35] S. Schochet, *Examples of measure-valued solutions*, Comm. Partial Differential Equations, 14 (1989), pp. 545–575.
[36] M. Schatzman, *Analyse numérique*, InterEditions, Paris, 1991.
[37] N. Seguin and J. Vovelle, *Analysis and approximation of a scalar consevation law with a flux function with discontinuous coefficients*, Math. Models Methods Appl. Sci., 13 (2003), pp. 221–257.
[38] D. Serre, *Systemes de lois de conservation* I *et* II, Diderot Editeur, Paris, 1996.

[39] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.
[40] A. Szepessy, *An existence result for scalar conservation laws using measure valued solutions*, Comm. Partial Differential Equations, 14 (1989), pp. 1329–1350.
[41] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, Volume IV, Res. Notes in Math. 39, Pitman, Boston, 1979, pp. 136–212.
[42] J. Blake Temple, *Systems of conservation laws with coinciding shock and rarefaction curves*, Contemp. Math., 17 (1983), pp. 143–151.
[43] L. Tong, *Well-posedness theory of an inhomogeneous traffic flow model*, Discrete Contin. Dyn. Syst. Ser. B, 2 (2002), pp. 401–414.
[44] J. D. Towers, *Convergence of a difference scheme for conservation laws with a discontinuous flux*, SIAM J. Numer. Anal., 38 (2000), pp. 681–698.
[45] A. Vasseur, *private communication*.
[46] D. H. Wagner, *Equivalence of the Euler and Lagrangian equations of gas dynamics for weak solutions*, J. Differential Equations, 68 (1987), pp. 118–136.
[47] G. B. Whitham, *Linear and Nonlinear Waves*, Pure and Applied Mathematics, Wiley-Interscience, New York, 1974.
[48] G. C. K. Wong and S. C. Wong, *A multi-class traffic flow model—an extension of LWR model with heterogeneous drivers*, Transportation Research Part A, 36 (2002), pp. 827–841.
[49] H. M. Zhang, *A non-equilibrium traffic model devoid of gas-like behavior*, Transportation Research Part B, 36 (2002), pp. 275–90.

SIAM J. MATH. ANAL.
Vol. 35, No. 4, pp. 974–986

A FREE BOUNDARY PROBLEM FOR AN ELLIPTIC-HYPERBOLIC SYSTEM: AN APPLICATION TO TUMOR GROWTH*

XINFU CHEN† AND AVNER FRIEDMAN‡

Abstract. We consider a system of two hyperbolic equations for p, q and two elliptic equations for c, σ, where p, q are the densities of cells within the tumor Ω_t in proliferating and quiescent states, respectively, c is the concentration of nutrients, and σ is the pressure. The pressure is a result of the transport of cells which proliferate or die. The motion of the free boundary $\partial\Omega_t$ is given by the continuity condition, and σ at the free boundary is proportional to the surface tension. We prove the existence, uniqueness, and regularity of the solution for a small time interval $0 \le t \le T$.

Key words. free boundary problem, elliptic-hyperbolic system, tumor growth

AMS subject classifications. 35R35, 35M10, 35Q80, 92D99

DOI. 10.1137/S0036141002418388

1. The model. In this paper we consider a free boundary problem for an elliptic-hyperbolic system which describes the evolution of a tumor. The cells within the tumor are in one of three states: proliferating, quiescent, or necrotic. We shall denote the corresponding cell densities by p, q, and n, respectively. Quiescent cells become proliferating at a rate $K_P(c)$, where $K_P(c)$ is a positive-valued function of the nutrient concentration c, and they become necrotic at another rate $K_D(c)$; $K_P(c)$ is a monotone increasing function of c and $K_D(c)$ is a monotone decreasing function of c, although these monotonicity properties will not be assumed in this paper. Proliferating cells become quiescent at a rate $K_Q(c)$, and they proliferate at a rate $K_B(c)$, where $K_Q(c)$ is monotone decreasing and $K_B(c)$ is monotone increasing in c, properties which again will not be assumed in this paper. Finally, necrotic cells are removed from the tumor at a constant rate K_R. We assume that the nutrient concentration c satisfies a diffusion equation $\Delta c - \lambda c = 0$, where λ is a positive constant.

Due to proliferation and removal of cells, there is a continuous motion of cells within the tumor. We shall represent this movement by a velocity field $\vec{v}$. We can then write the conservation of mass laws for the densities of the proliferating cells p, the quiescent cells q, and the necrotic cells n within the tumor region Ω_t in the following form:

$$\frac{\partial p}{\partial t} + \operatorname{div}(p\vec{v}) = [K_B(c) - K_Q(c)]p + K_P(c)q,$$

$$\frac{\partial q}{\partial t} + \operatorname{div}(q\vec{v}) = K_Q(c)p - [K_P(c) + K_D(c)]q,$$

$$\frac{\partial n}{\partial t} + \operatorname{div}(n\vec{v}) = K_D(c)q - K_R n.$$

*Received by the editors November 21, 2002; accepted for publication (in revised form) May 23, 2003; published electronically November 4, 2003.
http://www.siam.org/journals/sima/35-4/41838.html

†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (xinfu@pitt.edu). This author was supported by National Science Foundation grant DMS-0203991.

‡Department of Mathematics, Ohio State University, 231 West 18th Avenue, Columbus, OH 43210 (afriedman@mbi.osu.edu). This author was supported by National Science Foundation grant DMS-0098520.

We shall make the assumption that the tumor tissue is a porous medium so that, by Darcey's law,

$$\vec{v} = -\nabla\sigma, \quad \sigma = \text{pressure}.$$

We also assume that all cells are of the same volume and density and that the total density of cells is uniform throughout the tumor. Then

$$p + q + n = \text{const.} = B,$$

and, without loss of generality, we take $B = 1$. If we add the equations for p, q, and n, we get $\text{div}\vec{v} = K_B(c)p - K_R n$; this equation can be used to replace the conservation law for n. If we also substitute $n = 1 - p - q$ in the equation for $\text{div}\vec{v}$, we obtain the system of equations

$$\Delta c - \lambda c = 0 \text{ in } \Omega_t, \tag{1.1}$$

$$\frac{\partial p}{\partial t} + \nabla\sigma \cdot \nabla p = f(c,p,q) \text{ in } \Omega_t, \tag{1.2}$$

$$\frac{\partial q}{\partial t} + \nabla\sigma \cdot \nabla q = g(c,p,q) \text{ in } \Omega_t, \tag{1.3}$$

$$\Delta\sigma = -h(c,p,q) \text{ in } \Omega_t, \tag{1.4}$$

where

$$\begin{aligned} f(c,p,q) &= [K_B(c) - K_Q(c)]p + K_P(c)q + h(c,p,q)p, \\ g(c,p,q) &= K_Q(c)p - [K_P(c) + K_D(c)]q + h(c,p,q)q, \\ h(c,p,q) &= -K_R + [K_B(c) + K_R]p + K_R q. \end{aligned} \tag{1.5}$$

We take the boundary conditions to be

$$c = 1 \text{ on } \Gamma_t, \tag{1.6}$$

$$\sigma = \gamma\kappa \text{ on } \Gamma_t, \tag{1.7}$$

$$\frac{\partial\sigma}{\partial n} = -V_n \text{ on } \Gamma_t, \tag{1.8}$$

where $\Gamma_t = \partial\Omega_t, \gamma$ is a positive constant, κ is the mean curvature, $\frac{\partial}{\partial n}$ is the derivatives in the direction $\vec{n}$ of the outward normal, and V_n is the velocity of the free boundary Γ_t in the direction $\vec{n}$. The condition (1.7) is based on the assumption that the pressure σ on the surface of the tumor is proportional to the surface tension (see Greenspan [12]), and the condition (1.8) is a standard kinetic (or continuity) condition.

Finally, we supplement the above system with initial conditions:

$$p(x,0) = p_0(x), \quad q(x,0) = q_0(x) \text{ in } \Omega_0, \Gamma_0 \text{ is given}, \tag{1.9}$$

where

$$p_0(x) \geq 0, \quad q_0(x) \geq 0, \quad p_0(x) + q_0(x) \leq 1; \tag{1.10}$$

here Ω_0 is a bounded domain with boundary Γ_0.

The model (1.1)–(1.10) was introduced in [14] in the case where the initial data and the solution are spherically symmetric. Existence of a global solution, in this case, was proved by Cui and Friedman [6]; they also established uniform positive bounds from above and below for the radius $R(t)$ of the tumor.

The special case, where the system consists only of the equation

$$\Delta\sigma = 0$$

with the boundary conditions (1.7), (1.8), is known as the Hele–Shaw problem, or the quasi-static Stefan problem. In this case, given any initial domain Ω_0, there exists a unique local (in time) solution. There are several proofs of this result; see Chen [3], Duchon and Robert [7], and Constantin and Pugh [5] for the two-dimensional case, and Bazaliy [1], Bazaliy and Friedman [2], Chen, Hong, and Yi [4], and Escher and Simonett [9, 10] for the N-dimensional case ($N \geq 2$). Global existence and asymptotic stability for nearby spherical initial data were proved in the two-dimensional case by Chen [3] and for the N-dimensional case ($N \geq 2$) by Escher and Simonett [10] and Friedman and Reitich [11].

In the more general case

$$-c_t + \Delta c - \lambda c = 0, \tag{1.11}$$

$$\Delta\sigma = -h(c) \tag{1.12}$$

with boundary conditions (1.6)–(1.8), local existence and uniqueness were proved by Bazaliy and Friedman [2] and subsequently also by Escher [8] (under minimal regularity assumptions).

In this paper we prove local existence and uniqueness for the system (1.1)–(1.10) under minimal regularity assumptions. As will be pointed out in Remark 3.2, our proof extends to the case where (1.1) is replaced by (1.11).

One can easily show, using the special forms of f, g, h, that for any solution of (1.1)–(1.10) there holds

$$p \geq 0, \quad q \geq 0, \quad p + q \leq 1,$$

and clearly also $0 \leq c \leq 1$. In what follows we shall not use the special forms of f, g, h and instead assume only that

$$f, g, h \in C^{m+1} \quad \text{for some } m \text{ integer} \geq 0 \tag{1.13}$$

and, without loss of generality, that

$$f, g, h \text{ vanish if } |c| + |p| + |q| \text{ is sufficiently large.} \tag{1.14}$$

Since our results apply equally well in any number of dimensions, we shall take the domains Ω_t to be N-dimensional, $N \geq 2$.

The main result of this paper, namely, the existence and regularity of a local solution to (1.1)–(1.4), (1.6)–(1.9), is stated (Theorem 3.1) and proved in section 3. The proof depends upon viewing the solution of the system as a fixed point for a nonlinear mapping W. In section 2 we prove that W is regularizing (Theorem 2.3). The proof depends on (i) extending results for the Hele–Shaw problem to an "inhomogeneous" Hele–Shaw problem (Theorem 2.1), and (ii) regularity properties of solutions of hyperbolic systems (Lemma 2.2).

2. A bootstrap argument. For any vector $\beta = (\beta_0, \beta_1, \ldots, \beta_n)$, β_i integers ≥ 0, set $|\beta| = \beta_0 + \beta_1 + \cdots + \beta_N$ and

$$D^\beta \varphi = D^\beta_{(x,t)} \varphi = \frac{\partial^{|\beta|} \varphi}{(\partial t)^{\beta_0} (\partial x_1)^{\beta_1} \ldots (\partial x_N)^{\beta_N}}.$$

For any $0 < \alpha_1 < 1$, $0 < \alpha_2 < 1$, $0 < \alpha < 1$, m integer ≥ 0, define

$$\|\varphi\|_0 = \sup |\varphi|, \|\varphi\|_m = \sum_{|\beta| \leq m} \|D^\beta \varphi\|_0,$$

$$[\varphi]_{\alpha_1, \alpha_2} = \sup \frac{|\varphi(x,t) - \varphi(y,\tau)|}{|x-y|^{\alpha_1} + |t-\tau|^{\alpha_2}},$$

$$\|\varphi\|_{m+\alpha_1, m+\alpha_2} = \|\varphi\|_0 + \sum_{|\beta|=m} [D^\beta \varphi]_{\alpha_1, \alpha_2}, \tag{2.1}$$

$$\|\varphi\|_{3+\alpha, (3+\alpha)/3} = \|\varphi\|_0 + [D_x^3 \varphi]_{\alpha, \alpha/3} + [D_t \varphi]_{\alpha, \alpha/3}, \tag{2.2}$$

$$\|\varphi\|_{2+\alpha, (2+\alpha)/3} = \|\varphi\|_{0, \frac{2+\alpha}{3}} + [D_x^2 \varphi]_{\alpha, \alpha/3}. \tag{2.3}$$

We note that the right-hand side of (2.1) dominates $\|\varphi\|_m$, so that by adding this norm to the right-hand side of (2.1) we obtain a new norm which is equivalent to $\|\varphi\|_{m+\alpha_1, m+\alpha_2}$. Similarly, the right-hand side of (2.2) dominates

$$[D_x^2 \varphi]_{0, \frac{1+\alpha}{3}} + [D_x \varphi]_{0, \frac{2+\alpha}{3}}$$

and the right-hand side of (2.3) dominates

$$[D_x \varphi]_{1+\alpha, \frac{1+\alpha}{3}}.$$

We say that $\varphi \in C^{m+\alpha_1, m+\alpha_2}$ if $\|\varphi\|_{m+\alpha_1, m+\alpha_2} < \infty$. Similarly we define the notions $\varphi \in C^{3+\alpha, (3+\alpha)/3}$, $\varphi \in C^{2+\alpha, (2+\alpha)/3}$.

In what follows we assume that

$$\Gamma_0 \in C^{m+4+\alpha}, \tag{2.4}$$

where $0 < \alpha < 1$ and m is an integer ≥ 0. Denote by s a variable point in Γ_0 and by $\vec{n}(s)$ the unit outward normal to Γ_0 at s. We shall write Γ_t in the form

$$\Gamma_t = \{s + \rho(s,t) \vec{n}(s)\}.$$

Set $d = d(x) = d(x, \Gamma_0) =$ signed distance from x to Γ_0. Then, for x near Γ_0, we can write

$$x = s + d\vec{n}(s),$$

where s is uniquely determined by x.

In what follows we shall use a local coordinate transformation to flatten the boundary Γ_t. The procedure is as follows: Take local coordinates $y' = (y_1, \dots, y_{n-1})$ near 0 in $\mathbb{R}^{N-1}$, about a point s_0 in Γ_0, so that $s = S(y_1, \dots, y_{N-1})$ for $|s - s_0|$ small.

Then, for x near s_0,

$$x = S(y_1, \dots, y_{N-1}) + (\rho(s,t) + y_N)\vec{n}(S(y_1, \dots, y_{N-1})),$$

where $y_N = d(x, \Gamma_0) - \rho(s,t)$. This defines a local mapping $y \to x$ from a neighborhood of the origin in $\mathbb{R}^N$ into an $\mathbb{R}^N$-neighborhood of s_0 such that $x \in \Gamma_t$ corresponds to $(y', 0)$.

Assume that

$$c, p, q \in C^{m+\alpha, m+\alpha/3}(\mathbb{R}^N \times [0,T]) \tag{2.5}$$

and consider the "inhomogeneous" Hele–Shaw problem: Find σ and $\{\Omega_t; 0 \le t \le T\}$ such that

$$\Delta\sigma = -h(c,p,q) \equiv -h(x,t) \text{ in } \Omega_t, \tag{2.6}$$

$$\sigma = \gamma\kappa, \quad \frac{\partial\sigma}{\partial n} = V_n \text{ on } \Gamma_t$$

for $0 \le t \le T$ with $\rho_t|_{t=0} = r_0$. By (1.14), $h(x,t) \in C^{m+\alpha, m+\alpha/3}(\mathbb{R}^N \times [0,T])$.

THEOREM 2.1. *If T is sufficiently small, then the system* (2.6) *has a unique solution for $0 \le t \le T$ with*

$$D_s D^m_{(s,t)}\rho \in C^{3+\alpha,(3+\alpha)/3}(\mathbb{R}^{N-1} \times [0,T]), \tag{2.7}$$

$$D_x^2 D^m_{(x,t)}\sigma \in C^{\alpha,\alpha/3}(\mathbb{R}^N \times [0,T]).$$

Note that the case $h \equiv 0$ is the Hele–Shaw problem mentioned in section 1. The proof of Theorem 2.1 to be given here is an extension of the proof for the Hele–Shaw problem of Chen, Hong, and Yi [4].

Proof. The proof of existence and uniqueness of a solution with $\rho \in C^{3+\alpha,(3+\alpha)/3}$ follows by the same arguments as in [4]. Indeed, using the local mapping $x \to y$ described above and the Hanzawa transformation [13], we obtain a model problem which is the same as in the case $h \equiv 0$. The same basic Lemma 3.4 of [4] can then be applied to deduce existence, uniqueness, and Hölder norm estimates for the model problem and, subsequently, for the linearized problem (as in [4, section 4]). The completion of the proof with $\rho \in C^{3+\alpha,(3+\alpha)/3}$ then proceeds exactly as in [4].

We shall now establish the regularity assertions of (2.7). Denote by G the fundamental solution of the Laplace operator. Since h has a compact support, the convolution $G * h$ is well defined, and $D_x^2(G * h) \in C^{m+\alpha,m+\alpha/3}(\mathbb{R}^N \times [0,T])$. Set

$$\widetilde{\sigma} = \sigma + G * h.$$

Then

$$\Delta\widetilde{\sigma} = 0 \text{ in } \Omega_t,$$

$$\widetilde{\sigma} = \gamma\kappa + G * h, \quad V_n = \frac{\partial\widetilde{\sigma}}{\partial n} - \vec{n} \cdot (\nabla G * h) \text{ on } \Gamma_t.$$

We fix s_0 in Γ_0 and rewrite (2.6) in terms of the variable y:

$$\sum_{i,j=1}^{N}(a_{ij}^{\rho}\widetilde{\sigma}_{y_i})_{y_j}+\sum_{i=1}^{N}a_i^{\rho}\widetilde{\sigma}_{y_i}=0 \quad \text{in } \{y_N>0\},$$

$$\widetilde{\sigma}=\sum_{i,j=1}^{N-1}b_{ij}^{\rho}\rho_{y_iy_j}+b^{\rho}+G*h \text{ on } \{y_N=0\}, \tag{2.8}$$

$$\rho_t=\sum_{i=1}^{N}\ell_i^{\rho}\frac{\partial}{\partial y_i}(\widetilde{\sigma}-G*h) \text{ on } \{y_N=0\},$$

where

$$a_{ij}^{\rho}=a_{ij}(D_{y'}\rho), a_i^{\rho}=a_i(D_{y'}^2\rho), b_{ij}^{\rho}=b_{ij}(D_{y'}\rho), \tag{2.9}$$
$$b^{\rho}=b(D_{y'}\rho), \ell_i^{\rho}=\ell_i(D_{y'}\rho).$$

In particular, at $\rho=0, y'=0, y_N=0$

$$a_{ij}=\delta_{ij}, \quad a_i=0, \quad b_{ij}=\delta_{ij}, \quad b=\gamma\kappa(s_0), \quad \ell_i=\delta_{Ni}.$$

For $\beta=(\beta_0,\beta_1,\dots,\beta_N)$ with $|\beta|\le k+1$, $k\le m'$, $\beta_0\le k$, $\beta_N=0$ we formally have

$$\Sigma[a_{ij}^{\rho}(D^{\beta}\widetilde{\sigma}_{y_i})]_{y_j}=\mathrm{div}\vec{F}_1 \text{ in } \{y_N>0\},$$

$$D^{\beta}\widetilde{\sigma}=\Sigma b_{ij}^{\rho}(D^{\beta}\rho)_{y_iy_j}+F_2 \text{ in } \{y_N=0\}, \tag{2.10}$$
$$(D^{\beta}\rho)_t+\Sigma\ell_i^{\rho}\frac{\partial}{\partial y_i}(D^{\beta}\widetilde{\sigma})+F_3 \text{ in } \{y_N=0\},$$

where

$$\vec{F}_1=\vec{F}_1(D_yD^k\sigma, D_{y'}^2D^k\rho), \quad F_2=F_2(D_xD^kG*h, D_{y'}^2D^k\rho),$$
$$F_3=F_3(D_x^2D^kG*h, D_{y'}^2D^k\rho, D^k\sigma);$$

here we used abbreviations such as $D^k\sigma=\{D^{\lambda}\sigma;|\lambda|\le k\}$ to describe the structure of $\vec{F}_1, F_j$.

The system (2.10) is linear in $(D^{\beta}\widetilde{\sigma}, D^{\beta}\rho)$. By the Hölder estimates of the model problem [4, Lemma 3.4] we deduce that if

$$D^{\beta}\rho\in C^{3+\alpha,(3+\alpha)/3}, \quad D_yD^{\beta}\widetilde{\sigma}\in C^{\alpha,\alpha/3}$$

for all β as above, then

$$D_sD_{(s,t)}^{m'}\rho\in C^{3+\alpha,(3+\alpha)/3}, \quad D_x^2D_{(x,t)}^{m'}\widetilde{\sigma}\in C^{\alpha,\alpha/3}. \tag{2.11}$$

This result allows us to establish (2.11) inductively for all $m'\le m$ (since $\Gamma_0\in C^{m+4+\alpha}$ and $D_x^2(G*h)\in C^{m+\alpha,m+\alpha/3}$). □

Consider next the elliptic problem

$$\Delta\overline{c} - \lambda\overline{c} = 0 \text{ in } \Omega_t, \tag{2.12}$$
$$\overline{c} = 1 \text{ on } \Gamma_t.$$

In terms of the local coordinates y about $s_0 \in \Gamma_0$, we can write (2.12) in the form

$$\Sigma\widehat{a}_{ij}\overline{c}_{y_iy_j} + \Sigma\widehat{b}_i\overline{c}_{y_i} - \lambda\overline{c} = 0 \text{ in } \{y_N > 0\}, \tag{2.13}$$
$$\overline{c} = 1 \text{ on } \{y_N = 0\},$$

where

$$\widehat{a}_{ij} = \widehat{a}_{ij}(D_s\rho), \widehat{b}_i = \widehat{b}_i(D_s^2\rho). \tag{2.14}$$

Let $\beta = (\beta_0, \beta_1, \dots, \beta_{N-1}, 0)$, $|\beta| \le m' + 1$, $\beta_0 \le m'$. Then, formally,

$$\Sigma\widehat{a}_{ij}(D^\beta\overline{c})_{y_iy_j} = F(D_s^3D^{m'}\rho, D_y^2D^{m'}\overline{c}) \text{ in } \{y_N > 0\}, \tag{2.15}$$
$$D^\beta\overline{c} = 0 \text{ on } \{y_N = 0\},$$

where F is a smooth function in its variables.

Using Theorem 2.1 and elliptic theory, we deduce inductively on m' ($m' = 0, 1, \dots, m$) that the right-hand side F and its first x- and t-derivations are in $C^{\alpha,\alpha/3}$ and

$$D_xD^mD_x^2\overline{c}, D_xD^mD_t\overline{c} \quad \text{belong to } C^{\alpha,\alpha/3}. \tag{2.16}$$

We can now extend $\overline{c}$ into $\mathbb{R}^N \times [0, \mathrm{T}]$ (along normals to Γ_t, with a cutoff function) so that

$$\overline{c} \in C^{m+1+\alpha,\ m+1+\alpha/3}(\mathbb{R}^N \times [0, T]). \tag{2.17}$$

Remark 2.1. Suppose that instead of (2.12) we consider the parabolic problem

$$-\frac{\partial\overline{c}}{\partial t} + \Delta\overline{c} - \lambda\overline{c} = 0 \text{ in } \Omega_t, \tag{2.18}$$
$$\overline{c} = 1 \text{ on } \Gamma_t,$$
$$\overline{c}\ (x, 0) = c_0(x) \text{ in } \Omega_0.$$

If $c_0(x)$ is in $C^{m+1+\alpha}$ and it satisfies the compatibility conditions (at $\partial\Omega_0$) to order $m + 1$, then the above procedure leads to (2.16) with $C^{\alpha,\alpha/3}$ replaced by $C^{2\alpha/3,\alpha/3}$, so that (2.17) is again valid.

From (2.7) it follows that we can extend $\sigma(x, t)$ into $\mathbb{R}^N \times [0, T]$ (along normals to Γ_t, with a cutoff function) so that

$$D_x^2D^m\sigma \in C^{\alpha,\alpha/3}(\mathbb{R}^N \times [0, T]). \tag{2.19}$$

We now turn to the hyperbolic system

$$\overline{p}_t + \nabla\sigma \cdot \nabla\overline{p} = f(c, \overline{p}, \overline{q}) \quad \text{in } \mathbb{R}^N \times [0, T],$$
$$\overline{q}_t + \nabla\sigma \cdot \nabla\overline{q} = g(c, \overline{p}, \overline{q}) \quad \text{in } \mathbb{R}^N \times [0, T], \tag{2.20}$$
$$\overline{p}|_{t=0} = p_0,\ \overline{q}|_{t=0} = q_0 \quad \text{in } \mathbb{R}^N$$

and assume that

$$p_0, q_0 \in C^{m+1+\alpha}(\mathbb{R}^N). \tag{2.21}$$

Formally, for $\beta = (\beta_0, \beta_1, \dots, \beta_N)$, $|\beta| \le m$, we have

$$\begin{aligned} (D^\beta \overline{p})_t + \nabla\sigma \cdot \nabla(D^\beta \overline{p}) &= \overline{F}_1, \\ (D^\beta \overline{q})_t + \nabla\sigma \cdot \nabla(D^\beta \overline{q}) &= \overline{F}_2, \end{aligned} \tag{2.22}$$

where

$$\overline{F}_1 = D^\beta f - \sum_{\substack{|\mu|\ge 1 \\ \mu \le \beta}} \binom{\beta}{\mu} \nabla D^\mu \sigma \cdot \nabla D^{\beta-\mu} \overline{p},$$

$$\overline{F}_2 = D^\beta g - \sum_{\substack{|\mu|\ge 1 \\ \mu \le \beta}} \binom{\beta}{\mu} \nabla D^\mu \sigma \cdot \nabla D^{\beta-\mu} \overline{q}.$$

We shall need the following lemma.

LEMMA 2.2. *Consider a hyperbolic system of two equations,*

$$\begin{aligned} &\vec{w}_t + (\vec{b} \cdot \nabla_x)\vec{w} = G(x,t,\vec{w}) \text{ in } \mathbb{R}^N \times [0,T], \\ &\vec{w}|_{t=0} = \vec{w}_0 \text{ in } \mathbb{R}^N, \end{aligned} \tag{2.23}$$

where $\vec{w} = (w_1, w_2)$, *and assume that*

$$\begin{gathered} D_x\vec{b}, D_x G \in C^{\alpha_1,\alpha_2}(\mathbb{R}^N \times [0,T]), \\ D_{\vec{w}} G \in L^\infty(\mathbb{R}^N \times [0,T]), \\ D_x \vec{w}_0 \in C^{\alpha_1}(\mathbb{R}^N). \end{gathered}$$

Then there exists a unique solution of (2.23) *such that*

$$\vec{w}_t, D\vec{w} \text{ in } C^{\alpha_1,\alpha_2}(\mathbb{R}^N \times [0,T]).$$

Using (2.19) we can apply Lemma 2.2 inductively to deduce that

$$D^\beta \overline{p}_t, D^\beta D_x \overline{p}, D^\beta \overline{q}_t, D^\beta D_x \overline{q} \in C^{\alpha,\alpha/3}$$

for $|\beta| \le m$. Hence

$$\overline{p}, \overline{q} \in C^{m+1+\alpha,\ m+1+\alpha/3}. \tag{2.24}$$

Proof. We introduce the characteristic curves $X = X(\xi,t)$,

$$\frac{\partial X}{\partial t} = \vec{b}(X,t), \quad X(\xi,0) = \xi,$$

and the function $\vec{U}(x,t) = \vec{w}(X(x,t),t)$. Then

$$\frac{d\vec{U}}{dt} = \vec{G}(x,t,\vec{U}), \quad \vec{U}(x,0) = \vec{w}_0(x).$$

Denote by $\xi = \xi(\cdot, t)$ the inverse of $X = X(\cdot, t)$, i.e., $x = X(\xi(x,t), t)$. Then

$$\vec{w}(x,t) = \vec{U}(\xi(x,t), t).$$

Since

$$\frac{d}{dt}\left(\frac{\partial X}{\partial \xi}\right) = D_x \vec{b}\, \frac{\partial X}{\partial \xi}, \quad 0 \le t \le T,$$

$$\frac{\partial X}{\partial \xi}|_{t=0} = I,$$

we have

$$\left\| \frac{\partial X}{\partial \xi} \right\|_{L^\infty}, \left\| \frac{\partial}{\partial t} \frac{\partial X}{\partial \xi} \right\|_{L^\infty} \le C.$$

From the relation

$$\frac{d}{dt}\left(\left(\frac{\partial X}{\partial \xi}\right)^{-1}\right) = -\left(\frac{\partial X}{\partial \xi}\right) D_x \vec{b}$$

we also have

$$\left\| \frac{\partial \xi}{\partial X} \right\|_{L^\infty} \le C.$$

Similarly,

$$\left\| \frac{\partial \vec{U}}{\partial \xi} \right\|_{L^\infty}, \left\| \frac{\partial}{\partial t} \frac{\partial \vec{U}}{\partial \xi} \right\|_{L^\infty} \le C.$$

We next establish Hölder estimates. For $\xi_1 \ne \xi_2$,

$$\frac{d}{dt}(X_\xi(\xi_1,t) - X_\xi(\xi_2,t)) = D_x \vec{b}(X(\xi_1,t))(X_\xi(\xi_1,t) - X_\xi(\xi_2,t))$$
$$+ (D_x \vec{b}(X(\xi_1,t)) - D_x \vec{b}(X(\xi_2,t)))X_\xi(\xi_2,t).$$

Noting that

$$|D_x \vec{b}(X(\xi_2,t)) - D_x\, \vec{b}\,(X(\xi_1,t),t)| \le |D_x \vec{b}|_{\alpha,0} |X(\xi_2,t) - X(\xi_1,t)|$$
$$\le C|\xi_2 - \xi_1|^{\alpha_1},$$

we deduce that

$$|X_\xi(\xi_1,t) - X_\xi(\xi_2,t)| \le C|\xi_1 - \xi_2|^{\alpha_1}.$$

Similarly

$$|X_\xi(\xi_1,t_1) - X_\xi(\xi_2,t_2)| \le C|\xi_1 - \xi_2|^{\alpha_1} + |t_1 - t_2|^{\alpha_2}.$$

From the relations

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}, \quad \xi_x = X_\xi^{-1}$$

we then get

$$|\xi_x(x_1,t_1)-\xi_x(x_2,t_2)|\le C(|\xi_1-\xi_2|^{\alpha_1}+|t_1-t_2|^{\alpha_2}).$$

Since

$$|\xi_2-\xi_1|=|\xi(x_1,t_1)-\xi(x_2,t_2)|\le C(|x_1-x_2|+|t_1-t_2|),$$

we conclude that

$$|\xi_x(x_1,t_1)-\xi_x(x_2,t_2)|\le C(|x_1-x_2|^{\alpha_1}+|t_1-t_2|^{\alpha_2}).$$

Similarly one can prove that

$$\left|\frac{\partial\vec{U}}{\partial\xi}(\xi_1,t_1)-\frac{\partial\vec{U}}{\partial\xi}(\xi_2,t_2)\right|\le C(|\xi_1-\xi_2|^{\alpha_1}+|t_1-t_2|^{\alpha_2}).$$

Hence

$$\vec{w}_x=\vec{U}_\xi\xi_x\in C^{\alpha_1,\alpha_2},\quad \vec{w}_t=\vec{G}-(\vec{b}\cdot\nabla_x)\vec{w}\in C^{\alpha_1,\alpha_2}.\qquad \square$$

We summarize in the following theorem the results obtained in Theorem 2.1 and (2.17), (2.24).

THEOREM 2.3. *Given* Γ_0, p_0, q_0 *as in* (2.4), (2.21), *there exists a unique solution* $\{\sigma,\Omega_t\}_{0\le t\le T}$ *of* (2.6) *and a unique solution* $(\overline{c},\overline{p},\overline{q})$ *of* (2.12), (2.20) *such that* (2.7), (2.17), *and* (2.24) *hold.*

3. The main result. We shall henceforth assume that

$$\Gamma_0\in C^{m+1+\alpha},\quad p_0 \text{ and } q_0 \text{ belong to } C^{m+1+\alpha}(\overline{\Omega}_0), \tag{3.1}$$

where $m\ge 0$, $0<\alpha<1$. We can then extend p_0,q_0 so that they satisfy (2.21) and have compact supports.

Theorem 2.3 shows that the mapping

$$(c,p,q)\overset{W}{\to}(\overline{c},\overline{p},\overline{q})$$

is regularizing: It maps $C^{m+\alpha,\ m+\alpha/3}(\mathbb{R}^N\times[0,T])$ into $C^{m+1+\alpha,\ m+1+\alpha/3}(\mathbb{R}^N\times[0,T])$. Recall that this mapping was defined by first solving the "inhomogeneous" Hele–Shaw problem (2.6), thus determining $\{\sigma,\Omega_t;0\le t\le T\}$ as in Theorem 2.1, and then solving the elliptic and hyperbolic problems for $\overline{c}$ and $(\overline{p},\overline{q})$, respectively.

Consider now a set Y_M of functions (c,p,q) such that

$$\|(c,p,q)\|_{m+\alpha,\ m+\alpha/3,\ (\mathbb{R}^N\times[0,T])}\le M,$$

$$p|_{t=0}=p_0,\quad q|_{t=0}=q_0,$$

where M is a positive number to be chosen later on.

From the proof of Theorem 2.3 one can see that

$$\|(\overline{c},\overline{p},\overline{q})\|_{m+1+\alpha,\ m+1+\alpha/3,\ (\mathbb{R}^N\times[0,T])}\le C(M),$$

where $C(M)$ is a constant depending on M. Writing $w(x,t)=w(x,0)+\int_0^t w_t(x,t)dt$ for $w=\overline{c},\overline{p},\overline{q}$ and using the facts that

$$0\le\overline{c}\le 1\quad\text{and}\quad \overline{p}|_{t=0}=p_0,\overline{q}|_{t=0}=q_0,$$

we deduce that

$$\|(\overline{c},\overline{p},\overline{q})\|_{m+\alpha,\ m+\alpha/3,\ (\mathbb{R}^N\times[0,T])} \le B(A+T\ C(M)),$$

where

$$A = 1 + \|(p_0,q_0)\|_{C^{m+\alpha}(\mathbb{R}^N)}$$

and B is a universal constant. Choosing $M = BA+1$ we conclude that W maps Y_M into a compact subset of itself, provided T is so small that $BTC(M) < 1$.

We claim that W is also a contraction. To prove this we take two points

$$(c_1,p_1,q_1),\ (c_2,p_2,q_2) \text{ in } Y_M$$

and set

$$h_i = h(c_i,p_i,q_i) \quad (i=1,2),$$
$$\widehat{h} = h_1 - h_2.$$

Denote by $(\sigma_i,\Omega_{i,t},0\le t\le T\}$ the solution of (2.6) corresponding to (c_i,p_i,q_i) and by $(\overline{c}_i,\overline{p}_i,\overline{q}_i)$ the corresponding solutions of (2.12), (2.20). We shall first estimate the difference between the corresponding distance functions $\rho_i(s,t)$, using the same local coordinates y for $i=1$ and $i=2$.

Setting

$$\widehat{\sigma} \equiv \widetilde{\sigma}_1 - \widetilde{\sigma}_2, \quad \widehat{\rho} = \rho_1 - \rho_2$$

we have, from (2.10),

$$\Sigma(a_{ij}^{\rho_1}\ \widehat{\sigma}_{y_i})_{y_i} + \Sigma a_i^{\rho_1}\ \widehat{\sigma}_{y_i} = \operatorname{div}\vec{f}_1 + f_2 \text{ in } \{y_N>0\},$$

$$\widehat{\sigma} = \Sigma\ b_{ij}^{\rho_1}\ \widehat{\rho}_{y_iy_j} + f_3 \text{ on } \{y_N=0\}, \tag{3.2}$$

$$\widehat{\rho}_t = \Sigma\ \ell_i^{\rho_1}\frac{\partial}{\partial y_i}\ \widehat{\sigma} + f_4 \text{ on } \{y_N=0\},$$

where

$$\|\vec{f}_1\|_{\alpha,0} = \|D_{y'}\hat{\rho}\cdot D\widetilde{\sigma}_2\|_{\alpha,0} \le C\|\widehat{\rho}\|_{1+\alpha,0},$$

$$\|f_2\|_0 = \|(D^2_{y'}\widehat{\rho})\widetilde{\sigma}_2\|_0 \le C\|\widehat{\rho}\|_{2,0},$$

$$\|f_3\|_{1+\alpha,0} \le \|D_{y'}\widehat{\rho}\|_{1+\alpha,0} + |G*\widehat{h}\|_{1+\alpha,0},$$

$$\le C\|\widehat{\rho}\|_{2+\alpha,0} + C\|\widehat{h}\|_0,$$

$$\|f_4\|_{\alpha,0} \le C\|\widehat{\rho}\|_{1+\alpha,0} + C\|\widehat{h}\|_0.$$

All these norms are taken in a small neighborhood of $s=s_0$. Applying Lemma 3.4 of [4] we deduce that

$$\|D_s\widehat{\rho}\|_{2+\alpha,(2+\alpha)/3} \le C\|\widehat{\rho}\|_{2+\alpha,0} + C\|\widehat{h}\|_0$$

locally, and, by partition of unity and the fact that $\widehat{\rho}(x,0)=0$,

$$\|D_s\widehat{\rho}\|_{2+\alpha,(2+\alpha)/3} \le C\|\widehat{h}\|_0 \tag{3.3}$$

globally, provided T is sufficiently small. We can use this inequality to estimate the norms of $\vec{f}_1$ and f_2, f_3, f_4 above and then, applying elliptic estimates to the solution $\widehat{\sigma}$ of (3.2), we obtain the estimates

$$\|\widehat{\sigma}\|_{1+\alpha,0} \le C\|\widehat{h}\|_0, \quad \|\widehat{\sigma}\|_{0,(1+\alpha)/3} \le C\|h\|_0. \tag{3.4}$$

We next proceed to estimate $\widehat{c} \equiv c_1 - c_2$ by flattening the boundaries $\Gamma_{1,t}$, $\Gamma_{2,t}$ using (3.3); we easily obtain

$$\|\widehat{c}\|_{1+\alpha,0} + \|\widehat{c}\|_{0,(1+\alpha)/3} \le C\|\widehat{h}\|_0. \tag{3.5}$$

Next we estimate $\widehat{p} = p_1 - p_2$ and $\widehat{q} = q_1 - q_2$ using (3.4), (3.5), and the arguments used in Lemma 2.2. We get

$$\|\widehat{p}\|_{\alpha,\alpha/3} + \|\widehat{q}\|_{\alpha,\alpha/3} \le C\|\widehat{h}\|_0.$$

Combining this estimate with (3.5) we obtain the inequality

$$\|W(c_1,p_1,q_1) - W(c_2,p_2,q_2)\|_{\alpha,\alpha/3} \le C\|(c_1-c_2, p_1-p_2, q_1-q_2)\|_0$$

and, since $\overline{c}_1 - \overline{c}_2 = 0$, $\overline{p}_1 - \overline{p}_2 = 0$, $\overline{q}_1 - \overline{q}_2 = 0$ at $t=0$,

$$\|W(c_1,p_1,q_1) - W(c_2,p_2,q_2)\|_0 \le CT^{\beta}\|(c_1-c_2, p_1-p_2, q_1, -q_2)\|_0$$

for some $\beta > 0$. It follows that W is a contraction in L^∞ provided T is sufficiently small. This, combined with the fact that W maps Y_M into a compact subset, implies that W is a continuous map in Y_M and hence, by the Schauder fixed point theorem, W has a fixed point. The uniqueness of the fixed point follows from the fact that W is a contraction in the L^∞ norm.

We have thus completed the proof of the following theorem.

THEOREM 3.1. *Under the assumptions* (3.1), (1.13), (1.14), *there exists a unique solution to* (1.1)–(1.4), (1.6)–(1.9) *for* $0 \le t \le T$ *with* ρ, σ *as in* (2.7), *and*

$$c, p, q \text{ in } C^{m+1+\alpha,\ m+1+\alpha/3}(\mathbb{R}^N \times [0,T])$$

for some $T > 0$.

In particular, in terms of the local coordinates (s,t) of $\cup_{0\le t\le T}\Gamma_t \times \{t\}$, the free boundary has $D_sD_{s,t}^m$ derivatives which belong to $C^{3+\alpha,\ (3+\alpha)/3}$.

Remark 3.1. If Γ_0 is only assumed to belong to $C^{3+\alpha}$, we take a C^∞ manifold M which lies within a small $C^{3+\alpha}$-neighborhood of Γ_0 and parametrize Γ_t by

$$\Gamma_t = \{s + \rho(s,t)\vec{n}(s)\}, \quad s \in M,$$

where $\vec{n}(s)$ is the outward normal to M at s. Assuming that p_0, q_0 satisfy (2.21) we can then extend the bootstrap argument and conclude that there exists a unique solution to (1.1)–(1.4), (1.6)–(1.9) for $0 \le t \le T$ with

$$D_s\rho \in C^{3+\alpha,\ (3+\alpha)/3}(\mathbb{R}^N \times [0,T])$$

and

$$D_s D^M_{(s,t)}\rho \in C^{3+\alpha,\ (3+\alpha)3}(\mathbb{R}^N \times [t_0, T])$$

for any $t_0 > 0$; cf. [4] for the corresponding result for the problem (2.6).

Remark 3.2. Theorem 3.1 extends to the case where the elliptic equation (1.1) is replaced by the parabolic equation (1.11), provided the initial values $c(x, 0)$ satisfy the corresponding smoothness and compatibility conditions as indicated in Remark 2.1.

REFERENCES

[1] B. V. Bazaliy, *Stefan problem for the Laplace equation with regard for the curvature of the free boundary*, Ukrainian Math. J., 49 (1997), pp. 1465–1484.

[2] B. V. Bazaliy and A. Friedman, *A free boundary problem for an elliptic-parabolic system: Application to a model of tumor growth*, Comm. Partial Differential Equations, 28 (2003), pp. 517–560.

[3] X. Chen, *The Hele-Shaw problem and area preserving curve-shortening motions*, Arch. Ration. Mech. Anal., 123 (1993), pp. 117–151.

[4] X. Chen, J. Hong, and F. Yi, *Existence, uniqueness and regularity of classical solutions of the Mullins-Sekerka problem*, Comm. Partial Differential Equations, 21 (1996), pp. 1705–1727.

[5] R. Constantin and M. Pugh, *Global solutions for small data to the Hele–Shaw problem*, Nonlinearity, 6 (1993), pp. 393–415.

[6] S. Cui and A. Friedman, *A hyperbolic free boundary problem modeling tumor growth*, Interfaces Free Bound., 5 (2003), pp. 159–181.

[7] J. Duchon and R. Robert, *Evolution d'un interface par capillarité et diffusion de volume* I. *Existence locale en temps*, Ann. Inst. H. Poincaré Non Linéaire, 1 (1984), pp. 361–378.

[8] J. Escher, *Classical Solutions to a Moving Boundary Problem for an Elliptic-Parabolic System*, preprint.

[9] J. Escher and G. Simonett, *Classical solutions of multidimensional Hele–Shaw models*, SIAM J. Math. Anal., 28 (1997), pp. 1028–1047.

[10] J. Escher and G. Simonett, *A center manifold analysis for the Mullins-Sekerka model*, J. Differential Equations, 143 (1998), pp. 267–292.

[11] A. Friedman and F. Reitich, *Nonlinear stability of a quasi-state Stefan problem with surface tension: A continuation approach*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 30 (2001), pp. 341–403.

[12] H. Greenspan, *On the growth and stability of cell cultures and solid tumors*, J. Theor. Biology, 56 (1976), pp. 229–242.

[13] E. I. Hanzawa, *Classical solution of the Stefan problem*, Tohuku Math. J., 33 (1981), pp. 297–335.

[14] G. Pettet, C. P. Please, M. Tindall, and D. McElwain, *The migration of cells in multicell tumor spheroids*, Bull. Math. Biology, 63 (2001), pp. 231–257.

SIAM J. MATH. ANAL.
Vol. 35, No. 4, pp. 987–1004

WELL-POSEDNESS OF THE BOUNDARY LAYER EQUATIONS*

MARIA CARMELA LOMBARDO†, MARCO CANNONE‡, AND MARCO SAMMARTINO†

Abstract. We consider the mild solutions of the Prandtl equations on the half space. Requiring analyticity only with respect to the tangential variable, we prove the short time existence and the uniqueness of the solution in the proper function space. The proof is achieved applying the abstract Cauchy–Kowalewski theorem to the boundary layer equations once the convection-diffusion operator is explicitly inverted. This improves the result of [M. Sammartino and R. E. Caflisch, *Comm. Math. Phys.*, 192 (1998), pp. 433–461], as we do not require analyticity of the data with respect to the normal variable.

Key words. boundary layer, Prandtl equations

AMS subject classifications. 76N20, 76D03, 35A10

DOI. 10.1137/S0036141002412057

1. Introduction.

In this paper we shall be concerned with the unsteady Prandtl equations on the half space. They describe the behavior of an incompressible fluid close to a physical boundary in the limit of small viscosity [19]. The system we shall deal with is the following:

$$(\partial_t - \partial_{YY})\, u^P + u^P \partial_x u^P + v^P \partial_Y u^P + \partial_x p^P = 0 \,, \tag{1.1}$$

$$\partial_Y p^P = 0 \,, \tag{1.2}$$

$$\partial_x u^P + \partial_Y v^P = 0 \,, \tag{1.3}$$

$$u^P(x, Y = 0, t) = v^P(x, Y = 0, t) = 0 \,, \tag{1.4}$$

$$u^P(x, Y \to \infty, t) \longrightarrow U(x, t) \,, \tag{1.5}$$

$$p^P(x, Y \to \infty, t) \longrightarrow p^E(x, y = 0, t) \,, \tag{1.6}$$

$$u^P(x, Y, t = 0) = u^P_{in} \,. \tag{1.7}$$

In the above equations (u^P, v^P) and p^P represent the components of the fluid velocity and the pressure inside the boundary layer. Equation (1.3) is the incompressibility condition and equations (1.4) are the boundary conditions: $u^P(x, Y = 0, t) = 0$ is the no-slip condition and $v^P(x, Y = 0, t) = 0$ is the no-influx condition. Equation (1.5) is the matching condition between the flow inside the boundary layer and the outer Euler flow; $U(x, t)$ is the tangential component of the Euler flow at the boundary; $x = (x_1, x_2)$ is the tangential variable, and Y the normal variable.

The Prandtl equations can be regarded as asymptotic equations of the Navier–Stokes equations in the limit of vanishing viscosity ($\nu \to 0$). In the limit case $\nu = 0$, the higher derivative term is dropped from the Navier–Stokes system and one gets

*Received by the editors July 25, 2002; accepted for publication (in revised form) April 18, 2003; published electronically November 4, 2003. This paper was partially supported by a Galileo grant (Egide 2002). The work of the first and third authors was supported in part also by the MURST under the grant "Problemi Matematici Non Lineari di Propagazione e Stabilità nei Modelli del Continuo."
http://www.siam.org/journals/sima/35-4/41205.html

†Dipartimento di Matematica, Università di Palermo, Via Archirafi 34, 90123 Palermo, Italy (lombardo@math.unipa.it, marco@math.unipa.it).

‡U.F.R. Mathématiques, Université de Marne-la-Vallée, Equipe d'Analyse et de Mathématiques Appliquées, Cité Descartes–5, bd Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2, France (cannone@math.univ-mlv.fr).

the Euler equations, which rule the behavior of inviscid flows. Since the Euler system is first order, we have a reduction of the order of the equations, and a corresponding reduction must be done in the number of the boundary conditions: only the normal component of the velocity can be imposed at the boundary. Since the Navier–Stokes equations impose the value of both the velocity components at the boundary, one must allow a thin layer where there is a rapid variation of the fluid velocity from zero (imposed by the no-slip condition) to the value prescribed by the inviscid equations. Hence, in the boundary layer (whose size is $O(\sqrt{\nu})$), vorticity is generated so that the viscosity term $\nu\Delta\boldsymbol{u}$ is $O(1)$, even as the viscosity goes to zero. The fluid develops an internal length scale so that one is faced with a singular perturbation problem. Rescaling the normal variable with the square root of the viscosity, and writing the solution to the Navier–Stokes equations in the form of an asymptotic series, one gets the equations which rule the fluid inside the boundary layer, i.e., Prandtl equations.

The equations were first derived by Prandtl in 1904, and the practical success of the boundary layer theory was soon overwhelming. Nevertheless, the theoretical foundation of the boundary layer theory was rather unsatisfactory, and many fundamental questions are still debated. For instance, the problem of establishing a well-founded mathematical connection to the Navier–Stokes equation has been solved only recently, and neither existence, uniqueness, nor well-posedness of the boundary layer equation is proved in the general case.

Regarding the problem of the convergence of the Prandtl equations to the Navier–Stokes equations, a major complication is given by the fact that no uniqueness theorem with Sobolev-type initial data for the three-dimensional Navier–Stokes (nor Euler) equations has been proved, and the time of existence of a regular solution depends on the data and on the viscosity (see Marsden [13] and the monographs Constantin and Foias [7] and Temam [21]). In the absence of boundaries the convergence of viscous planar flow to ideal planar flow was shown by Swann [20] for a time which is independent of the viscosity and, lately, in the case of concentrated vorticity, by Constantin and Wu [8].

In the presence of boundaries the problem is harder. Kato [10] proved that a necessary and sufficient condition for the convergence of $\boldsymbol{u}^{NS}$ to the solution of Euler equations, $\boldsymbol{u}^{E}$, in $L^2(\Omega)$ uniformly in $t \in [0,T]$ is that the energy dissipation for $\boldsymbol{u}^{NS}$ in a small layer close to the boundary of size $O(\nu)$, during the interval $[0,T]$, tends to zero. However, such result gives no ultimate solution to the problem because of the unverified energy estimate on the Navier–Stokes solution. With a similar condition on the L^2-norm of the gradient of the pressure, Temam and Wang [22] proved the convergence of the Navier–Stokes solution to the solution of the Euler equation in a strip.

Analogously it is also hard to prove the convergence of the Navier–Stokes solution to the Prandtl solution under satisfactory hypotheses: the few existence and uniqueness theorems proved for the unsteady case hold in particular cases. For instance, Oleinik proved the existence and uniqueness of the Prandtl equations on the half space requiring prescribed horizontal velocities positive and strictly increasing. See [14] for a review.

The first results which do not require monotonicity of the initial data were proved by Sammartino and Caflisch, after the earlier work of Asano [2]. In [17], assuming analyticity of the initial data with respect to the spatial variables, they proved the existence and uniqueness of the Prandtl equations on the half space. They achieved the

result using an abstract formulation of the Cauchy–Kowalewski theorem in the Banach spaces of analytic functions. In [18] they performed the asymptotic analysis of the Navier–Stokes equation in the limit of zero viscosity. They constructed the solution in the form of an asymptotic series in $\sqrt{\nu}$, whose zeroth order term is constituted by the sum of the Euler and the Prandtl solutions. The norm of the first order correction term is then proved to be bounded in the proper function space. They also proved an analogous result in the case of a curved boundary (see [5]).

In the linear case it has been possible to prove the convergence of the linearized Navier–Stokes equations to the corresponding inviscid equations for Sobolev-type initial data. The asymptotic analysis has been successfully performed for the Stokes equations on the half space (Sammartino [16]) and on the exterior of a disk (Lombardo, Caflisch, and Sammartino [11]). Similar results were achieved for the Oseen equations, i.e., the Navier–Stokes equations linearized around a nonzero flow, on a strip (see Lombardo and Sammartino [12] and Temam and Wang [23]).

Temam and Wang analyzed the linear case for a general $2 - D$ exterior domain (see [24] and [25]), but they obtained weaker convergence results. In the nonlinear case, with blowing and suction boundary conditions [26], they were able to prove that these boundary conditions stabilize the boundary layer.

In the opposite direction Grenier [9] proved that a solution of the Prandtl equations is linearly and nonlinearly unstable, and, therefore, it does not converge in H^1 to the Navier–Stokes solutions.

A review about the mathematical aspects of the boundary layer theory can be found in [4].

In this paper we extend the result of [17] to a wider class of initial data, namely, the functions which are analytic only with respect to the tangential variable and L^2, together with their derivatives, with respect to the normal variable. Through the explicit expression of the Green's function, we invert the second order parabolic operator appearing in the Prandtl equation, including the first order Y-derivative. We are thus able to obtain a mild form of the system. The existence and the uniqueness of the solution are then proved using a slightly modified version of the abstract Cauchy–Kowalewski (ACK) theorem in the Banach spaces.

The results presented in this paper were previously announced in [6].

The paper is organized as follows. In section 2 we define the function spaces where existence and uniqueness will be proved. In section 3 we state the abstract Cauchy–Kowalewski theorem in the Banach spaces. In section 4 the parabolic initial-boundary value problem is explicitly solved and the norm of the corresponding operators bounded in the proper function spaces. The mild form of the Prandtl equation is given in section 5. In sections 6 and 7 the source term of the Prandtl equation is proved to satisfy the hypotheses of the ACK theorem. Finally the main theorem is stated in section 8. For convenience two appendices are inserted. In Appendix A a sketch of the proof of the ACK theorem is given. In Appendix B the estimates of the pseudodifferential operator defined in section 4 are proved.

2. Function spaces. In this section we introduce the function spaces used in the proof of the existence and uniqueness of the Prandtl equations. We first define the domain of analyticity with respect to the tangential variable:

$$D(\rho) = \{x \in \mathbb{C} \ : \Im x \in (-\rho, \rho)\}.$$

We now introduce the ambient spaces for the Prandtl equations.

DEFINITION 2.1. *The space $K^{l,\rho}$ is the space of the functions $f(x)$ such that*

- *f is analytic in $D(\rho)$;*
- *if $\Im x \in (-\rho,\rho)$ and $0 \le j \le l$, then $\partial_x^j f(\Re x + i\Im x)$ is square integrable in $\Re x$;*
- *$|f|_{l,\rho} \equiv \sum_{j=0}^{l} \sup_{\Im x\in(-\rho,\rho)} \|\partial_x^j f(\cdot + i\Im x)\|_{L^2(\Re x)} < \infty$.*

DEFINITION 2.2. *The space $K^{l,\rho,\mu}$, with $\mu > 0$, is the space of the functions $f(Y,x)$ such that*

$$e^{\mu Y}\partial_x^i\partial_Y^j f \in L^\infty(\mathbb{R}^+, K^{0,\rho}) \text{ when } i+j \le l \text{ and } j \le 2.$$

The norm in $K^{l,\rho,\mu}$ is defined as

$$|f|_{l,\rho,\mu} \equiv \sum_{j\le 2}\sum_{i\le l-j} \sup_{Y\in\mathbb{R}^+} e^{\mu Y}\ |\partial_Y^j\partial_x^i f(Y,\cdot)|_{0,\rho}.$$

DEFINITION 2.3. *The space $K^{l,\rho}_{\beta,T}$, with $\beta > 0$ and $\rho - \beta T > 0$, is the space of the functions $f(x,t)$ such that*

$$\partial_t^i\partial_x^j f(x,t) \in K^{l,\rho-\beta t} \quad \forall 0 \le t \le T, \text{ where } 0 \le i+j \le l \text{ and } 0 \le i \le 1.$$

Moreover,

$$|f|_{l,\rho,\beta,T} \equiv \sum_{0\le j\le 1}\sum_{i\le l-j} \sup_{0\le t\le T} |\partial_t^j\partial_x^i f(\cdot,t)|_{0,\rho-\beta t} < \infty\ .$$

DEFINITION 2.4. *The space $K^{l,\rho,\mu}_{\beta,T}$, with $\beta > 0$, $\rho - \beta T > 0$ and $\mu - \beta T > 0$, is the space of the functions $f(x,Y,t)$ such that*

$$f \in K^{l,\rho-\beta t,\mu-\beta t} \text{ and } \partial_t\partial_x^i f \in K^{0,\rho-\beta t,\mu-\beta t} \quad \forall 0 \le t \le T, \text{ where } 0 \le i \le l-2.$$

Moreover,

$$\begin{aligned}|f|_{l,\rho,\mu,\beta,T} \equiv &\sum_{0\le j\le 2}\sum_{i\le l-j} \sup_{0\le t\le T} |\partial_Y^j\partial_x^i f(\cdot,\cdot,t)|_{0,\rho-\beta t,\mu-\beta t}\\ &+ \sum_{i\le l-2} \sup_{0\le t\le T} |\partial_t\partial_x^i f(\cdot,\cdot,t)|_{0,\rho-\beta t,\mu-\beta t} < \infty\ .\end{aligned}$$

3. The abstract Cauchy–Kowalewski theorem. To prove the existence and the uniqueness of the mild solution to the Prandtl equations, we shall give a slightly modified version of the abstract Cauchy–Kowalewski (ACK) theorem as given in [15] or [1] and [3].

For t in $[0,T]$, consider the equation

$$u + F(t,u) = 0. \tag{3.1}$$

Let $\{X_\rho :\ 0 < \rho \le \rho_0\}$ be a Banach scale with norms $|\cdot|_\rho$ such that $X_{\rho'} \subset X_{\rho''}$ and $|\cdot|_{\rho''} \le |\cdot|_{\rho'}$ when $\rho'' \le \rho' \le \rho_0$.

THEOREM 3.1 (ACK theorem). *Suppose that $\exists R > 0$, $\rho_0 > 0$, and $\beta_0 > 0$ such that if $0 < t \le \rho_0/\beta_0$, the following properties hold:*

(1) $\forall\, 0 < \rho' < \rho \le \rho_0$ *and* $\forall u$ *such that* $\{u \in X_\rho : \sup_{0\le t\le T} |u(t)|_\rho \le R\}$ *the map* $F(t,u) : [0,T] \mapsto X_{\rho'}$ *is continuous.*

(2) $\forall\, 0 < \rho < \rho_0$ *the function* $F(t,0) : [0, \rho_0/\beta_0] \mapsto \{u \in X_\rho : \sup_{0\le t\le T} |u(t)|_\rho \le R\}$ *is continuous and*

$$|F(t,0)|_\rho \le R_0 < R\,. \tag{3.2}$$

(3) $\forall 0 < \rho' < \rho(s) < \rho_0$ *and* $\forall\ u^1$ *and* $u^2 \in \{u \in X_\rho : \sup_{0\le t\le T} |u(t)|_{\rho-\beta_0 t} \le R\}$,

$$|F(t,u^1) - F(t,u^2)|_{\rho'} \le C \int_0^t ds \left(\frac{|u^1 - u^2|_{\rho(s)}}{\rho(s) - \rho'} + \frac{|u^1 - u^2|_{\rho'}}{\sqrt{t-s}} \right). \tag{3.3}$$

Then $\exists \beta > \beta_0$ *such that* $\forall 0 < \rho < \rho_0$, (3.1) *has a unique solution* $u(t) \in X_\rho$ *with* $t \in [0, (\rho_0 - \rho)/\beta]$*; moreover* $\sup_{\rho<\rho_0-\beta t} |u(t)|_\rho \le R$.

The proof of the above theorem is given in Appendix A.

4. A parabolic equation. The next section will be devoted to writing Prandtl equations in the form given by (3.1). The main difficulty in doing this is in the parabolic nature of the Prandtl equation. We shall solve this difficulty by inverting the parabolic operator $(\partial_t - \partial_{YY} + \alpha Y\, \partial_Y)$, giving the explicit expression of the Green's function.

We introduce the kernels

$$F_\alpha(x,Y,t) = \frac{1}{\sqrt{4\pi}} \frac{1}{\Psi(x,t)} \exp\left(-\frac{Y^2 e^{-2A(x,t)}}{4(\Psi(x,t))^2} \right), \tag{4.1}$$

$$E_\alpha(x,Y,t) = \int_0^\infty dY' \left[F_\alpha(x, Y - Y', t) - F_\alpha(x, Y + Y', t) \right], \tag{4.2}$$

$$H_\alpha(x,Y,t) = -\frac{\partial F_\alpha}{\partial Y}(x,Y,t) + \alpha(x,t) Y F_\alpha(x,Y,t) - \frac{1}{2}\alpha(x,t) E_\alpha(x,Y,t), \tag{4.3}$$

where α is a function of x and t, and $A(x,\tau)$ is defined as

$$A(x,\tau) = \int_0^\tau d\theta\ \alpha(x,\theta) \tag{4.4}$$

and

$$\Psi(x,t) = \left(\int_0^t d\tau\ e^{-2A(x,\tau)} \right)^{1/2}. \tag{4.5}$$

The operator M_0 is the convolution of the kernel F_α with the odd extension to $Y < 0$ of the function $u_0(x,Y)$:

$$M_0 u_0 = \int_0^\infty dY'\ \left[F_\alpha(Y - Y', t) - F_\alpha(Y + Y', t) \right]\ u_0(x, Y'). \tag{4.6}$$

It solves the following system:

$$(\partial_t - \partial_{YY} + \alpha Y\, \partial_Y) M_0 u_0 = 0, \tag{4.7}$$

$$M_0 u_0(x, Y = 0, t) = 0, \tag{4.8}$$

$$M_0 u_0(x, Y, t = 0) = u_0. \tag{4.9}$$

We now introduce the operator M_2:

$$M_2 f = \int_0^t ds \int_0^\infty dY' \; [F_\alpha(Y - Y', t - s) - F_\alpha(Y + Y', t - s)] \; f(x, Y', s). \tag{4.10}$$

It solves the parabolic equations with zero boundary and initial data:

$$(\partial_t - \partial_{YY} + \alpha\, Y\, \partial_Y) M_2 f = f, \tag{4.11}$$

$$M_2 f(x, Y = 0, t) = 0, \tag{4.12}$$

$$M_2 f(x, Y, t = 0) = 0. \tag{4.13}$$

The operator M_1 acts on functions defined on the boundary, namely,

$$M_1 g = 2 \int_0^t ds \; H_\alpha(Y, t - s)\, g(x, s), \tag{4.14}$$

and solves the following system:

$$(\partial_t - \partial_{YY} + \alpha\, Y\, \partial_Y) M_1 g = 0, \tag{4.15}$$

$$M_1 g(x, Y = 0, t) = g, \tag{4.16}$$

$$M_1 g(x, Y, t = 0) = 0. \tag{4.17}$$

Finally we define the operator $M_3 h$:

$$M_3 h = - \int_0^t ds \int_0^\infty dY' \, \partial_Y \; [F_\alpha(x, Y - Y', t - s) - F_\alpha(x, Y + Y', t - s)] \; h(x, Y', s). \tag{4.18}$$

Notice that if $h(x, Y = 0, t) = 0$, then, integrating by parts, one gets $M_3 h \equiv M_2 \partial_Y h$.

We shall now give some estimates on the above operators. We begin with the estimates on the operator M_2.

PROPOSITION 4.1. *Let $\alpha \in K^{l,\rho}_{\beta,T}$, $f \in K^{l,\rho,\mu}_{\beta,T}$ with $f|_{Y=0} = 0$. If $\rho' < \rho - \beta t$ and $\mu' < \mu - \beta t$, then the following estimate holds:*

$$|M_2 f|_{l,\rho',\mu'} \le c \int_0^t ds \; |f(\cdot, \cdot, s)|_{l,\rho',\mu'} \le c\, |f|_{l,\rho,\mu,\beta,T},$$

where the constant c depends on $|\alpha|_{l,\rho,\beta,T}$.

PROPOSITION 4.2. *Let $\alpha \in K^{l,\rho}_{\beta,T}$, $f \in K^{l,\rho,\mu}_{\beta,T}$. Then $M_2 f \in K^{l,\rho,\mu}_{\beta,T}$ and the following estimate holds:*

$$|M_2 f|_{l,\rho,\mu,\beta,T} \le c\, |f|_{l,\rho,\mu,\beta,T} \, .$$

The following estimate of $M_3 h$ will be crucial in handling the nonlinear term containing the Y-derivative.

PROPOSITION 4.3. *Suppose $\alpha \in K^{l,\rho}_{\beta,T}$, $h \in K^{l,\rho,\mu}_{\beta,T}$ with $h|_{Y=0} = 0$, $\partial_Y h|_{Y=0} = 0$. If $0 < \mu' < \mu(s) < \mu - \beta s$, then $M_3 h \in K^{l,\rho,\mu'}$ for each $0 < t < T$ and the following estimate holds:*

$$|M_3 h|_{l,\rho,\mu'} \le c \int_0^t ds \; \left(\frac{|h(\cdot, \cdot, s)|_{l,\rho,\mu'}}{\sqrt{t - s}} + \frac{|h(\cdot, \cdot, s)|_{l,\rho,\mu(s)}}{\mu(s) - \mu'} \right).$$

The proofs of the above propositions are given in Appendix B.

We finally give some bounds on the operators M_0 and M_1.

PROPOSITION 4.4. *Let $\alpha \in K^{l,\rho}_{\beta,T}$ and $u_0(x,Y) \in K^{l,\rho,\mu}$. Moreover let the compatibility condition $u_0(x,Y=0)=0$. Then $M_0 u_0 \in K^{l,\rho,\mu}_{\beta,T}$ and the following estimate holds:*

$$|M_0 u_0|_{l,\rho,\mu,\beta,T} \le c\,|u_0|_{l,\rho,\mu}\ .$$

PROPOSITION 4.5. *Let $\alpha, g \in K^{l,\rho}_{\beta,T}$ and $g(x,t=0)=0$. Then $M_1 g \in K^{l,\rho,\mu}_{\beta,T}$ and the following estimate holds:*

$$|M_1 g|_{l,\rho,\mu,\beta,T} \le c\,|g|_{l,\rho,\beta,T}\ .$$

We will also need the following lemma.

LEMMA 4.6. *Let $\alpha \in K^{l,\rho}_{\beta,T}$, $w = u+g$ with $u \in K^{l,\rho,\mu}$, and $g \in K^{l,\rho}$, i.e., constant with respect to Y and t. Moreover, let $u(x,Y=0) = -g(x)$. Then $M_0(t)w - g \in K^{l,\rho,\mu}\ \forall t$ and the following estimate holds:*

$$\sup_{0\le t\le T} |M_0(t)w - g\ |_{\,l,\rho,\mu} \le c\,(|\alpha|_{l,\rho,\beta,T} + |u|_{l,\rho,\mu} + |g|_{l,\rho})\,.$$

5. The mild form of the Prandtl equations. In this section, following the same procedure used in [17], we shall recast the Prandtl equations in a form suitable for the application of the ACK theorem.

First, one can get rid of the pressure gradient introducing the new variable u:

$$u = u^P - U\ . \tag{5.1}$$

In fact, written in terms of the variable u and using the Euler equation at the boundary,

$$\partial_t U + U\partial_x U + \partial_x p^E|_{y=0} = 0, \tag{5.2}$$

equations (1.1)–(1.7) become

$$(\partial_t - \partial_{YY} + Y\partial_x U\,\partial_Y)\,u + u\,\partial_x u - \left(\int_0^Y dY'\partial_x u\right)\partial_Y u + U\,\partial_x u + u\,\partial_x U = 0, \tag{5.3}$$

$$u(x,Y=0,t) = -U, \tag{5.4}$$

$$u(x,Y\to\infty,t) = 0, \tag{5.5}$$

$$u(t=0) = u^P_{in} - U(t=0) \equiv u_0, \tag{5.6}$$

where we have also used the incompressibility condition, written as

$$v^P = -\int_0^Y \partial_x u^P dY' = -\left(\int_0^Y \partial_x u\ dY' + Y\partial_x U\right). \tag{5.7}$$

We can now define the quantities

$$K_1(u,t) = -\,(2u\,\partial_x u + U\,\partial_x u + u\,\partial_x U)\,, \tag{5.8}$$

$$K_2(u,t) = \partial_Y\left(u\int_0^Y dY'\ \partial_x u,\right) \tag{5.9}$$

and the operator $F(u,t)$ as

$$F(u,t) = M_2 K_1(u,t) + M_2 K_2(u,t) + \mathcal{C}, \tag{5.10}$$

where we have identified the $\alpha(x,t)$ appearing in the kernel F_α with $-\partial_x U(x,t)$, and where $\mathcal{C}$ is defined by

$$\mathcal{C} = M_0(t)\,(u_0 + U(t=0)) - M_1\,(U - U(t=0)) - U(t=0). \tag{5.11}$$

Given that $(u \int_0^Y dY'\ \partial_x u)\,|_{Y=0} = 0$, $F(u,t)$ can be written as

$$F(u,t) = M_2 K_1(u,t) + M_3 K_3(u,t) + \mathcal{C}, \tag{5.12}$$

where $K_3(u,t)$ is defined as

$$K_3(u,t) = u \int_0^Y dY'\ \partial_x u. \tag{5.13}$$

Therefore (5.3), together with the boundary and initial condition (5.4)–(5.6), can finally be written in the form

$$u = F(u,t). \tag{5.14}$$

We call (5.14) with $F(u,t)$ defined in (5.12), and with M_2, M_3, K_1, K_3 defined in (4.10), (4.18), (5.8), (5.13), respectively, the mild form of the Prandtl equations. We are now left to prove that the operator $F(u,t)$, given by (5.12), satisfies the hypotheses of the ACK theorem.

6. The forcing term. It is obvious that the operator $F(u,t)$ satisfies assumption 1 of the ACK theorem. In this section we shall show that it satisfies assumption 2, namely, that $F(0,t) \in K^{l,\rho,\mu}$ and that $\forall t \in [0,t]$

$$|F(0,t)|_{l,\rho,\mu} \le R_0. \tag{6.1}$$

Since

$$F(0,t) = \mathcal{C}, \tag{6.2}$$

using Lemma 4.6 and Proposition 4.5, one gets the following.

PROPOSITION 6.1. *Suppose that $u_0 \in K^{l,\rho,\mu}$ with $u_0(\cdot, Y=0) = -U(t=0)$ and $U \in K^{l,\rho}_{\beta,T}$. Then $F(0,t) \in K^{l,\rho,\mu}_{\beta,T}$ and the following estimate holds:*

$$|F(0,t)|_{l,\rho,\mu,\beta,T} \le c\,(|U|_{l,\rho,\beta,T} + |u_0|_{l,\rho,\mu})\,.$$

This proves that the forcing term can be estimated in terms of the initial condition for Prandtl equations and the outer Euler flow. Notice that the compatibility condition $u_0(\cdot, Y=0) = -U(t=0)$ is necessary for the hypotheses of Lemma 4.6 to be verified.

7. The contractiveness property of the operator F. In this section we shall prove that the operator F, given by (5.10), satisfies assumption 3 of the ACK theorem. Namely, we shall prove the following.

THEOREM 7.1. *Suppose that* u^1 *and* u^2 *are in* $K^{l,\rho_0,\mu_0}_{\beta_0,T}$. *Suppose* $0 < \rho' < \rho(s) < \rho_0 s$ *and* $0 < \mu' < \mu(s) < \mu_0$. *Then the following estimate holds:*

$$\begin{aligned} &\left| F(u^1,t) - F(u^2,t) \right|_{l,\rho',\mu'} \\ &\leq c \int_0^t ds \left(\frac{|u^1-u^2|_{l,\rho(s),\mu}}{\rho(s)-\rho'} + \frac{|u^1-u^2|_{l,\rho,\mu(s)}}{\mu(s)-\mu'} + \frac{|u^1-u^2|_{l,\rho',\mu'}}{\sqrt{t-s}} \right). \end{aligned} \tag{7.1}$$

To prove the above theorem we have to bound the operators M_2K_1 and M_3K_3. The first one contains two different kinds of terms: the nonlinear term, $u\partial_x u$, and two linear terms. They all will be estimated through the Cauchy estimate in the x-variable. The operator M_3K_3, which contains the nonlinear term involving the Y-derivative, will be estimated using the properties of the kernel of the operator M_3.

7.1. The operator M_2K_1. We start with the estimate of the nonlinear term involving the x-derivative. One has the following Cauchy estimate for the derivative of an analytic function.

PROPOSITION 7.2. *Let* $f \in K^{l,\rho''}$. *If* $\rho' < \rho''$, *then*

$$|\partial_x f|_{l,\rho'} \leq \frac{|f|_{l,\rho''}}{\rho''-\rho'} \,. \tag{7.2}$$

Therefore the following proposition can be proved.

PROPOSITION 7.3. *Suppose that* u^1 *and* u^2 *are in* $K^{l,\rho_0,\mu_0}_{\beta_0,T}$. *Suppose* $0 < \rho' < \rho(s) < \rho_0$. *Then the following estimate holds:*

$$\left| u^1\partial_x u^1 - u^2\partial_x u^2 \right|_{l,\rho',\mu'} \leq c \, \frac{|u^1-u^2|_{l,\rho,\mu}}{\rho-\rho'}, \tag{7.3}$$

where the constant c *depends only on* $|u^1|_{l,\rho_0,\mu_0,\beta,T}$ *and* $|u^2|_{l,\rho_0,\mu_0,\beta,T}$.

The proof of the above proposition can be found in [17].

The estimate of the linear terms is easily achieved using the following lemma.

LEMMA 7.4. *Let* $U \in K^{l,\rho}_{\beta,T}$ *and let* $\rho' < \rho$; *then* $\forall\ 0 < t \leq T$

$$\sup_{x \in D(\rho')} |\partial_x^l U(\cdot,t)| \leq c \, |U|_{l,\rho,\beta,T}.$$

The proof of the above lemma is a consequence of the Cauchy estimate for an analytic function and of the Sobolev inequality.

Finally, using Proposition 4.1 and the above lemmas, we get the following.

PROPOSITION 7.5. *Suppose that* u^1 *and* u^2 *are in* $K^{l,\rho,\mu}_{\beta,T}$. *Suppose* $0 < \rho' < \rho(s) < \rho$. *Then the following estimate holds:*

$$\left| M_2K_1(u^1,t) - M_2K_1(u^2,t) \right|_{l,\rho',\mu} \leq c \int_0^t ds \, \frac{|u^1-u^2|_{l,\rho(s),\mu}}{\rho(s)-\rho'}, \tag{7.4}$$

where the constant c *depends only on* $|u^1|_{l,\rho,\mu,\beta,T}$ *and* $|u^2|_{l,\rho,\mu,\beta,T}$.

Notice that the difference $K_1(u^1,t) - K_1(u^2,t)$ has to be considered only for functions which satisfy the condition $u(x,Y=0,t) = -U$, so that $K_1(u^1,t) - K_1(u^2,t)\,|_{Y=0} = 0$. Therefore the requirement of Proposition 4.1 is fulfilled.

7.2. The operator M_3K_3. In this subsection we shall estimate the term containing the Y-derivative using Proposition 4.3. Since it involves also the x-derivative, one must pay attention to the way the derivatives are distributed. In the estimate of the term involving the $\partial_Y^2\partial_x^{l-2}$-derivatives, one has to invoke Proposition 4.3. On the other hand, in the estimate of the term involving the $\partial_Y\partial_x^{l-1}$-derivatives, one has to Cauchy estimate the x-derivative.

The following proposition then holds.

PROPOSITION 7.6. *Suppose that u^1 and u^2 are in $K^{l,\rho,\mu}_{\beta,T}$. Suppose $0<\rho'<\rho(s)<\rho$, $0<\mu'<\mu(s)<\mu$. Then the following estimate holds:*

$$\begin{aligned}(7.5)\quad &\left|M_3K_3(u^1,t)-M_3K_3(u^2,t)\right|_{l,\rho',\mu'}\\ &\qquad\le c\int_0^t ds\left(\frac{|u^1-u^2|_{l,\rho(s),\mu'}}{\rho(s)-\rho'}+\frac{|u^1-u^2|_{l,\rho',\mu(s)}}{\mu(s)-\mu'}+\frac{|u^1-u^2|_{l,\rho',\mu'}}{\sqrt{t-s}}\right),\end{aligned}$$

where the constant c depends only on $|u^1|_{l,\rho,\mu,\beta,T}$ and $|u^2|_{l,\rho,\mu,\beta,T}$.

We stress the fact that we are allowed to use Proposition 4.3, as both the hypotheses are satisfied. In fact the first hypothesis reads $[u^1\int_0^Y dY'\partial_x u^1-u^2\int_0^Y dY'\partial_x u^2]_{Y=0}=0$ and the second one

$$\begin{aligned}&\left[\partial_Y\left(u^1\int_0^Y dY'\partial_x u^1-u^2\int_0^Y dY'\partial_x u^2\right)\right]_{Y=0}\\ &=\left[\partial_Y u^1\int_0^Y dY'\partial_x u^1-\partial_Y u^2\int_0^Y dY'\partial_x u^2\right]_{Y=0}+\left[u^1\partial_x u^1-u^2\partial_x u^2\right]_{Y=0}\\ &=\left[(u^1-u^2)\partial_x u^1+u^2\partial_x(u^1-u^2)\right]_{Y=0}=0,\end{aligned}$$

where the last equality holds since both u^1 and u^2 have the same datum at the boundary.

This concludes the proof of Theorem 7.1.

8. The main result. In the previous sections we have proved that the operator F satisfies all the hypotheses of the ACK theorem. Hence the following theorem, which is the main result of this paper, has been proved.

THEOREM 8.1. *Suppose $U\in K^{l,\rho_0}_{\beta_0,T}$ and $u^P_{in}-U\in K^{l,\rho_0,\mu_0}$. Moreover let the compatibility conditions*

$$u^P_{in}(x,Y=0) = 0, \tag{8.1}$$

$$u^P_{in}(x,Y\to\infty)-U\longrightarrow 0 \tag{8.2}$$

hold. Then there exist $0<\rho_1<\rho_0$, $0<\mu_1<\mu_0$, $\beta_1>\beta_0>0$, and $0<T_1<T$ such that (1.1)–(1.7) admit a unique mild solution u^P. This solution can be written as

$$u^P(x,Y,t)=u(x,Y,t)+U, \tag{8.3}$$

where $u\in K^{l,\rho_1,\mu_1}_{\beta_1,T_1}$.

9. Concluding remarks. In this paper we have proved short time existence and uniqueness of the solution of the Prandtl equations. The main hypothesis we have imposed is the analyticity of the initial data and of the prescribed (Euler) flow with respect to the tangential variable. This improves the results of [17], where analyticity with respect to the normal variable was also imposed.

The main ideas in our proof are the following.

First, we inverted the convection-diffusion (in the normal variable) operator. This led us to introduce the mild form of the Prandtl equations and allowed us to put the Prandtl equations in a form (see (5.14)) suitable for the application of the ACK theorem.

Second, we introduced a modified form of the ACK theorem to deal with a term which has a mild singularity in time (see (3.3)). The origin of this mild singularity is in the fact that, due to the lack of analyticity with respect to the normal variable, we had to use the regularizing properties of the Green's function of the diffusion operator. The gain of regularity in the normal space variable was paid with a mild singularity in time.

Third, the analyticity in the tangential variable was used to deal with the nonlinear convection in the tangential direction. Application of our version of the ACK theorem gave the existence and uniqueness of the solution.

The result of this paper is more general than the results of [17]. Moreover it seems a necessary step toward a rigorous mathematical analysis of the boundary layer theory for curved boundaries. In fact, when the curvature is present, the requirement of analyticity with respect to the normal variable would not allow the asymptotic matching between the exterior and the interior solutions. Therefore the problem of proving the well-posedness of the boundary layer equations when geometries other than very special ones (e.g., the half space or the exterior of a circular domain) are involved does not seem to be out of reach. This would open the possibility of the analysis of the zero viscosity problem for a fluid confined in a general bounded domain.

Appendix A. Proof of the ACK theorem. The proof of Theorem 3.1 follows along the same lines as that of [15].

In fact we prove the ACK theorem by proving that $F(u,t)$ is contractive in an auxiliary Banach space $\mathbb{S}^\gamma$.

For $\gamma > 0$, we consider the weighted Banach space $\mathbb{S}^\gamma$ of continuous functions $u(t)$ with values in X_ρ, where $\rho + \beta t < \rho_0$. The norm in $\mathbb{S}^\gamma$ is defined as

$$\|u\|^{(\gamma)} = \sup_{\rho+\beta t<\rho_0} (\rho_0 - \rho - \beta_0 t)^\gamma \ |u(t)|_\rho. \tag{A.1}$$

The contractiveness of the $F(u)$ in $\mathbb{S}^\gamma$ can be proved as follows.

Let $0 < \rho' < \rho(s) < \rho_0$. We set

$$\rho(s) = \rho' + \frac{\lambda(s)}{2}, \tag{A.2}$$

where

$$\lambda(s) = \rho_0 - \rho' - \beta s. \tag{A.3}$$

Therefore

$$\rho_0 - \rho(s) - \beta s = \frac{\lambda(s)}{2} = \rho(s) - \rho'. \tag{A.4}$$

We can now make the estimate

$$
\begin{aligned}
|F(t,u^1)-F(t,u^2)|_{\rho'} &\le C\int_0^t ds\left(\frac{|u^1-u^2|_{\rho'}}{\sqrt{t-s}}+\frac{|u^1-u^2|_{\rho(s)}}{\rho(s)-\rho'}\right)\\
&\le C\int_0^t ds\left(\frac{|u^1-u^2|_{\rho'}}{\sqrt{t-s}}\frac{(\rho_0-\rho'-\beta s)^\gamma}{(\rho_0-\rho'-\beta t)^\gamma}+\frac{|u^1-u^2|_{\rho(s)}}{\rho(s)-\rho'}\frac{(\rho_0-\rho(s)-\beta s)^\gamma}{(\rho_0-\rho(s)-\beta s)^\gamma}\right)\\
&\le C\|u^1-u^2\|^{(\gamma)}\left[2\sqrt{t}(\rho_0-\rho'-\beta t)^{-\gamma}+\int_0^t ds\frac{2^{\gamma+1}}{(\rho_0-\rho'-\beta s)^{\gamma+1}}\right]\\
&\le C\frac{\|u^1-u^2\|^{(\gamma)}}{(\rho_0-\rho'-\beta t)^\gamma}\left[2\sqrt{\frac{\rho_0}{\beta}}+\frac{2^{\gamma+1}}{\gamma\beta}\right], \qquad \text{(A.5)}
\end{aligned}
$$

where C is the constant appearing in assumption 3. Passing from the second to the third line, we have used (A.3) and (A.4).

Taking the sup of (A.5) over $\rho'+\beta t<\rho_0$, we get

$$
\|F(t,u^1)-F(t,u^2)\|^{(\gamma)} \le 2\left(\sqrt{\frac{\rho_0}{\beta}}+\frac{2^\gamma}{\gamma\beta}\right)\|u^1-u^2\|^{(\gamma)}. \qquad \text{(A.6)}
$$

Therefore, to prove that the operator F is contractive in the (γ)-norm, it is enough to choose β big enough so that $\sqrt{\frac{\rho_0}{\beta}}+\frac{2^\gamma}{\gamma\beta}<\frac{1}{2}$. ∎

Appendix B. Proofs of Propositions 4.1, 4.2, 4.3, 4.4, and 4.5. We first prove some simple lemmas. Set

$$
\Psi(x,t)=\left(\int_0^t d\tau\; e^{-2A(x,\tau)}\right)^{1/2}. \qquad \text{(B.1)}
$$

LEMMA B.1.

$$
\sup_{x\in D(\rho)}\left|\frac{e^{-2A(x,t)}}{(\Psi(x,t))^2}\right| \le \frac{e^{4T\sup_{x,t}|\alpha|}}{t}.
$$

Proof.

$$
\begin{aligned}
\sup_{x\in D(\rho)}\left|\frac{e^{-2A(x,t)}}{(\Psi(x,t))^2}\right| &\le \frac{e^{2T\sup_{x,t}|\alpha|}}{\inf_{x\in D(\rho)}\left|\int_0^t d\tau\; e^{-2A(x,\tau)}\right|} \le \frac{e^{2T\sup_{x,t}|\alpha|}}{\left|\int_0^t d\tau\; e^{-2\sup_{x\in D(\rho)}A(x,\tau)}\right|}\\
&\le \frac{e^{2T\sup_{x,t}|\alpha|}}{\int_0^t d\tau\; e^{-2T\sup_{x\in D(\rho)}|\alpha(x,\tau)|}} \le \frac{e^{4T\sup_{x,t}|\alpha|}}{t}. \qquad \blacksquare
\end{aligned}
$$

Using the above bound it is straightforward to prove the following lemmas.

LEMMA B.2.

$$
\sup_{x\in D(\rho)}\left|\partial_x^l F_\alpha(\cdot,Y,t)\right| \le c\,\frac{\exp\left(-\frac{Y^2\; e^{-4T\sup_{x,t}|\alpha|}}{4t}\right)}{\sqrt{t}}\sum_{i=0}^{l}\left(\frac{Y^2\; e^{4T\sup_{x,t}|\alpha|}}{2t}\right)^i.
$$

LEMMA B.3.

$$
\sup_{x\in D(\rho)}|\partial_Y F_\alpha(\cdot,Y,t)| \le c\,\frac{Y\; e^{4T\sup_{x,t}|\alpha|}}{t}\,\frac{\exp\left(-\frac{Y^2\; e^{-4T\sup_{x,t}|\alpha|}}{4t}\right)}{\sqrt{t}}.
$$

LEMMA B.4.

$$\sup_{x\in D(\rho)}\left|\partial_Y\partial_x^l F_\alpha(\cdot,Y,t)\right| \le c\,\frac{\exp\left(-\frac{Y^2\ e^{-4T\sup_{x,t}|\alpha|}}{4t}\right)}{\sqrt{t}}\sum_{i=0}^{l}\left\{\left(\frac{Y^2\ e^{4T\sup_{x,t}|\alpha|}}{2t}\right)^i\frac{Y\ e^{-4T\sup_{x,t}|\alpha|}}{2t} + \left(\frac{Y^2\ e^{4T\sup_{x,t}|\alpha|}}{2t}\right)^{i-1}\frac{Y\ e^{4T\sup_{x,t}|\alpha|}}{2t}\right\}.$$

In the proof of Proposition 4.5 we shall also need the following two lemmas.

LEMMA B.5.

$$\sup_{x\in D(\rho)}\left|\ \exp\left(-\frac{Y^2e^{-2A(\cdot,Y^2/4\eta^2)}}{4\Psi^2(\cdot,Y^2/4\eta^2)}\right)\ \right| \le\ c\ e^{-\eta^2}.$$

LEMMA B.6.

$$\sup_{x\in D(\rho)}\ \left|\ \Psi^n(\cdot,Y^2/4\eta^2)\ \right| \ge\ c\ \frac{Y^n}{2^n\eta^n}\ e^{-nT\sup|\alpha|}.$$

We now start with the proof of Proposition 4.3.

Proof of Proposition 4.3. In order to estimate $|M_3h|_{l,\rho,\mu'}$ we have to estimate $|\partial_x^iM_3h|_{0,\rho,\mu'}$ with $i\le l$, $|\partial_Y\partial_x^iM_3h|_{0,\rho,\mu'}$ with $i\le l-1$, $|\partial_t\partial_x^iM_3h|_{0,\rho,\mu'}$ with $i\le l-1$, and $|\partial_{YY}\partial_x^iM_3h|_{0,\rho,\mu'}$ with $i\le l-2$.

We begin with $|\partial_x^iM_3h|_{0,\rho,\mu'}$ with $i\le l$.

$$\begin{aligned}
&|\partial_x^iM_3h|_{0,\rho,\mu'}\\
&=\sup_{Y\ge0}e^{\mu'Y}\sup_{|\Im x|\le\rho}\left\|\partial_x^i\int_0^t ds\int_0^\infty dY'\partial_Y\left[F_\alpha(x,Y{-}Y',t-s){-}F_\alpha(x,Y{+}Y',t-s)\right]h(x,Y',s)\right\|_{L^2}\\
&\le\sup_{Y\ge0}e^{\mu'Y}\int_0^t ds\int_0^\infty dY'\sum_{k=0}^{i}\sup_x\left|\partial_x^k\partial_Y\left[F_\alpha(\cdot,Y-Y',t-s)-F_\alpha(\cdot,Y+Y',t-s)\right]\right|\\
&\qquad\times\sup_{|\Im x|\le\rho}\|\partial_x^{i-k}h(\cdot,Y',s)\|_{L^2}\\
&\le c\sup_{Y\ge0}e^{\mu'Y}\int_0^t\frac{ds}{\sqrt{t-s}}\sum_{k=0}^{i}\left\{\int_{\frac{-Y\ e^{-2T\sup|\alpha|}}{2\sqrt{t-s}}}^\infty d\eta\,e^{-\eta^2}\eta^{2k+1}\right.\\
&\qquad\times\sup_{|\Im x|\le\rho}\left\|\partial_x^{i-k}h(x,Y+2\eta e^{2T\sup|\alpha|}\sqrt{t-s},s)\right\|_{L^2}\\
&\quad+\int_{\frac{-Y\ e^{-2T\sup|\alpha|}}{2\sqrt{t-s}}}^\infty d\eta\ e^{-\eta^2}\ k\ \eta^{2k-1}\sup_{|\Im x|\le\rho}\left\|\partial_x^{i-k}h(x,Y+2\eta e^{2T\sup|\alpha|}\sqrt{t-s},s)\right\|_{L^2}\\
&\quad+\int_{\frac{Y\ e^{-2T\sup|\alpha|}}{2\sqrt{t-s}}}^\infty d\eta\ e^{-\eta^2}\eta^{2k+1}\sup_{|\Im x|\le\rho}\left\|\partial_x^{i-k}h(x,-Y+2\eta e^{2T\sup|\alpha|}\sqrt{t-s},s)\right\|_{L^2}\\
&\quad\left.+\int_{\frac{Y\ e^{-2T\sup|\alpha|}}{2\sqrt{t-s}}}^\infty d\eta\ e^{-\eta^2}\ k\ \eta^{2k-1}\sup_{|\Im x|\le\rho}\left\|\partial_x^{i-k}h(x,-Y+2\eta e^{2T\sup|\alpha|}\sqrt{t-s},s)\right\|_{L^2}\right\}\\
&\le c\int_0^t ds\ \frac{1}{\sqrt{t-s}}\ |\partial_x^ih|_{0,\rho,\mu}\le c\int_0^t ds\ \frac{1}{\sqrt{t-s}}\ |h|_{l,\rho,\mu},
\end{aligned}$$

where, in passing from the third to the fourth line, we have used Lemma B.4 and have posed $\eta = \frac{(Y'-Y)e^{-2T\sup|\alpha|}}{2\sqrt{t-s}}$ in the first two integrals and $\eta = \frac{(Y'+Y)e^{-2T\sup|\alpha|}}{2\sqrt{t-s}}$ in the third and fourth integrals.

We now pass to the estimates of $|\partial_Y \partial_x^i M_2 \partial_Y h|_{0,\rho,\mu'}$ with $i \le l-1$.

$$
\begin{aligned}
&|\partial_Y \partial_x^i M_3 h|_{0,\rho,\mu'} \\
&= \sup_{Y\ge 0} e^{\mu' Y} \sup_{|\Im x|\le\rho} \Big\| \partial_x^i \int_0^t ds \int_0^\infty dY'\ \partial_Y \left[F_\alpha(x, Y-Y', t-s) - F_\alpha(x, Y+Y', t-s) \right] \\
&\qquad\qquad \times \partial_{Y'} h(x, Y', s) \Big\|_{L^2} \\
&\le \sup_{Y\ge 0} e^{\mu' Y} \int_0^t ds \int_0^\infty dY' \sum_{k=0}^{i} \sup_x \left| \partial_x^k \partial_Y \left[F_\alpha(\cdot, Y-Y', t-s) - F_\alpha(\cdot, Y+Y', t-s) \right] \right| \\
&\qquad\qquad \times \sup_{|\Im x|\le\rho} \left\| \partial_x^{i-k} \partial_{Y'} h(\cdot, Y', s) \right\|_{L^2} \\
&\le c \sup_{Y\ge 0} e^{\mu' Y} \int_0^t \frac{ds}{\sqrt{t-s}} \sum_{k=0}^{i} \Bigg\{ \int_{\frac{-Y\ e^{-2T\sup|\alpha|}}{2\sqrt{t-s}}}^{\infty} d\eta\ e^{-\eta^2} \eta^{2k+1} \\
&\qquad\qquad \times \sup_{|\Im x|\le\rho} \left\| \partial_x^{i-k} \partial_Y h(x, Y + 2\eta e^{2T\sup|\alpha|}\sqrt{t-s}, s) \right\|_{L^2} \\
&\quad + \int_{\frac{-Y\ e^{-2T\sup|\alpha|}}{2\sqrt{t-s}}}^{\infty} d\eta\ e^{-\eta^2}\ k\ \eta^{2k-1} \sup_{|\Im x|\le\rho} \left\| \partial_x^{i-k} \partial_Y h(x, Y + 2\eta e^{2T\sup|\alpha|}\sqrt{t-s}, s) \right\|_{L^2} \\
&\quad + \int_{\frac{Y\ e^{-2T\sup|\alpha|}}{2\sqrt{t-s}}}^{\infty} d\eta\ e^{-\eta^2} \eta^{2k+1} \sup_{|\Im x|\le\rho} \left\| \partial_x^{i-k} \partial_Y h(x, -Y + 2\eta e^{2T\sup|\alpha|}\sqrt{t-s}, s) \right\|_{L^2} \\
&\quad + \int_{\frac{Y\ e^{-2T\sup|\alpha|}}{2\sqrt{t-s}}}^{\infty} d\eta\ e^{-\eta^2}\ k\ \eta^{2k-1} \sup_{|\Im x|\le\rho} \left\| \partial_x^{i-k} \partial_Y h(x, -Y + 2\eta e^{2T\sup|\alpha|}\sqrt{t-s}, s) \right\|_{L^2} \Bigg\} \\
&\le c \int_0^t ds\ \frac{1}{\sqrt{t-s}}\ |\partial_x^i \partial_Y h|_{0,\rho,\mu} \le c \int_0^t ds \frac{1}{\sqrt{t-s}}\ |h|_{l,\rho,\mu}.
\end{aligned}
$$

The estimate of $|\partial_{YY} M_2 \partial_Y h|_{0,\rho,\mu'}$ is easily achieved by transforming the derivative ∂_{YY} acting on the kernel into $\partial_{Y'}\partial_Y$ and integrating by parts. It then proceeds analogously to the one given above, as the appearance of singular boundary terms is prevented by the condition $\partial_Y h(x, Y=0, t) = 0$.

Finally we have to bound the term $|\partial_t M_3 h|_{0,\rho,\mu'}$. We notice that $\partial_t M_3 h = \partial_{YY} M_3 h - \alpha Y \partial_Y M_3 h$; hence we need to estimate $|Y\partial_Y \partial_x^i M_3 h|_{0,\rho,\mu'}$ with $i \le l-2$ and use the estimate given above.

$$
\begin{aligned}
&|Y\partial_Y \partial_x^i M_3 h|_{0,\rho,\mu'} \\
&= \sup_{Y\ge 0} e^{\mu' Y} Y \sup_{|\Im x|\le\rho} \Big\| \partial_Y \partial_x^i \int_0^t ds \int_0^\infty dY' \left[F_\alpha(x, Y-Y', t-s) - F_\alpha(x, Y+Y', t-s) \right] \\
&\qquad\qquad \times \partial_{Y'} h(x, Y', s) \Big\|_{L^2} \\
&\le \sup_{Y\ge 0} e^{\mu' Y} \sup_{|\Im x|\le\rho} \Big\| \partial_x^i \int_0^t ds \int_0^\infty dY'\ Y \left[F_\alpha(x, Y-Y', t-s) + F_\alpha(x, Y+Y', t-s) \right] \\
&\qquad\qquad \times \partial_{Y'}^2 h(x, Y', s) \Big\|_{L^2}
\end{aligned}
$$

$$
\begin{aligned}
&\le \sup_{Y\ge 0} e^{\mu' Y} \sup_{|\Im x|\le\rho} \left\| \partial_x^i \int_0^t ds \int_0^\infty dY' \ (Y-Y') \ F_\alpha(x, Y-Y', t-s) \ \partial_{Y'}^2 h(x, Y', s) \right\|_{L^2} \\
&+ \sup_{Y\ge 0} e^{\mu' Y} \sup_{|\Im x|\le\rho} \left\| \partial_x^i \int_0^t ds \int_0^\infty dY' \ (Y+Y') \ F_\alpha(x, Y+Y', t-s) \ \partial_{Y'}^2 h(x, Y', s) \right\|_{L^2} \\
&+ \sup_{Y\ge 0} e^{\mu' Y} \sup_{|\Im x|\le\rho} \left\| \partial_x^i \int_0^t ds \int_0^\infty dY' \ Y' \left[F_\alpha(x, Y-Y', t-s) - F_\alpha(x, Y+Y', t-s) \right] \right. \\
&\qquad\qquad \left. \times \partial_{Y'}^2 h(x, Y', s) \right\|_{L^2} \\
&\le c \sup_{Y\ge 0} e^{\mu' Y} \int_0^t \frac{ds}{\sqrt{t-s}} \left\{ \sum_{k=0}^{i} \int_{\frac{-Y \ e^{-2T\sup|\alpha|}}{2\sqrt{t-s}}}^{\infty} d\eta \ e^{-\eta^2} \eta^{2k+1} \right. \\
&\qquad\qquad \times \sup_{|\Im x|\le\rho} \left\| \partial_x^{i-k} \partial_Y^2 h(x, Y + 2\eta e^{2T\sup|\alpha|}\sqrt{t-s}, s) \right\|_{L^2} \\
&\quad \left. + \sum_{k=0}^{i} \int_{\frac{Y \ e^{-2T\sup|\alpha|}}{2\sqrt{t-s}}}^{\infty} d\eta \ e^{-\eta^2} \eta^{2k+1} \sup_{|\Im x|\le\rho} \left\| \partial_x^{i-k} \partial_Y^2 h(x, -Y + 2\eta e^{2T\sup|\alpha|}\sqrt{t-s}, s) \right\|_{L^2} \right\} \\
&+ c \sup_{Y\ge 0} \sup_{|\Im x|\le\rho} \left\| \partial_x^i \int_0^t ds \int_0^\infty dY' \frac{e^{\mu'(Y-Y')}}{\mu-\mu'} \left[F_\alpha(x, Y-Y', t-s) - F_\alpha(x, Y+Y', t-s) \right] \right. \\
&\qquad\qquad \left. \times \sup_{Y'\ge 0} e^{\mu Y'} \ \partial_{Y'}^2 h(x, Y', s) \right\|_{L^2} \\
&\le c \int_0^t ds \left(\frac{|\partial_x^i \partial_Y^2 h|_{0,\rho,\mu}}{\sqrt{t-s}} + \frac{|\partial_x^i \partial_Y^2 h|_{0,\rho,\mu}}{\mu-\mu'} \right),
\end{aligned}
$$

where, in passing from the second to the third line, we have integrated by parts and used the condition $\partial_Y h(x, Y = 0, t) = 0$. In the last step, the third integral was estimated using Lemma B.2 and the boundedness of the integral with respect to Y'. □

Proofs of Propositions 4.1, 4.2, and 4.4. The proofs of Propositions 4.1, 4.2, and 4.4 are easily achieved by adopting the same techniques used to prove Proposition 4.3.

Proof of Proposition 4.5. To prove Proposition 4.5 it is useful to introduce the following change of variable into the expression (4.14) for the operator $M_1 g$:

$$
\eta = \frac{Y}{2\Psi(x, t-s)}, \tag{B.2}
$$

where the function $\Psi(x, t-s)$ has been defined by (B.1). Since $\Psi(x, t-s)$ is a monotone function of the time variable, one can express $t-s$ as a function of η. Namely, it exits the function Φ such that

$$
s = t - \Phi(Y/2\eta).
$$

Therefore the expression (4.14) becomes

$$
\begin{aligned}
(\mathrm{B.3}) \quad M_1 g = 4 \int_{\frac{Y}{2\Psi(x,t)}}^{\infty} d\eta \, \exp\left(-\eta^2 e^{-2A(x,\,\Phi(Y/2\eta))}\right) g(x,\, t-\Phi(Y/2\eta)) \\
\times \left[1 + \frac{Y^2}{2\eta^2}\, \alpha(x,\, \Phi(Y/2\eta))\, e^{2A(x,\,\Phi(Y/2\eta))}\right] \\
- \int_0^t dz\, g(x,t-z)\, \alpha(x,z) \left[\int_{-\frac{Ye^{-2A}}{2\Psi(x,z)}}^{\infty} d\eta\, e^{-\eta^2} - \int_{\frac{Ye^{-2A}}{2\Psi(x,z)}}^{\infty} d\eta\, e^{-\eta^2}\right],
\end{aligned}
$$

where, in the last integral, we have also posed $t - s = z$.

To estimate $|M_1 g|_{l,\rho,\mu}$ we have to estimate $|\partial_x^i M_1 g|_{0,\rho,\mu}$ with $i \le l$, $|\partial_t \partial_x^i M_1 g|_{0,\rho,\mu}$ with $i \le l-1$, $|\partial_Y \partial_x^i M_1 g|_{0,\rho,\mu}$ with $i \le l-1$, and $|\partial_{YY} \partial_x^i M_1 g|_{0,\rho,\mu}$ with $i \le l-2$.

The estimate of the term $|\partial_x^i M_1 g|_{0,\rho,\mu}$ with $i \le l$ is easily achieved by letting the operator ∂_x^i act and by using the same techniques of Proposition 4.3.

Analogously, one can get the estimate of the term $|\partial_t \partial_x^i M_1 g|_{0,\rho,\mu}$ with $i \le l-1$, noticing that, in the expression (B.3), the time derivative commutes with the integral because $g(x, t=0) = 0$.

We now estimate the term $|\partial_Y \partial_x^i M_1 g|_{0,\rho,\mu}$ with $i \le l-1$. Recalling that if $f = f(\Phi(Y/2\eta))$, one has

$$
\partial_Y f = \frac{\partial f}{\partial \Phi} \frac{\partial \Phi}{\partial (Y/2\eta)} \frac{1}{2\eta} = -\frac{Y e^{2A(x,\,\Phi(Y/2\eta)\,)}}{2\eta^2} \frac{\partial f}{\partial \Phi},
$$

we obtain the expression for $\partial_Y M_1 g$:

$$
\begin{aligned}
\partial_Y M_1 g = 8Y \int_{\frac{Y}{2\Psi(x,t)}}^{\infty} d\eta \, \exp\left(-\eta^2 e^{-2A(x,\,\Phi(Y/2\eta))}\right) g(x,\, t-\Phi(Y/2\eta)) \\
\times \left[1 + \frac{Y^2}{2\eta^2}\, \alpha(x,\, \Phi(Y/2\eta))\, e^{2A(x,\,\Phi(Y/2\eta))}\right] \\
+ 2 \int_{\frac{Y}{2\Psi(x,t)}}^{\infty} d\eta \, \exp\left(-\eta^2 e^{-2A(x,\,\Phi(Y/2\eta))}\right) \frac{Y e^{2A(x,\,\Phi(Y/2\eta))}}{\eta^2}\, \partial_t g(x,\, t-\Phi(Y/2\eta)) \\
\times \left[1 + \frac{Y^2}{2\eta^2}\, \alpha(x,\, \Phi(Y/2\eta))\, e^{2A(x,\,\Phi(Y/2\eta))}\right] \\
+ 4 \int_{\frac{Y}{2\Psi(x,t)}}^{\infty} d\eta \, \exp\left(-\eta^2 e^{-2A(x,\,\Phi(Y/2\eta))}\right) \frac{Y e^{2A(x,\,\Phi(Y/2\eta))}}{\eta^2}\, g(x,\, t-\Phi(Y/2\eta)) \\
\times \left[\alpha - \frac{Y^2}{\eta}\, e^{2A(x,\,\Phi(Y/2\eta))}\, \left(\alpha - \frac{\partial_t \alpha}{2}\right)\right] \\
- \int_0^t g(x,t-z)\, \alpha(x,\, z)\, \frac{\exp\left(-\frac{Y^2 e^{-2A(x,z)}}{4\Psi^2(x,z)}\right)}{\Psi(x,z)}.
\end{aligned}
$$

Using the above expression and Lemmas B.5 and B.6, the estimate of the terms $|\partial_Y \partial_x^i M_1 g|_{0,\rho,\mu}$ with $i \le l-1$ and $|Y \partial_Y \partial_x^i M_1 g|_{0,\rho,\mu}$ with $i \le l-1$ is straightforward. The proof of Proposition 4.5 is thus achieved.

Acknowledgments. The second author acknowledges K. Asano, C. Bardos, and T. Yanagisawa for fruitful suggestions and enlightening discussions on the topic during his stay at Kyoto University in November 2001.

REFERENCES

[1] K. Asano, *A note on the abstract Cauchy-Kowalewski theorem,* Proc. Japan Acad. Ser. A, 64 (1988), pp. 102–105.

[2] K. Asano, *Zero-viscosity limit of the incompressible Navier-Stokes equations.* II, in Mathematical Analysis of Fluid and Plasma Dynamics, Sūrikaisekikenkyūsho Kōkyūroku 656, Kyoto University, Research Institute for Mathematical Sciences, Kyoto, 1988, pp. 105–128.

[3] R.E. Caflisch, *A simplified version of the abstract Cauchy-Kowalewski theorem with weak singularities*, Bull. Amer. Math. Soc., 23 (1990), pp. 495–500.

[4] R.E. Caflisch and M. Sammartino, *Existence and singularities for the Prandtl boundary layer equations*, Z. Angew. Math. Mech., 80 (2000), pp. 733–744.

[5] R.E. Caflisch and M. Sammartino, *Navier-Stokes equations on an exterior circular domain: Construction of the solution and the zero viscosity limit*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 861–866.

[6] M. Cannone, M.C. Lombardo, and M. Sammartino, *Existence and uniqueness for the Prandtl equations*, C. R. Acad. Sci. Paris Sér. I Math., 332 (2001), pp. 277–282.

[7] P. Constantin and C. Foias, *Navier-Stokes Equations,* University of Chicago Press, Chicago, 1988.

[8] P. Constantin and J. Wu, *Inviscid Limit for Vortex Patches*, Nonlinearity, 8 (1988), pp. 735–742.

[9] E. Grenier, *On the nonlinear instability of Euler and Prandtl equations*, Comm. Pure Appl. Math., 53 (2000), pp. 1067–1091.

[10] T. Kato, *Remarks on the zero viscosity limit for nonstationary Navier-Stokes flows with boundary*, in Seminar on Partial Differential Equations, Mathematical Sciences Research Institute, Berkeley, CA, 1984, pp. 85–98.

[11] M.C. Lombardo, R.E. Caflisch, and M. Sammartino, *Asymptotic analysis of the linearized Navier-Stokes equation on an exterior circular domain: Explicit solution and the zero viscosity limit*, Comm. Partial Differential Equations, 26 (2001), pp. 335–354.

[12] M.C. Lombardo and M. Sammartino, *Zero viscosity limit of the Oseen equations in a channel*, SIAM J. Math. Anal., 33 (2001), pp. 390–410.

[13] J.E. Marsden, *Nonlinear Semigroups Associated with the Equations for a Non-homogeneous Fluid*, University of California, Berkeley, 1970.

[14] O.A. Oleinik and V.N. Samokhin, *Mathematical Models in Boundary Layer Theory*, Chapman & Hall/CRC, Boca Raton, FL, 1999.

[15] M.V. Safonov, *The abstract Cauchy-Kowalewski theorem in a weighted Banach space*, Comm. Pure Appl. Math., 48 (1995), pp. 629–637.

[16] M. Sammartino, *The boundary layer analysis for Stokes equations on a half-space,* Comm. Partial Differential Equations, 22 (1997), pp. 749–771.

[17] M. Sammartino and R.E. Caflisch, *Zero viscosity limit for analytic solutions of the Navier-Stokes equation on a half-space* I. *Existence for Euler and Prandtl equations,* Comm. Math. Phys., 192 (1998), pp. 433–461.

[18] M. Sammartino and R.E. Caflisch, *Zero viscosity limit for analytic solutions of the Navier-Stokes equation on a half-space* II. *Construction of the Navier-Stokes solution*, Comm. Math. Phys., 192 (1998), pp. 463–491.

[19] H. Schlichting, *Boundary Layer Theory*, 4th ed., McGraw-Hill Series in Mechanical Engineering, Karlsruhe, Germany, 1960.

[20] H.S.G. Swann, *The convergence with vanishing viscosity of nonstationary Navier-Stokes flow to ideal flow in* $\mathbb{R}^3$, Trans. Amer. Math. Soc., 157 (1971), pp. 373–397.

[21] R. Temam, *Navier-Stokes Equations: Theory and Numerical Analysis*, North-Holland, Amsterdam, 1977.

[22] R. Temam and X. Wang, *The convergence of the solutions of the Navier-Stokes equations to that of the Euler equations*, Appl. Math. Lett., 10 (1997), pp. 29–33.

[23] R. Temam and X. Wang, *On the behavior of the solutions of the Navier-Stokes equations at vanishing viscosity,* Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 25 (1997), pp. 807–828.

[24] R. Temam and X. Wang, *Asymptotic analysis of the linearized Navier-Stokes equations in a general* $2D$ *domain*, Asymptot. Anal., 14 (1997), pp. 293–321.

[25] R. Temam and X. Wang, *Boundary layers for Oseen's type equation in space dimension three*, Russian J. Math. Phys., 5 (1998), pp. 227–246.

[26] R. Temam and X. Wang, *Remarks on the Prandtl equation for a permeable wall. Special issue on the occasion of the* 125*th anniversary of the birth of Ludwig Prandtl,* ZAMM Z. Angew. Math. Mech., 80 (2000), pp. 835–843.

SIAM J. MATH. ANAL.
Vol. 35, No. 4, pp. 1005–1028

DARCY'S LAW AND THE THEORY OF SHRINKING SOLUTIONS OF FAST DIFFUSION EQUATIONS*

JUAN L. VÁZQUEZ†

Abstract. The behavior of the solutions of the degenerate parabolic equation

$$v_t = v\, v_{xx} + \kappa |v_x|^2, \qquad \kappa \in \mathbb{R},$$

a basic model in the theory of flows in porous media, depends strongly on the parameter κ. We show here a striking example of that variability in the case of compactly supported solutions having free boundaries. We consider the initial-value problem with continuous and compactly supported initial data $v(x,0) = v_0(x) \geq 0$. When $\kappa > 0$ it is well known that this problem admits a unique weak solution which is compactly supported and the following free-boundary conditions are satisfied: $v = 0$, $v_t = \kappa\, |v_x|^2$. The latter relation is a form of Darcy's law and determines a unique solution, hence a unique choice of the interface.

We prove here that for $\kappa \leq 0$ there exist infinitely many solutions $v \geq 0$ with the same initial data and an interface on which Darcy's law holds. Actually, the interfaces can be chosen as arbitrary Lipschitz continuous curves as long as the support shrinks. Therefore, Darcy's law does not play a selecting role for this free-boundary problem.

Key words. Darcy's law, fast diffusion, pressure equation, free boundaries, uniqueness, shrinking supports

AMS subject classifications. 35K55, 35K65

DOI. 10.1137/S0036141001396540

1. Introduction.

This paper is devoted to showing that the behavior of the solutions of the degenerate parabolic equation

$$v_t = v\, v_{xx} + \kappa |v_x|^2, \tag{1.1}$$

called the *pressure equation* because of its well-known application in modeling the flow of compressible fluids through porous media [36, 38, 11], depends strongly on the parameter $\kappa \in \mathbb{R}$. We show here a striking instance of that variability in the case of compactly supported solutions having free boundaries for (1.1) with continuous and compactly supported initial data

$$v(x,0) = v_0(x) \geq 0, \quad x \in \mathbb{R}. \tag{1.2}$$

Case $\kappa > 0$. This is the case that appears in gas theory. It also appears in filtration theory ($\kappa = 1$), in lubrication ($\kappa = 1/3$), and many other applications; cf. [3, 8, 37, 44]. It is well known that for all positive κ the initial-value problem posed in $Q = \mathbb{R} \times (0, \infty)$ admits a unique weak solution which turns out to be continuous in (x,t) and compactly supported in x for every t. Moreover, the regions $\{v > 0\}$ and $\{v = 0\}$ are separated by locally Lipschitz interfaces (also known as free boundaries) where the free-boundary conditions are satisfied:

$$v = 0, \qquad v_t = \kappa\, |v_x|^2 . \tag{1.3}$$

*Received by the editors September 27, 2001; accepted for publication (in revised form) March 14, 2003; published electronically November 4, 2003. This work was supported by the European network HYKE, was funded by the EC under contract HPRN-CT-2002-00282, and was performed during the author's stay at the TICAM Institute of the University of Texas, Austin, in 2001, with partial support from M.E.C., Spain.

http://www.siam.org/journals/sima/35-4/39654.html

†Departamento de Matemáticas, Universidad Autónoma de Madrid, 28046 Madrid, Spain (juanluis.vazquez@uam.es).

The second relation is a form of writing Darcy's law on the interface (see below). If we assume for simplicity that $v_0(x) > 0$ on an interval $I = (a, b)$ and zero outside, then the following facts are well known:

(i) There are two continuous interfaces $x = L(t)$ and $x = R(t)$, where $L(t)$ is nonincreasing with $L(0) = 0$, and $R(t)$ nonincreasing with $R(0) = b$.

(ii) $v > 0$ precisely in the set

$$\Omega = \{(x, t) : L(t) < x < R(t)\}, \tag{1.4}$$

which is noncontracting in time (in fact, strictly expanding after a finite waiting time).

(iii) v is C^∞ smooth in the positivity set Ω and on to the lateral boundary after the waiting time.

(iv) The solution is unique in several other formulations. Thus, it is unique as a viscosity solution in the sense of [21]. It is also uniquely determined as the solution of the following classical free-boundary problem.

Problem (P_0). Given v_0 as above, to find a subset $\Omega \subset Q$ of the form (1.4) with continuous L and R, $L(0) = a$, $R(0) = b$, and a smooth function v defined and positive in Ω, continuous at $t = 0$, and Lipschitz continuous up to the lateral boundary, and such that (1.1) is satisfied in Ω, the initial data are taken, v is zero on the lateral boundary (the interfaces), and finally Darcy's law is true as a limit for all $t > 0$.

Moreover, the unique solution has a noncontracting support. Let us recall some facts about Darcy's law. It is a basic law of fluids through porous media which states that seepage velocity $\mathbf{q}$ is proportional to the gradient of pressure, which is identified with v (up to a constant). In (1.1), (1.6) it is well known that the law means that $\mathbf{q} = -\kappa \nabla v$. For a problem with free boundaries there can be difficulty in identifying these quantities at the free boundary because of lack of regularity. For a solution which is C^1 up to a C^1 interface, say $R(t)$, we have

$$R'(t) = -\kappa\, v_x(R(t), t)\,. \tag{1.5}$$

Since $v(R(t), t) = 0$ so that $v_t + R'(t)\, v_x = 0$ on the interface $x = R(t)$, this implies $v_t = \kappa\, |v_x|^2$. Indeed, both forms are equivalent if $v_x \neq 0$ on the interface. Actually, in the PME theory it is proved that the last form happens on the moving parts of the interface [19], and it happens trivially at resting interfaces. The only analytical difficulty happens at the corner point that may be formed at the waiting time [5], where the stronger form of the law must be reformulated with side derivatives, $D_t^+ R(t_*) = -\kappa\, D_x^- v(R(t_*), t_*)$; see section 4. The same happens for the left interface: we have $L'(t) = -\kappa\, v_x(L(t), t)$ and $v_t = \kappa\, |v_x|^2$ on $x = L(t)$.

Equation (1.1) is equivalent to the well-known *porous medium equation* (PME)

$$u_t = (u^{m-1} u_x)_x \tag{1.6}$$

through the transformation $v = u^{m-1}$, and then the parameters are related by

$$m = \frac{\kappa + 1}{\kappa}, \qquad \kappa = \frac{1}{m - 1}, \tag{1.7}$$

a correspondence that excludes only $m = 1$ and $\gamma = 0$. In the typical gas application u is the density, so we will refer to (1.6) as the density equation for all exponents m. For $m < 1$ it is usually called the *fast diffusion equation*.

Case $\kappa < 0$. A completely different situation appears for $\kappa < 0$. It is known, through the work of Bertsch et al. [13, 14, 15, 16], that in the range $\kappa < 0$ there

exist infinitely many solutions $v \geq 0$ with the same initial data, since the bounding interfaces can be quite arbitrarily chosen as long as the support does not expand. The solution with stationary support is identified as the maximal solution and also as the limit of the vanishing viscosity approximation, a result that is extended to completely arbitrary measurable initial data by Chasseigne and Vázquez in [23].

The problem of identification of the classes of solutions with contracting support is addressed by Angenent [1], who constructs (local in time) unique solutions which have a high regularity near the free boundary that allows for a Taylor expansion. Recent work of Barenblatt et al. [9] points out the use of Darcy's law to numerically identify the solutions with shrinking supports. But these authors also point out that further investigation of the issue is needed.

We will show here that both conditions in (1.3) hold on the interfaces of infinitely many constructed solutions for any initial data. More precisely, we pose the following problem, formally similar to Problem (P_0).

Problem (P_1). Let $\kappa < 0$. Given a continuous function v_0 which is positive in a bounded interval $I = (a, b)$ and zero outside, find a subset of Q of the form

$$\Omega = \{(x,t) : L(t) < x < R(t), 0 < t < T\} \tag{1.8}$$

with continuous L and R, $L(0) = a$, $R(0) = b$, and a smooth function v defined and *positive in* Ω, continuous up to the lateral boundaries, and such that the initial data are taken continuously, v is zero on the lateral boundaries, and (1.1) is satisfied in Ω.

We prove that for every $\kappa < 0$ this problem can be solved in infinitely many ways by prescribing Ω as a nonexpanding domain.

THEOREM 1.1. *Let us prescribe not only v_0 as before but also the monotone curves $x = R(t)$ and $x = L(t)$ in a (finite or infinite) interval $t \in [0, T]$ with $L(t) < R(t)$ and $L(0) = a$, $R(0) = b$. We also assume that L and R are Lipschitz continuous and monotone so that Ω is nonexpanding at all times. Then there exists a unique solution to Problem (P_1) in Ω. Moreover, the solution is classical in Ω, and the maximum principle applies to the solutions of this mixed problem; i.e., the map $(v_0, L, R) \mapsto v$ preserves order in the obvious way. Finally, there are no solutions of this problem if the domain Ω fails to be nonexpanding.*

Existence of solutions being more or less known from the work of previous authors, uniqueness is one of the main points of this paper. Let us make some useful comments: the monotonicity of R and L is not necessarily strict; it must be directed so that Ω is nonexpanding; i.e., R must be monotone nonincreasing and L monotone nondecreasing. The conditions on v_0, L, and R can be weakened, but only to a certain degree since there is a limitation: certain choices of continuous boundaries produce discontinuous solutions, a question that is discussed in [48].

The last assertion of the theorem means that positive solutions do not exist when Ω fails to be nonexpanding. It will be explained in section 3 as follows: when we apply the process of construction of a maximal solution outlined in that section, then the limit v vanishes inside Ω near the interface at the points where Ω expands, so that an actual free boundary appears inside Ω which is monotone as prescribed by the theorem.

The behavior near the boundary and the relation with Darcy's law constitute another main contribution.

THEOREM 1.2. *If L and R are Lipschitz continuous, then the solution behaves in a linear way near the free boundary and Darcy's law is satisfied on the interfaces for almost all $0 < t < T$. If $R(t)$ (resp., $L(t)$) is C^1 in some time interval J, then*

v is C^1 up to the boundary in J and conditions (1.3) *hold in the classical sense. If $L'(t) \neq 0$ (resp., $R'(t) \neq 0$), then the solution is linear nondegenerate near the free boundary.*

Darcy's law is stated at (1.5) for a C^1 right-hand side interface. It must be written with the help of side-derivatives in the general case. This is carefully stated in Corollary 4.3. Similarly for the left-hand side interface.

In other words, the result implies that *imposing the zero value plus Darcy's law on the interfaces does not characterize the solution* for any initial data, as it did for Problem (P_0) when $\kappa > 0$. Identifying physically significant classes of solutions with shrinking supports is still an open problem. We recall that prescribing a suitable pair of conditions on the free boundary involving the solution u, its derivatives, u_x and u_t, and the speed of the interface is the usual recipe for obtaining a well-posed problem in free boundary problems related to parabolic equations, such as the Stefan problem and its many variants, the porous medium problem, or the Hele–Shaw problem; cf. [28]. It fails in this case when using the seemingly natural choice of conditions.

Notation and outline. For convenience we prefer to write $\gamma = -\kappa$ when studying the equation with negative κ. Then the equation reads

$$v_t = v\, v_{xx} - \gamma |v_x|^2, \tag{1.9}$$

the relation between γ and the porous medium exponent is

$$m = \frac{\gamma - 1}{\gamma}, \qquad \gamma = \frac{1}{1-m}, \tag{1.10}$$

and density is related to pressure by $u = v^{-\gamma}$. Our scheme is to obtain maximal and minimal solutions by approximation and standard comparison (section 3) and then obtain Darcy's law at all Lipschitz points of the interfaces (section 4). For the full proof of uniqueness for data with our generality we need the higher machinery of nonlocal transformations using the so-called mass variable. In this way we pass from the original equation into a couple of associated equations (p-Laplacian equations) for some associated solutions; cf. section 5. Uniqueness is proved in terms of the second p-Laplacian formulation in section 6. The relation with the theory of fast diffusion equations is further investigated in section 7, where an improvement of previous existence and uniqueness results is obtained, Theorem 7.2.

The case $\gamma = 0$. The equation under consideration is seemingly "simpler," $v_t = v\, v_{xx}$, but it offers a quite interesting variation of the above line of argument. The general result of Theorem 1.1 is essentially recovered, but there are a number of significant modifications in Theorem 1.2. The main one is that Darcy's law can be applied but it takes a quite different from. It cannot be derived from the equation for v as in the case $\gamma > 0$, since we lack the term $(v_x)^2$. There is, however, a change of variables $\pi = \pi(v)$, given by formula (8.3), that transforms the equation into

$$\pi_t = v\, \pi_{xx} - (\pi_x)^2. \tag{1.11}$$

Then Darcy's law means $R'(t) = \pi_x$ on the interface, and $\pi \sim v/\log(1/v)$ replaces v in the regularity statement of the main result above. This means that on smooth strictly contracting interfaces we will have $|v_x| = \infty$, instead of the finite value of the range $\gamma > 0$. Incidentally, we find both quadratic and linear pressure profiles for stationary interfaces. See the complete technical details (simple but rather beautiful in the author's opinion) in section 8.

Further results. We devote an additional section to the question of limited regularity: we show existence of self-similar solutions representing a piecewise linear interface with a corner point in section 9. We end with a section of conclusions, comments, and open problems.

2. Traveling wave solutions. The simplest explicit solutions are of course the constants, $v = C$, $C \geq 0$. The simplest family of nonclassical (weak) solutions with moving free boundaries consists of the *traveling waves* (TWs) with speed $c \neq 0$, which exist for $\gamma \neq 0$ and have the form

$$V_c(x,t) = \left\{ \frac{c}{\gamma}(x - ct) \right\}_+ . \tag{2.1}$$

Then V_c behaves linearly near the interface $\{x_1 = ct\}$ and Darcy's law is obviously satisfied. By changing x into $-x$ we find TWs moving to the left with speed c with interface $x = -ct$. It can also be obtained by changing the sign of c: $V_{-c} = (-(c/\gamma)(x+ct))_+$. Supports expand for $\gamma < 0$ ($\kappa > 0$) and contract for $\gamma > 0$ ($\kappa < 0$). For $\gamma = 0$ we have stationary solutions, $V(x,t) = (ax)_+$, $a \neq 0$.

TWs are the model of behavior of all the solutions near a moving interface with nonzero speed and will be used below in proving that such behavior holds, as explained in Theorem 1.2. At first sight, there is something to be worried about in the family of traveling waves, namely, the relative behavior of waves with different speeds.

LEMMA 2.1. *For $\gamma > 0$ the maximum principle does not hold for any initially ordered pair of TWs with different speeds.*

Indeed, if they are ordered at $t = 0$ they move in the same direction. The reason the maximum principle breaks down later is that the slope equals the receding speed, and hence an initially larger solution moves back faster. Note that the maximum principle holds for TWs as well as for the rest of the solutions if $\gamma < 0$.

The linear family does not exhaust the set of TWs. If we try the form $v(x,t) = V(x+ct)$, we find the equation $cV' = VV'' - \gamma\,(V')^2$. In order to integrate it we pass to the density function $U = V^{-\gamma}$ and obtain $(U^{m-1}U')' = cU'$, hence $U^{m-1}U' = K_0 + cU$, and finally

$$V' = -\frac{c}{\gamma} - K\,V^{\gamma}, \quad K \in \mathbb{R}. \tag{2.2}$$

The case $K = 0$ is the linear TW seen before: $V = -c\,s/\gamma$ for $s < 0$, and zero for $s \geq 0$. For $K > 0$ integrate this equation from $V(0) = 0$ to obtain a solution $V = V_{c,K}(s)$ for $s < 0$. We get a solution which goes to infinity as s decreases. This happens at a finite distance if $\gamma > 1$, at an infinite distance if $0 < \gamma \leq 1$. The rates are easily calculated from the asymptotic behavior $V' \sim -KV^{\gamma}$. In any case, $V_{c,K} \to \infty$ as $K \to \infty$ uniformly for $s \leq \varepsilon < 0$.

On the contrary, when $K = -H < 0$, if we start with $V(0) = 0$, we get a bounded TW which increases as $s \to -\infty$ up to the level

$$V(-\infty) = h(c,K) \equiv \left(\frac{c}{\gamma\,H} \right)^{1/\gamma},$$

so that $V_{c,-H} \to 0$ uniformly as $H \to \infty$. But we can also obtain increasing profiles if we start with $V(0) = h_0 > h(c,K)$. Then $V > h$, $V' > 0$ for all $s \in \mathbb{R}$ and $V(-\infty) = h$. These profiles have a vertical asymptote at the right-hand end of the

domain of definition if $\gamma > 1$. Keeping h_0 and c fixed and letting $H \to \infty$ we get a limit zero if $s < 0$, infinity if $s > 0$.

Remark. Other types of explicit examples of solutions are known in the literature; cf. [9, 23, 34] and their references. For $\gamma = 0$ see more in section 8.

3. Maximal and minimal solution. First uniqueness. We know that for $\kappa = -\gamma > 0$ the expansive character of the solutions of the Cauchy problem implies that there is a unique choice of the interfaces compatible with the zero-flux condition. We now start proving that this fails for $\gamma > 0$. We review in this section the existence question; we point out the existence of both a *maximal* and a *minimal* solution for Problem (P_1) with a shrinking prescribed domain; we start the analysis of the boundary behavior and prove a partial uniqueness result.

Let us then fix $\gamma \geq 0$. We first recall some well-known facts: since the equation is only degenerate parabolic at the value $v = 0$, a theory of classical solutions with bounded initial data such that $v_0(x) \geq \varepsilon > 0$ is standard [35], the problem is well posed, the solutions are C^∞ smooth, the maximum principle applies, and these facts hold independently of the value of γ.

Such a nice panorama breaks down if the value $v = 0$ is admitted. Let us review the approximation procedure for the Cauchy problem in the whole space ($x \in \mathbb{R}$) as done by Bertsch and collaborators [16]. Given a bounded initial function $v_0 \geq 0$ defined in the whole line, the initial data are approximated by adding $\varepsilon > 0$ into $v_{0\varepsilon}(x) = v_0(x) + \varepsilon$, $x \in \mathbb{R}$. This method is equivalent to the standard vanishing viscosity approximation (cf. [16]) and produces in the limit $\varepsilon \to 0$ a function $v \in \mathcal{C}(\mathbb{R} \times [0, T])$ which is the maximal solution with respect to all possible weak solutions of the Cauchy problem (a different use of the word maximal, as the reader will notice). The following property is important in what follows: the maximal solutions thus constructed have *constant supports* for $t > 0$. Let us also remark that maximal solutions with the same initial data depend monotonically on κ because of the form of the equation and the simple comparison theorem for the approximating problems.

We now address the question of solving the mixed problem (P_1).

PROPOSITION 3.1. *Let $\gamma \geq 0$ and let v_0, L, R be as stated in Theorem* 1.1. *There exists a solution to Problem* (P_1) *which is positive everywhere in Ω if the support is nonexpanding at all times. Indeed, there exist both a maximal and a minimal solution. The problem cannot be solved if the boundaries have any expanding parts.*

Proof. (i) In order to obtain a solution we extend $v_0(x)$ to the whole line by putting its value to zero outside of I_0. The results of [16] imply that the maximal solution of the Cauchy problem also solves our problem in a domain Ω with constant boundaries $R(t) = b$, $L(t) = a$. We will use this solution, $V(x, t)$, which is positive in $(a, b) \times (0, \infty)$, as an upper bound for the solutions with all other kinds of boundaries.

(ii) We now exclude expanding boundaries: let u be a possible solution with a boundary that reaches a point $R(t_1) > b$ at some $t_1 > 0$. Let us argue by comparison with the approximations V_ε to the maximal solution of the Cauchy problem, which is a smooth and positive solution defined in $Q = \mathbb{R} \times (0, \infty)$. From the maximum principle applied in $\Omega \cap \{t \leq t_1\}$ we have $v(x, t) \leq V_\varepsilon$ in Ω; hence in the limit $v \leq V$. But $V(x, t_1)$ vanishes for $x \geq b$, hence a contradiction. The same argument excludes the left boundary less than a. Summing up, $L(t) \geq a$ and $R(t) \leq b$ for all solutions and all times $t > 0$.

In order to avoid nonmonotone parts of the boundaries at later times, displace the origin of time at any $t_0 \in (0, T)$. It follows that $R(t) \leq R(t_0)$ and $L(t) \geq L(t_0)$ for $t > t_0$. Summing up, the lateral boundaries must be nonexpanding at all times if

we want them to be the lateral free boundaries of a solution which is positive inside domain Ω.

(iii) *Maximal solution.* We now construct the maximal solution to the mixed problem with given nonexpanding lateral boundaries. We approximate the problem in the standard sense by adding $\varepsilon > 0$ to the initial and boundary data and solving in Ω. By classical theory (cf. [35]) we get a family of solutions $v_\varepsilon \geq \varepsilon$ which are ordered.

In order to obtain an a priori bound from below for any solution $\overline{v}$ we compare with the maximal solution of the problem posed in a rectangle $\mathcal{R} = I_1 \times (0, \tau)$, where I_1 is a small subinterval of I_0 and τ is chosen as large as possible as long as $\mathcal{R}$ is strictly included in Ω. In this way we get an a priori uniform bound from below $w(x,t) > 0$ inside Ω independent of the particular solution.

Since this bound applies to all the approximations v_e of a given problem, standard comparison shows that $w \leq v_\varepsilon$; we can take the monotone limit as $\varepsilon \to 0$ to get a function v, which is still equal to or larger than w, and hence positive in Ω. Since the equation is nondegenerate when $v > 0$, we conclude that v is smooth in Ω. Continuity up to the boundary is obtained by the method of barriers.

Accepting this result, which we prove next, the construction of the maximal solution is finished. Indeed, any other solution v_2 having the stated properties can be compared with v_ε so that $v_2 \leq v_\varepsilon$, and in the limit $v_2 \leq v$. This means that v is maximal for (P_1).

LEMMA 3.2 (Lipschitz continuity at the boundary). *Suppose that $R(t)$ is a nonincreasing and Lipschitz continuous function in an interval $[t_1, t_2]$, $0 < t_1 \leq t_2$. Then v grows at most linearly near the interface $x = R(t)$: there is a constant $C_1 > 0$ such that for h small enough*

$$0 < v < C_1 \, (R(t) - x) \tag{3.1}$$

whenever $R(t) - h < x < R(t)$ and $t_1 \leq t \leq t_2$. Similarly for the left interface $L(t)$. The constant C_1 depends only on the Lipschitz constant of R, t_1, t_2, and the initial data.

Proof. Fix any $t_0 \in [t_1, t_2]$ and let $x_0 = R(t_0)$. As an upper bound we consider a linear TW $V_c(x,t) = (c/\gamma)(d - x - ct)_+$ and choose c and d so that its straight interface, $x = d - ct$, touches the curve $x = R(t)$ for the first time at $t = t_0$, while $d - ct > R(t)$ for $t < t_0$. We also want $v_0(x) < V_c(x, 0)$. Both requirements are easily satisfied if we first choose c large enough and then let $d = ct_0 + x_0$.

Suppose that we happen to know that v is continuous up to the boundary. Then we compare V_c and v in the domain $\Omega_0 = \{(x,t) : L(t) < x < R(t), 0 < t < t_0\}$. We discover that, locally inside Ω_0, both V_c and v are classical solutions of a nondegenerate parabolic equation and that they are continuous up to the parabolic boundary and are strictly ordered there. Hence, v cannot touch V_c in any interior point, and we conclude that $v(x, t_0) \leq V_c(x, t_0) = (c/\gamma)(x_0 - x)$ for $x < R(t)$. This gives the upper bound, and $C_1 = c/\gamma$ depends only on the Lipschitz constant $K(R)$ of $R(t)$ and the initial data. The dependence on v_0 can be reduced to $R(0)$ and $\|v_0\|_\infty$.

If we do not know if v is continuous, the argument requires a bit more patience. We compare the approximation v_ε used in the construction of the solution with a displaced TW $V^*(x,t) = V_c(x - \delta, t) = (c/\gamma)(d + \delta - x - ct)$. It is easy to see from the same maximum principle argument used above that for a suitable δ that depends on ε, and the same c and d, we have $v \leq V^*$ in Ω_0 . Passing to the limit $\varepsilon \to 0$ and taking into account that $\delta \to 0$ as $\varepsilon \to 0$, we conclude that $v \leq V_c$ as before. In particular, we have established the desired continuity of v near $(R(t_0), t_0)$. ∎

(iv) *Minimal solution.* In order to obtain the minimal solution we construct the maximal solution in domains Ω_δ strictly included in Ω and converging monotonically to Ω as $\delta \to 0$. We have for these solutions the ordering $u_\delta \le \overline{u}$, and $u_\delta \le u_{\delta'}$ if $\Omega_\delta \subset \Omega_{\delta'}$. Passing to the monotone limit $\delta \to 0$ we obtain a solution which is easily shown to be continuous and the minimal solution. The previous argument is rather sketchy, but it uses ideas that can be found in the literature; cf. [22]. □

First uniqueness results. We will prove in section 6 that the maximal and minimal solutions coincide, which means uniqueness for the overdetermined version of Problem (P_1) stated in Theorem 1.1. The general proof needs more sophisticated tools, which we introduce in section 5. However, the result can be obtained very easily for symmetric data, and we present that proof here. First, we need a comparison result. We say that two solutions u_1 and u_2 are strictly ordered if their domains of definition are strictly included in the sense that $R_1(t) > R_2(t)$ and $L_1(t) < L_2(t)$ for all $t \in [0,T]$, and $v_1 > v_2$ in $\overline{\Omega_2}$. In the same way we define strictly ordered data.

LEMMA 3.3. *If the data of two solutions are strictly ordered, so are the solutions.*

The proof is just an application of the classical maximum principle at interior points of Ω_2. Using this result we can get a first uniqueness result.

PROPOSITION 3.4. *If v_0 has only one maximum point, then the solution is unique.*

Proof. After shifting the x-axis if necessary, we may assume that the maximum of v_0 is taken at $x = 0$. Given a solution v_1 we can use the scaling law

$$v_{1,\lambda}(x,t) = \frac{1}{\lambda^2} v(\lambda x, t)$$

with $\lambda = 1 - \varepsilon \in (0,1)$ to produce another solution of (1.9) with strictly larger data. If v_2 is the second solution of the original problem, then by the previous lemma $v_{\lambda,1} \ge v_2$ in Ω. In the limit $\lambda \to 1$ we get $v_1 \ge v_2$. Reversing the roles, we conclude that $v_1 = v_2$. □

4. Lipschitz continuity near the interfaces and Darcy's law. Here we study in greater detail the linear behavior of solutions near moving interfaces. We make full use of the families of TW solutions studied in section 2 as comparison functions.

PROPOSITION 4.1. *Suppose that $R(t)$ is a Lipschitz continuous function with derivative bounded above and below away from zero in absolute value in an interval $[t_1, t_2]$, $0 < t_1 < t_2$. Then v behaves linearly in a nondegenerate way near the interface $x = R(t)$: there are constants $C_1, C_2 > 0$ such that for h small enough*

$$C_2\,(R(t) - x) < v < C_1\,(R(t) - x) \tag{4.1}$$

whenever $R(t) - h < x < R(t)$ and $t_1 \le t \le t_2$. Besides, v_x and v_t are uniformly bounded near the right interface. Similarly for the left interface $L(t)$.

Proof. The upper bound has been established in Lemma 3.2. We proceed in a similar manner for the lower bound. Now a convenient TW is placed locally below as follows: we select a small speed $c \sim 0$ and set the value of d as before, $d - ct_0 = x_0$. By the assumption on R, if c is small enough, we can obtain separate interfaces for v and the TW $V(x,t) = V_c(x - d + ct)$ in the interval $0 < t < t_0$. We also have $V(x,0) < v_0(v)$ in a small interval $d - h \le x \le d < b = R(0)$. We then compare v and V in the region

$$S_h = \{(x,t) : d - h - ct < x < d - ct\,,\ 0 < t < t_0\},$$

which is a subset of Ω if h is small enough. In order to apply the maximum principle we need to check that $v(x,t) > V(x,t) = V_c(-h) > 0$ on the line $H(t) = d - h - ct$, $0 < t < t_0$. This will be true by continuity if c and h are small enough. Note that the line $x = H(t)$ is contained in the positivity set of v and approaches the zero set of v only as $t \to t_0$, where it stays at a distance not less than h independent of c. We conclude for c, h small $v(x,t) \geq V(x,t)$ in $S - h$ so that

$$v(x,t_0) \geq V(x,t_0) = V_c(x - x_0)$$

for $x_0 - h \leq x \leq x_0$. This establishes the lower bound, and the constant depends on the data as in Lemma 3.2 and is uniform in the whole interval $[t_1, t_2]$.

(ii) The estimate on v_x and v_t comes now from standard rescaling arguments with

$$v_\lambda(x,t) = \frac{1}{\lambda} v(\lambda\,(x - x_0), \lambda\,(t - t_0)) \tag{4.2}$$

and letting $\lambda \to 0$, as performed by Caffarelli and collaborators for the porous medium equation; cf. [19, 4]. □

In what follows we need to make clear the notation for side derivatives. For a function $x = R(t)$ which is continuous at a point t_0 we denote by $D_t^+(t_0)$ and $D_t^-(t_0)$ the limits

$$D_t^+ R(t_0) = \lim_{t \searrow t_0} \frac{R(t) - R(t_0)}{t - t_0}, \qquad D_t^- R(t_0) = \lim_{t \nearrow t_0} \frac{R(t) - R(t_0)}{t - t_0}$$

whenever they exist. They are called lateral derivatives. The notation $D_x^+ v(x_0,t_0)$, $D_x^- v(x_0,t_0)$, $D_t^- v(x_0,t_0)$, $D_x^+ v(x_0,t_0)$ for lateral partial derivatives should now be clear.

PROPOSITION 4.2. *Under the above conditions on v and $R(t)$, at every time t_0 where $D_t^- R(t_0)$ exists there also exist at $P_0 = (R(t_0), t_0)$ the limits $D_x^- v(P_0)$ and $D_t^- v(P_0)$, and*

$$\begin{aligned} D_x^- v\,(P_0) &= \lim_{x \nearrow R(t_0)} v_x(x,t_0) = \frac{1}{\gamma} D_t^- R(t_0)\,, \\ -D_t^- v\,(P_0) &= -\lim_{t \nearrow t_0} v_t(x_0,t) = \frac{1}{\gamma} (D_t^- R(t_0))^2\,. \end{aligned} \tag{4.3}$$

The convergence of v_x and v_t is uniform on nontangential directions approaching P_0 within the support with $t < t_0$ (backward directions). Similarly for the left interface $L(t)$.

Proof. Let $x_0 = R(t_0)$. We are assuming that $c_0 = -D_t^- R(t_0)$ is positive and finite.

(i) In order to prove the inequalities from above in (4.3) for the left derivatives $D_x^- v$, $D_t^- v$ we repeat at t_0 the argument of Proposition 4.1 after replacing the comparison function $V_c(s)$ by a nonlinear TW $V_{c,K}$ of type (2.2), where $c = c_0 + \varepsilon$ and $\varepsilon > 0$ is (arbitrarily) small. The assumption on R at t_0 says that by selecting a small $\tau > 0$ we can have

$$R(t_0) + (c_0 - \varepsilon)(t_0 - t) < R(t) < R(t_0) + (c_0 + \varepsilon)(t_0 - t)$$

for all $t \in [t_0 - \tau, t_0)$. If we now shift the origin of times to $t_1 = t_0 - \tau$ and select d as before, the function $V(x,t) = V_{c,K}(x + ct - d)$ will have an interface that touches

$x = R(t)$ at $t = t_0$ and is larger than $R(t)$ in the time interval $t_0 - \tau \le t < t_0$. Next, we take a large parameter $K > 0$ to ensure that $V(x,t) > v(x,t)$ at time $t = t_1 = t_0 - \tau$ and for $x \le R(t)$. This is where the parameter K is needed. Comparison as in the previous proposition then gives $V(x,t) > v(x,t)$ for $t_0 - \tau < t < t_0$ and $x \in (L(t), R(t))$. In the limit $t = t_0$, recalling that the parameter K does not affect the boundary behavior of the TW in the first approximation, we get

$$\frac{v(x,t_0)}{x_0 - x} \le \frac{1}{\gamma}(c + 2\varepsilon), \quad \frac{v(x_0,t)}{t_0 - t} \le \frac{1}{\gamma}((c + 2\varepsilon)^2)$$

for every $x \in (x_0 - h, x_0)$, and h depends only on the data and ε.

(ii) For the lower inequality we use the nonlinear TWs $V_{c,K}$ of (2.2) with $c = c_0 - \varepsilon$ and $H = -K \gg 1$ so that $V_{c,K} \sim 0$ and compare in a strip of the form S_h for small h and for $t_0 - \tau \le t < t_0$ to get $V(x,t_0) \le v(x,t_0)$ for $x \in (x_0 - h, x_0)$; hence

$$\frac{v(x,t_0)}{x_0 - x} \ge \frac{1}{\gamma}(c - 2\varepsilon).$$

The comparison on the lateral boundary $x = R(t) - h$ is correct if we take H large enough thanks to the minimal linear growth proved in Proposition 4.1. The error in the lower bound is then uniform in a strip of width h that depends on the constants of the data.

(iii) We still need to prove that the slope value at the interface is the limit of the values of v_x at the interior along nontangential directions, and the same for v_t. A proof using the scaling arguments is as follows. By the previous result we get the linear behavior

$$v(x,t) = \frac{c}{\gamma}\,\rho\,(1 + \varepsilon(t,\rho)), \quad \rho = R(t) - x, \tag{4.4}$$

where $\varepsilon(t,\rho) \to 0$ as $\rho \to 0$, $t \to t_0$ in a way that depends only on the stated constants. Translating the origin of coordinates to a point (x_0, t_0), scaling as in Proposition 4.1, and passing to the limit $\lambda \to 0$ we get

$$v_\lambda(x,t) \to V_{c_0}(x,t), \tag{4.5}$$

uniformly on compact subsets of $\{(x,t) : t \le 0\}$. By standard parabolic theory we get $v_{\lambda,x} \to -(c/\gamma)$, $v_{\lambda,t} \to -(c^2/\gamma)$ on a parabolic neighborhood of the point $(-1, 0)$. Changing back the scaling we conclude that v_x is continuous up to $x = R(t)$, at $t = t_0$ with uniform convergence along sets of the form

$$K_{n,h}(x_0,t_0) = \{(x,t) :\ t \le t_0,\ x_0 - h \le x \le x_0 + n(t_0 - t)\}$$

whenever $n < c_0$. Analogously for v_t. □

COROLLARY 4.3. (i) *At every point where $R(t)$ is left-differentiable and $R'(t_0) < 0$ we have Darcy's law in the form*

$$R'(t) = \gamma\, D_x^- v(R(t),t), \quad D_t^- v(R(t),t) = -\gamma\,(D_x^- v(R(t),t))^2\,.$$

(ii) *If $R(t)$ is C^1 with $R' < 0$ in an interval $t_1 < t < t_2$, then v is locally uniformly linear near that part of the interface, v_x and v_t are continuous up to $x \le R(t)$, and*

$$v_x(x,t) = \frac{1}{\gamma}R'(t), \quad v_t = -\gamma\,(v_x)^2 \quad \textit{for } x = R(t) \quad \textit{and } t \in (t_1,t_2). \tag{4.6}$$

Analogously for the left interface.

Proof. In the latter case the scaling argument holds locally uniformly in t_0. □

Extension. In order to cover the results of Theorem 1.2 we need to consider also the case $D_t^- R(t_0) = 0$ and prove that $D_x^- v(R(t_0), t_0) = D_t^- v(R(t_0), t_0) = 0$. This result follows easily: we do not need a lower estimate, and the upper one is proved by approximation and comparison, as an easy modification of Proposition 3.2 shows.

The reader should note that points at which $R'(t) = 0$ are more difficult to analyze in detail, because such generality includes stationary points and points where the boundary function $R(t)$ decreases with a rate that is less than linear. In those cases the lower part of the linear estimates may be false, and the free-boundary problem becomes *degenerate* in the usual free-boundary terminology. For instance, the solutions with stationary interfaces have a quadratic growth near the stationary boundaries. This class of solutions has been studied in detail in [23].

5. The mass variable and the conjugate *p*-Laplacian equations. We still have to prove the uniqueness of the solutions of Problem (P_1) for general data. This will be done after we perform some radical transformations of the problem. As a motivation we consider the mass function of the initial distribution $u_0(x)$ which is defined as

$$M_u(x) = \int_{x_0}^{x} u_0(x)dx$$

and is continuous and increasing. The origin x_0 is chosen at will. In probability, and taking as origin $x = -\infty$, this is called the distribution function for the density u_0 (which is then normalized to have integral 1).

In order to construct a function of both x and t with nice properties we argue as follows: let $u(x,t) > 0$ be a classical solution of (1.6) in a cylindrical space-time domain $\mathcal{R}$ of the form (1.8) with lateral boundaries given continuous functions $x = L(t)$, $x = R(t)$. We introduce the *space-time mass function*

$$X(x,t) = \int_\Gamma u dx + u^{m-1}u_x dt, \tag{5.1}$$

where the differential form $\omega = udx + u^{m-1}u_x dt$ is integrated along any curve Γ lying in the domain of u and joining the points (x_0, t_0) and (x,t) for some fixed $(x_0, t_0) \in \mathcal{R}$. Since (1.6) holds, the integrand defining X is an exact differential, and hence X is a potential that can be calculated equivalently along any path, for instance,

$$X(x,t) = \int_{x_0}^{x} u(x,t)dx + \int_{t_0}^{t} (u^{m-1}u_x)|_{x=0} dt, \tag{5.2}$$

as long as this broken path lies in $\mathcal{R}$ (which happens at least locally). The mass function has interesting properties. Let us consider some of them.

• *p-Laplacian equation.* Since $\partial X/\partial x = u$ and $\partial X/\partial t = u^{m-1}u_x$, the partial derivatives of $X = X(x,t)$ are related by equation $X_t = (X_x)^{m-1}X_{xx}$. Taking into account that $X_x = u > 0$ in $\mathcal{R}$, the equation can be written in the more standard form

$$X_t = \frac{1}{m}(|X|_x^{m-1}X_x)_x, \tag{5.3}$$

which is the so-called *p-Laplacian equation* with exponent $p = m + 1$. (Remark: if $m = 0$, the form is $X_t = X_{xx}/X_x = (\log X_x)_x$.) We will use these facts as an essential

tool in the uniqueness proof. But before going into that, we have to understand the inverse transformation.

• *Inversion.* Consider the transformation $(x,t) \mapsto (X,t)$ associated to the positive smooth solution u of (1.6) defined in the domain $\mathcal{R}$.

LEMMA 5.1. *The transformation $T : (x,t) \mapsto (X,t)$ is smooth in $\mathcal{R}$, one-to-one onto its image, and C^1 invertible.*

Proof. The Jacobian matrix of the transformation is

$$J = \frac{\partial(X,t)}{\partial(x,t)} = \begin{pmatrix} u & X_t \\ 0 & 1 \end{pmatrix}, \tag{5.4}$$

where $X_t = u^{m-1}u_x$. Since $\det(J) = u \neq 0$, the transformation is locally invertible. Moreover, time is conserved and $\partial X/\partial x = u > 0$; this monotonicity implies that we can define $x = w(X,t)$, and that the function w is also smooth in the domain $\mathcal{R}'$, the image of $\mathcal{R}$ by the direct transformation. Summing up, the transformation is a global diffeomorphism onto the image, which is also a cylindrical open domain. □

In principle, $\mathcal{R}'$ need not have continuous lateral boundary functions, like $\mathcal{R}$. This would depend on appropriate behavior of u near $L(t)$ and $R(t)$.

• *Conjugate equation.* Let us look closer at the inverse transformation. Its Jacobian is given by the matrix

$$J^{-1} = \frac{\partial(x,t)}{\partial(X,t)} = \begin{pmatrix} u^{-1} & w_t \\ 0 & 1 \end{pmatrix}, \tag{5.5}$$

where $w_t = -u^{m-2}u_x$ is the derivative of $x = w(X,t)$ with respect to t for fixed X. We can then join this result with $w_X = 1/u$ in the formula

$$w_t = (w_X)^{-1-m} w_{XX} = -\frac{1}{m}((w_X)^{-m})_X, \tag{5.6}$$

which is another p-Laplacian equation, with exponent $p' = 1 - m$. We call (5.3) and (5.6) *conjugate equations*. They have exponents $p' + p = 2$. (Note that when negative values of the gradient are allowed, the equation takes on the more general form $w_X = -(1/m)(|w_X|^{-1-m}w_X)_X$, but we do not need this generality here.)

• *Differentiated conjugate equations. Bäcklund transform.* We have seen that $u = \partial X/\partial x$ satisfies a porous medium-type equation (1.6) with exponent $m = (\gamma - 1)/\gamma$. In a similar way, differentiation with respect to X of the equation for $w = w(X,t)$ gives the PME equation

$$U_t = (U^{-1-m}U_X)_X = -\frac{1}{m}(U^{-m})_{XX} \tag{5.7}$$

for $U = w_X(X,t) > 0$ (U^{-m} is replaced by $\log(U)$ if $m = 0$). Note that by the rule of differentiation of the inverse function, we have the relations

$$U(X,t) = \frac{1}{u(x,t)} = v(x,t)^{\gamma}. \tag{5.8}$$

The two PME equations with exponents m and $m' = -m$ are usually obtained from each other in the literature by means of the so-called Bäcklund transform, which is essentially equivalent to the approach developed here but is less convenient in our opinion for the present application. The process of passing from (x,u) to (X,U) can be done in the reverse way, and both processes are inverses of each other.

Historical remark. The Bäcklund transform is well known in the case $m = -1$, where it linearizes the nonlinear equation into the heat equation $u_t = u_{xx}$; cf. [18]. This does not happen in the rest of the cases. It applies (1.6) onto itself when $m = 0$. For uses of the transform in related problems, see [41], where other references are given. We study in detail these transformations for the general filtration equation in [49].

• *Initial behavior.* Let us examine the extension of the transformation to the initial line $t = 0$. This is a trace argument that can be done under quite general circumstances. We give here for the reader's convenience a simple direct argument that is not usually found and is suited to our situation.

We assume that v takes the initial data v_0 in a continuous way and that v_0 is positive in the initial interval I_0. By restricting this interval if needed, we may also assume that v_0 is bounded below, $0 < c < v_0 < M$ in I_0' is a strict subset of $[a, b]$. It easily follows from comparison arguments using the theory explained in section 3 that in a small rectangle of the form $\mathcal{R}_1 = I_0' \times [0, \tau]$ v is bounded below and above in the form

$$0 < c_1 < v < M_1 \quad \text{in } \mathcal{R}_1.$$

Then the equation is uniformly parabolic in $\mathcal{R}_1$. Now we use the formula for X along a broken path

$$X(x,t) = X(x_0,t_0) + \int_{x_0}^{y} u(s,t_0)ds + \int_{t_0}^{t} (u^{m-1}u_x)|_{x=y}dt + \int_{y}^{x} u(s,t)\,ds\,,$$

where $0 < t_0 \le t < \tau$ and y is any intermediate point between x_0 and x, points of I_0'. We assume without loss of generality (w.l.o.g.) that $x_0 < x$. Let us now take a function $\phi(y)$ which is smooth, nonnegative, and supported in (x_0, x). Then

$$\begin{aligned}(X(x,t) - X(x_0,t_0)) \int \phi(y)\,dy &= \left(\int_{x_0}^{y} u(s,t_0)ds + \int_{y}^{x} u(s,t)\,ds\right)\phi(y)\,dy \\ &\quad + \int dy \int_{t_0}^{t} dt\,(u^{m-1}u_x)(y,t)\,.\end{aligned}$$

We now put the normalization $\int \phi(y)\,dy = 1$ to get

$$|X(x,t) - X(x_0,t_0)| \le A + B\,,$$

where $A \le \int_{x_0}^{x} M_1 ds = M_1(x - x_0)$, while

$$B \le \left|\iint \phi(y)\,(u^{m-1}u_x)(y,t)\,dtdy\right| = \left|\frac{1}{m}\iint \phi_x(y)u^m(y,t)\,dtdy\right|\,.$$

By taking $|x - x_0| \le \delta$ and $|t - t_0| \le \tau$ small and observing then that $|\phi_y| = O(1/\delta)$ we get

$$|X(x,t) - X(x_0,t_0)| \le C_1\delta + C_2\tau/\delta\,.$$

Now take $\tau = O(\delta^{1+\varepsilon})$ and let $t_0 \to 0$ to conclude that X can be extended continuously to $(x_0, 0)$. Moreover, X is uniformly Lipschitz continuous in x. It is easily seen that $w(X,t)$ is also continuous down to $t = 0$.

• *Relation with Darcy's law.* Since $\partial X/\partial x = u$, and $\partial X/\partial t = u^{m-1}u_x$, we get along a curve $x = x(t)$ the relation

$$dX = u\,dx + u^{m-1}u\,dt = u\,(dx - \gamma v_x\,dt)\,, \tag{5.9}$$

which shows how the variation of X is related to Darcy's law.

LEMMA 5.2. *Away from the boundary the curves $X(t) =$ constant are precisely the curves on which Darcy's law holds: $dx = \gamma v_x\, dt$.*

It has to be noted that the Darcy law on the boundaries $x = L(t), x = R(t)$ reflects the fact that the lines $X =$ constant approach the boundary when X tends to its end values. We have seen in section 4 that Darcy's law holds in more or less weak form on the boundary depending on its regularity.

6. Uniqueness. We are ready to prove the uniqueness result for Problem (P_1) with specified lateral boundaries.

THEOREM 6.1. *Let v be a solution of Problem (P_1) in a domain Ω with Lipschitz continuous lateral curves in the interval $0 \le t \le T$, $L(t) < R(t)$, and assume that it takes continuously initial data v_0, a continuous and positive function in $I_0 = (L(0), R(0))$, with limit zero at both ends. Then v is unique.*

Proof. We know that $X(x,t)$ is well defined in Ω and ranges in the interval $(A(t), B(t))$ for every $t > 0$, where the limit values $A(t) = X(L(t),t)$, $B(t) = X(R(t),t)$ exist by monotonicity. We have seen that $x = w(X,t)$ satisfies the equation

$$w_t = \frac{1}{m'}(|w_X|^{m'-1} w_X)_X \quad \left(m' = -m = \frac{1-\gamma}{\gamma}\right), \tag{6.1}$$

defined in $\Omega' = \{(X,t) : A(t) < X < B(t)\}$. Moreover, w is continuous at $t = 0$, and the initial data $w(X,0)$ is given by a continuous increasing function, the inverse of $X(x,0)$. The rest of the argument is different for the cases $\gamma \ge 1$ and $0 < \gamma < 1$.

Case $\gamma \ge 1$. It means $0 \le m < 1$, and hence $-1 < m' \le 0$. The local analysis has already shown us that v behaves in a linear way near the interface: $v = O(\rho)$, $\rho = R(t) - x$ for every t. Hence, $u \ge c\,\rho^{-\gamma}$ for $\rho \approx 0$. For $\gamma \ge 1$ we get $B(t) = X(R(t),t) = \infty$. Also $A(t) = X(L(t),t) = -\infty$. We get an important preliminary conclusion for the solutions in that case: the domain Ω' is the strip $\mathbb{R} \times (0,T]$.

Now take two solutions v_1 and v_2 with the same data and perform the operations of passing to X_1, X_2 using (5.1), and then to the conjugate functions w_1 and w_2, to obtain two continuous solutions of (6.1) in Ω' with the same data at $t = 0$. Moreover, both functions are strictly monotone in the variable $X \in \mathbb{R}$ for all t and have the same finite end-values at $X = \pm\infty$, namely, $L(t)$ and $R(t)$.

Now comes the technical point: by adding $\varepsilon > 0$ to w_1, (6.1) is not changed, but the data are now strictly ordered with respect to w_2. The strong maximum principle implies that the solutions $w_1 + \varepsilon$ and w_2 cannot touch for any $t > 0$. Hence, $w_1 + \varepsilon > w_2$. Reversing the roles, we get $w_2 + \varepsilon > w_1$. Letting $\varepsilon \to 0$, we conclude that $w_1(X,t) = w_2(X,t)$. Inverting the transformation, we arrive at $X_1(x,t) = X_2(x,t)$, differentiating at $u_1 = u_2$, and finally at $v_1 = v_2$.

Case $0 < \gamma < 1$. In this case $m' > 0$, the boundary behavior analysis shows that the domain Ω' may be bounded, and we need a further argument. In fact, the problem can be formulated as a parabolic variational inequality (obstacle problem); cf. [27, 28, 33]. It has upper obstacle $w = R(t)$ and lower obstacle $w = L(t)$, and the function $w(X,t)$ solves

$$w_t = \frac{1}{m'}(|w|_X^{m'-1} w_X)_X \qquad \text{whenever } L(t) < w(X,t) < R(t), \tag{6.2}$$

i.e., in the noncoincidence set which is the set Ω', and there the solution is smooth. Note that w and $w_X = U = 1/u = v^\gamma > 0$ are continuous, and

$$w_t = U^{m'-1} U_X = \gamma\, v_x$$

is bounded. The coincidence set is formed by the sets $\{w = L(t)\}$ and $\{w = R(t)\}$, where $w_X = 0$.

The uniqueness argument can now proceed as in the case $\gamma > 1$, and it is easily seen that $w_1 + \varepsilon$ and w_2 cannot touch at a noncoincidence point for both solutions. Since at a touching point $\partial w_1/\partial X = \partial w_2/\partial X$, if one of them is a coincidence point, so is the other, which is impossible in view of the obstacle values. □

7. Equivalent problems. Fast diffusion with $-1 < m' \le 0$. The line of proof of uniqueness followed in section 6 transforms the problem first into the p-Laplacian problem for $X = X(x,t)$ and then into the study of the conjugate p-Laplacian equation (5.6) for $x = w(X,t)$. It can be equivalently done in terms of the differentiated conjugate equation (5.7). In the case $\gamma \ge 1$ we land into the conjugate equations with exponent $m' \in (-1, 0]$. Now, this problem had been thoroughly studied in [26, 39, 40, 41]. The transformations indicate the interesting fact that all these problems are equivalent when properly formulated. Thus, the results of those papers give an alternative proof of the existence of solutions of Problem (P_1) in the range $\gamma \ge 1$ after applying backward the aforementioned transformations.

Conversely, the results of previous sections recover and slightly improve the results of [39]. Thus, according to the theory developed in that paper (and using the present notation for convenience) we can solve the mixed initial and boundary-value

$$U_t = (U^{m'-1}U_X)_X \quad \text{for} \quad -1 < m' \le 0\,, \tag{7.1}$$

posed in $Q = \{(X,t) : X \in \mathbb{R}\,, t > 0\}$, taking initial data

$$U(x,0) = U_0(x)\,, \quad X \in \mathbb{R}. \tag{7.2}$$

Moreover, for every fixed $t > 0$ we can impose the *flux conditions*

$$\begin{cases} U^{m'-1}U_X \to -f(t) & \text{as} \quad x \to \infty\,, \\ U^{m'-1}U_X \to g(t) & \text{as} \quad x \to -\infty\,. \end{cases} \tag{7.3}$$

THEOREM 7.1. *Let U_0 be a nonnegative and integrable function, and let f and g be a pair of flux functions which are nonnegative and belong to $L^1(0,\infty) \cap BV_{loc}(0,\infty)$. Then there exists one and only one solution u of the mixed problem* (7.1)–(7.3), *which is a positive and C^∞ smooth function defined in a strip $Q_T = \mathbb{R} \times (0,T)$ for some $T > 0$. We have $U \in C([0,T] : L^1(\mathbb{R}))$, and the following formula holds:*

$$\int_{\mathbb{R}} U(X,t)\,dx = \int_{\mathbb{R}} U_0(x)\,dx - \int_0^t (f+g)\,dt\,. \tag{7.4}$$

The solution with $f = g = 0$ is the *maximal solution* and conserves mass. It has a different behavior as $|X| \to \infty$ than the solutions with nonzero flux. The integral $\int U(X,t)\,dX = M(t)$ is the *mass* of the solution at time t, and (7.4) is then called the *global mass balance*, for $f = g = 0$ it is called conservation of mass. The formula is valid for all times $t > 0$ if $\int_{\mathbb{R}} U_0(X)\,dX$ is larger than $\int_0^\infty (f+g)\,dt$. Otherwise, there exists a time at which the second member of (1.1) becomes 0. This time is given by

$$T = \sup\left\{ t > 0 : \int U_0(X)\,dX > \int_0^t (f+g)\,dt \right\}. \tag{7.5}$$

If T is finite, our solution vanishes identically as $t \to T$, and the equation ceases to have a meaning for $t \ge T$. Such T is called the *extinction time*.

A very interesting property of this mixed problem is the stability property, which can be stated as follows. Let U_i, $i = 1, 2$, be two solutions with initial data U_{0i} and flux functions f_i, g_i, all of them satisfying the properties stated above. Then

$$\int_{\mathbb{R}} (U_1(t) - U_2(t))_+ \, dx \le \int_{\mathbb{R}} (U_{01} - U_{02})_+ \, dx + \int_0^t \{(f_2 - f_1)_+ + (g_2 - g_1)_+\} \, dt \tag{7.6}$$

as long as both solutions exist. As a consequence, we have the following maximum principle: *If* $U_{01} \le U_{02}$, $f_2 \le f_1$, *and* $g_2 \le g_1$, *then* $T_1 \le T_2$ *and* $U_1 \le U_2$ *in* Q_{T_1}.

Use of the conjugate equations. If we start from the set of solutions $v(x,t)$ of Problem (P_1) described in preceding sections in the range $0 \le m < 0$, passing to the density $u(x,t)$ and then applying the transformations, we obtain solutions $U(X,t)$ of the fast diffusion equation for $-1 < m' = -m \le 0$, and $U = X_x$, where X is the mass variable associated to u that satisfies a p-Laplace equation with $p = m' + 1$. Using our results we have the following extension of the previous existence result.

THEOREM 7.2. *The result of Theorem* 7.1 *holds for all flux functions* f *and* g *which are nonnegative, integrable, and locally bounded. The boundary data are taken in an integral sense. The stability property holds.*

Proof. We start from the pressure equation with initial data

$$v_0(t) = U_0(X)^{1/\gamma}, \quad x = \int_0^X U_0(s) \, ds$$

for (1.9) and interfaces given by $R'(t) = -f(t)$, $R(0) = \int_0^\infty U_0(s) \, ds$, and $L'(t) = g(t)$, $L(0) = \int_0^{-\infty} U_0(s) \, ds$. The rest is easy. □

8. The case $\gamma = 0$. There are many similarities between the range $\gamma > 0$ studied in previous sections and the limit case $\gamma = 0$, where (1.1) is written

$$v_t = v \, v_{xx}. \tag{8.1}$$

Actually, it has been proved some years ago that Problem (P_1) can be solved for different contracting interfaces [43, 25] and that the maximal solutions of the Cauchy problem have stationary interfaces. However, there are also strong differences with the case $\gamma > 0$ since the absence of the term in $(v_x)^2$ implies that the same form of Darcy's law cannot hold. There is, however, a *natural formulation of Darcy's law* that we want to present here. But we have to work a bit more and introduce "correct definitions" of density and pressure.

Indeed, we can pass from (8.1) to a porous-medium-type equation by means of the transformation $v = e^{-u}$, $u = \log(1/v)$, so that the new "density" u satisfies

$$u_t = (e^{-u} u_x)_x = (-e^{-u})_{xx}. \tag{8.2}$$

Observe that the correspondence $v \mapsto u$ is monotone decreasing (as in the fast diffusion cases $\gamma < 0$, which makes sense in view of the subsequent results). In particular, $v = 0$ implies $u = \infty$. It will be convenient to work with positive solutions of (8.2), $u > 0$. This means that we have to restrict our attention to solutions of (8.1) with $0 \le v < 1$. This is really no loss of generality since given a bounded solution v we can always consider its scaled version

$$\tilde{v}(x,t) = \frac{1}{\lambda} \, v(\lambda x, t) \, ,$$

which can be made less than 1 by suitably choosing $\lambda > 1$ and is again a solution of (8.1).

A natural pressure. We need an argument from physics. When we revise the model equation for gases in porous media and write it as conservation of mass, $u_t + (u\,\mathbf{V})_x = 0$, we see that it predicts for the case of (8.2) a particle velocity $\mathbf{V} = (e^{-u}/u)u_x$. Darcy's law is then written as $\mathbf{V} = -\nabla\pi$ for a "pressure function" π which is given in terms of u or v by

$$\pi(u) = -\int_{\infty}^{u} \frac{e^{-u}}{u}\,du = \int_{0}^{v} \frac{dv}{\log(1/v)} \tag{8.3}$$

(up to a constant). We notice that π behaves for $v \sim 0$ like

$$\pi(v) \sim \frac{v}{\log(1/v)},$$

which is $o(v)$, while $\pi(v) \to \infty$ as $v \to 1$. This reminds us of our restriction on v.

Going back to our analysis, it can be seen that $\pi(x,t)$ satisfies

$$\pi_t = v\,\pi_{xx} - (\pi_x)^2, \tag{8.4}$$

which points to the correct form of Darcy's law as $\pi_t = -(\pi_x)^2$ or, in terms of the interface, as $R'(t) = \pi_x$.

Traveling waves. We consider solutions of the form $v(x,t) = V(x+ct)$ and then V satisfies the equation $VV'' = cV'$, which upon integration gives

$$V' = c\,\log(V) - K, \quad K \in \mathbb{R}. \tag{8.5}$$

We integrate again from $V(0) = 0$. For $K = 0$ we get *linear waves for the pressure*, $\pi = c\,(-s)_+$. For $K \neq 0$ the behavior near $s = 0$ is the same, but the behavior for $s < 0$ changes monotonically with K exactly as described in section 2. Darcy's law is satisfied for these solutions in the modified form just explained.

Existence and regularity. There is no difficulty in redoing the analysis of section 3 to get maximal and minimal solutions of Problem (P_1), nor that of section 4 to describe the behavior near a Lipschitz free boundary. We get the following.

THEOREM 8.1. *If $R(t)$ and $L(t)$ are Lipschitz continuous monotone curves in the interval $0 \le t \le T$ with $L(t) < R(t)$, there exists a unique solution to Problem (P_1) in the nonexpanding domain Ω. The pressure π behaves in a linear way near the Lipschitz parts of the free boundary, and moreover the Darcy law is satisfied on the interfaces for almost all $0 < t < T$. If $R' < 0$ (resp., $L' > 0$) the behavior is nondegenerate. If $R(t)$ (resp., $L(t)$) is C^1, then π is C^1 up to the boundary and Darcy's law holds in the classical sense,*

$$R'(t) = \pi_x \quad \text{on } x = R(t), \qquad L'(t) = \pi_x \quad \text{on } x = L(t). \tag{8.6}$$

The maximum principle applies to the solutions of this mixed problem.

Uniqueness. We still have to prove this part of the theorem, which entails describing the modifications that have to be performed on the mass variable and the Bäcklund transform. These are quite interesting parts of the theory. The mass variable associated to u is

$$X(x,t) = \int_{\Gamma} u dx + e^{-u}\,u_x dt, \tag{8.7}$$

integrated along any curve Γ lying in the domain of u and joining the points $(0,0)$ and (x,t). Therefore, the function $X = W(x,t)$ satisfies

$$W_x = u, \quad W_t = e^{-u}\, u_x,$$

so that a generalization of the p-Laplacian equation (5.3) gives

$$W_t = \mathtt{exp}\,(-W_x)\, W_{xx}. \tag{8.8}$$

The transformation is invertible and the inverse function $x = w(X,t)$ satisfies

$$w_X = 1/u, \quad w_t = -(e^{-u}/u)\, u_x = -e^{-u}\, u_X,$$

so that

$$w_t = e^{-1/w_X}\, w_X^{-2}\, w_{XX} = \left(e^{-1/w_X}\right)_X, \tag{8.9}$$

which is the conjugate equation to (8.8). We remark that the map $s \mapsto -e^{-1/s}$ is a monotone increasing function defined for all $s \geq 0$, which generalizes the porous medium equation. Since all its derivatives at zero are zero it has a very slow diffusion property, and it has been studied as such in a number of papers; cf. [31, 32]. To end this excursion we list the Bäcklund transform that maps (x,t,u) into (X,t,U) according to the above formulas, and then $U = U(X,t) = w_X(X,t)$ satisfies the equation

$$U_t = \left(e^{-1/U}\right)_{XX}, \tag{8.10}$$

the Bäcklund conjugate to (8.2).

With all this material the proof of uniqueness performed in section 6 can be copied with only obvious changes.

9. Local behavior near a corner of the interface. In this section we introduce the study of the regularity of the solutions near a nonsmooth interface with a typical example. We consider the case of a piecewise linear interface with an angle point. W.l.o.g. we assume that the angle is located at $x = 0$, $t = 0$. For $t \leq 0$ we consider the linear TW $v = A(x - c_1\, t)_+$, with $A > 0$ and $c_1 = \gamma A$. In order to continue such a solution for $t > 0$ we will prove that there exists a unique self-similar solution that solves the problem

$$\begin{cases} v_t = v v_{xx} - \gamma |v_x|^2, \\ v_0(x) = A\, x_+, \\ L(t) = ct, R(t) = \infty \end{cases} \tag{9.1}$$

for all A and $c > 0$. Existence is easy after passing to the limit on solutions with compact initial data and the same interface, which are ordered. Then uniqueness comes from the mass transformation, and even from the argument at the end of section 2, replacing scaling by shifting (we leave this exercise to the reader). Self-similarity comes from invariance, a well-known argument.

We now remark that in the case $c = c_1$ the solution is the same linear wave already described for $t \leq 0$. But for other values of c the interface suddenly bends (it has a corner at $t = 0$, $x = 0$). The profile has the form

$$v = t f(x/t)$$

for a monotone function $f = f(s)$ such that

$$f(s)f''(s) = \gamma(f'(s))^2 + f(s) - sf'(s),$$

with $f(c) = 0$, $f(s) > 0$ for $s > c$, and $f'(\infty) = A$. We also have $f'(c+) = c/\gamma$. The existence and uniqueness theory for the PDE problem implies that this ODE problem can then be solved for all $A, c > 0$ if $\gamma > 0$, whereas it can only be solved for $c = -A$ if $\gamma < 0$. A simple scaling allows us to set $A = 1$ w.l.o.g., and then c is replaced by c/A.

Stationary interface. We can put $c = 0$ in the above formula to get the *maximal solution* after the angle, i.e, the one having a fixed (stationary) interface. It is given by $f(s) = s - 1 + e^{-s}$, which produces

$$v(x,t) = x - t + t\, e^{-x/t} \quad \text{for} \quad x > 0, \tag{9.2}$$

and $v = 0$ for $x \le 0$, $t > 0$. We see the typical quadratic form of stationary interfaces [23]

$$v(x,t) \sim x^2/(2t_0) \quad \text{as} \quad (x,t) \to (0,t_0)$$

with $t_0 > 0, x > 0$. This means that at $t_0 > 0$ the formula realizes the *connection* between the linear profile $V_1(x,t) = x - t$, which is the behavior for $x \sim \infty$, and the quadratic solution $V_2(x,t) = |x|^2/(2t)$ for $x \sim 0$. Note also that $v = V_1$ for $t \le 0$, and that v agrees with V_1 with a C^∞ contact at $t = 0+$ (not analytically, of course). On the other hand, $v_x = 1 - e^{x-t}$ and v_t are discontinuous at the corner $(0,0)$; this is the way the corner of the interface is reflected on the regularity of the solution. Note also that $v_{xx} = e^{x-t}/t$ is positive in Ω. For a reference to behavior near a corner point when $\kappa > 0$, see [5].

Explicit solution for $c \neq 0$. Betelú [17] pointed out that the case $\gamma = 1$ can be explicitly solved. Indeed, in this case the equation has a first integral to $f' = s - kf$ for an arbitrary constant k, and we arrive at

$$f(s) = A\, \left(s - A - B\, e^{-s/A}\right)_+,$$

where $A = 1/k$ and B are free constants. Let us put w.l.o.g. $A = 1$. The value of c comes from the relation $f(c) = 0$, which gives for c the relation $B = (c-1)e^c$ so that f can be written in terms of the moving coordinate $y = s - c$ as

$$f(s) = \left(y + (c-1)(1 - e^{-y})\right)_+. \tag{9.3}$$

We see that $c = 1$ is the standard linear TW. We can let c vary over the range $0 < c < \infty$ to get all possible moving interfaces. A final observation: it is easily seen that for $x - ct = o(t)$ we get the estimate

$$v(x,t) \sim c(x - ct),$$

which shows that this solution is a *continuous connection* between the linear profile ax_+ of a TW and the linear TW profile that corresponds to the interface we have chosen (as expected).

10. Conclusions and comments. Here is a list of additional comments on the theory developed in the article.

• *On the dependence with κ.* From the point of view of the theory of (1.6) there are three important critical values: $m = 1$, which marks the limit of the property of finite propagation; $m = 0$, which marks the onset of nonuniqueness of the Cauchy problem; and $m = -1$, which marks the limit of the region of nonexistence for small data. It seems that only $m = 1$, i.e., $\gamma = 0$, is important for the present discussion of (1.9). A deeper understanding is needed of the case $\gamma = 0$, transition from $\gamma > 0$ to $\gamma < 0$; cf. the work of Bertsch and collaborators and also [10, 42].

• *On the difference of free boundary behavior between $\kappa > 0$ and $\kappa < 0$.* The free-boundary problem for the case $\kappa = -\gamma > 0$ is uniquely determined in its classical free-boundary formulation by the weaker conditions on the free boundaries:

$$v = 0, \quad (v^{1+\kappa})_x = 0. \tag{10.1}$$

The latter is easily recognized in the form $(u^m)_x = 0$, i.e., zero flux for (1.6). Even weaker is the integral formulation called weak solution. On the other hand, contracting solutions for $\kappa < 0$ have $(u^m)_x = (v^{1+\kappa})_x \sim uv_x \to \infty$.

• *Incorrect problems.* There is a very interesting way of comparing the theories for positive and negative κ. Thus, let us try to solve the overdetermined problem (1.1)–(1.3) with a preassigned choice of interface, say $R(t)$, on which we want the full boundary conditions (1.3) to hold. When $\kappa > 0$, if the choice is not the precise free-boundary curve (which is a unique choice as we have said), then the boundary behavior will be of the form $(u^m)_x = (v^{1+\kappa})_x = -c(t) \neq 0$ if we commit an error from below, while we will have $v = 0$ in a neighborhood of $R(t)$ if we overshoot (i.e., if we have located the boundary too far). This last situation also happens if we try to solve the problem for $\kappa < 0$ with any expanding support. However, in the fast diffusion case we can solve *all problems* with a shrinking interface and full-boundary conditions (1.3).

• *On the class of initial data.* We have imposed a certain number of restrictions on the problem in order to make the presentation easier to read. Thus, there is no need to have two interfaces; the calculations can be repeated without difficulty if v_0 is supposed to be continuous, positive, and bounded on the interval $(-\infty, 0)$ and zero for $x > 0$. Then we get only one interface.

The case of multiple interfaces offers no special difficulty, but extra attention is needed in defining the mass variable (or the Bäcklund transform), as happens when data are not continuous but only integrable.

Moreover, the condition of boundedness on the data is made for convenience, and there are optimal growth conditions on the data that guarantee the existence.

The fact that the main regularity estimates for these nonlinear equations are local is well known; cf., for instance, [22, 46].

• *More general bounding curves.* We have assumed that the lateral curves $x = L(t)$, $x = R(t)$ are monotone and Lipschitz continuous. There is a natural extension of our study to the case where these curves are assumed only to be monotone and continuous. This study can be done and much of the above theory holds. However, there appears a problem with the regularity of the solutions, namely, that the maximal solution may cease to be continuous at non-Lipschitz points of the interfaces. This entails a number of interesting geometrical consequences, like the existence of so-called *needles* for the conjugate fast diffusion equation, which deserve a separate study that will appear soon [48].

• *Nonuniqueness with constant support.* Interesting nonuniqueness questions, related to the way initial traces are defined, may appear even for solutions with constant support. This happens for more complicated initial data which vanish inside the initial interval. The subject is discussed in [15, 23]. In [22], joint work with Chasseigne, the question of expanding singular sets for the fast diffusion equation is addressed in terms of nontrivial singular measures as second members. Upon translation to the pressure formulation we thus obtain an alternative explanation for the classes of solutions with contracting supports.

• *Nonshrinking solutions.* On the other hand, it must be remarked that solutions which do not touch the level $v = 0$ do not necessarily shrink in size or go to zero with time somewhere. The simplest examples are of course the constant solutions. But we may also consider the TWs, $v = V(x + ct)$, of the family (2.2) which satisfy

$$V' = -\frac{c}{\gamma} + H\,V^{\gamma}, \quad H > 0.$$

There is always the possibility, mentioned in section 2, of considering solutions with $V(-\infty) = (c/\gamma\,H)^{1/\gamma}$ and $V(s) > V(-\infty)$ for $s \in \mathbb{R}$. In that case $V_t = c\,V' > 0$. For the choice $\gamma = 1$ the solution is explicit and has been known for some time. If $H = c = 1$, we obtain $v(x,t) = 1 + e^{x+t}$.

• *On well-posedness.* It should be noted that when we take initial data of the forms given by the explicit solutions, which are analytic, there are infinitely many solutions with high degrees of smoothness depending on the choice of contracting support. In particular, the solution produced by the vanishing viscosity method is never the contracting solution.

We conclude that, in the line of references [1] and [9], there is an interest in finding the conditions that make the problems with contracting supports fully determined without the actual specification of the future support evolution. Stability questions can play a role in this analysis; cf. [2, 9, 12, 24].

• *On more general nonlinear diffusion equations.* The main tools we use in the proofs, like the relation pressure-density, the mass variable, the p-Laplacian equations and the conjugate equations, and Bäcklund Transform can be applied to a more general situation, the so-called *filtration equation*

$$u_t = \Phi(u)_{xx}, \tag{10.2}$$

where Φ is a continuous nondecreasing function or even a maximal monotone graph. Much of the above theory translates, but general Φ request some delicate technicalities. We hope to develop the detailed machinery in a forthcoming publication [49].

• *On several-dimensional problems.* There is no difficulty in generalizing our problem to the several-dimensional setting by considering the pressure equation

$$v_t = v\,\Delta v + \kappa\,|\nabla v|^2 \tag{10.3}$$

and the associated PME after the change $v = u^{m-1}$

$$u_t = \nabla \cdot (u^{m-1}\nabla u). \tag{10.4}$$

However, the magic of the associations with other equations using the mass variable and the Bäcklund transform has no counterparts for $N > 1$. The study of the N-dimensional problem therefore needs new tools. For stationary interfaces, see [23].

As a manifestation of the wealth of solutions of this problem in several dimensions, Betelú has examined the existence of TWs with nonplanar fronts and found for $\gamma = 1$ the family of explicit solutions

$$v(x,y,t) = c(x - ct - ar)_+, \quad r = ((x-ct)^2 + y^2)^{1/2}, \tag{10.5}$$

where $c > 0$ is the wave speed and $a < 1$ is a real parameter. For $|a| < 1$ the domain of positivity of the wave is the angular region $\{(x,y) : |y| < d(x-ct)\}$, $d = (1-a^2)^{1/2}/a$, which moves with speed c in the direction of the positive x-axis. For $a = 0$ we recover the standard plane TW $v(x,y,t) = c(x-ct)_+$. Otherwise, the region is convex if $a > 0$, concave if $a > 0$. The explicit family does not generalize to $\gamma \neq 1$, but still interesting solutions exist.

Acknowledgments. Discussions with G. Barenblatt and collaborators were possible by an invitation of the University of California at Berkeley. I thank A. Chertok for a copy of her unpublished work and S. Betelú for his explicit examples and for numerical work to confirm our assertions. I would also like to acknowledge fruitful discussions with D. G. Aronson, L. Caffarelli, and A. Friedman, and am grateful to the referees for several useful comments that have allowed me to improve the previous version of this work.

REFERENCES

[1] S. B. Angenent, *Local existence and regularity for a class of degenerate parabolic equations*, Math. Ann., 280 (1988), pp. 465–482.
[2] S. B. Angenent, *Large time asymptotics for the porous media equation*, in Nonlinear Diffusion Equations and Their Equilibrium States, I, Math. Sci. Res. Inst. Publ. 12, Springer-Verlag, New York, 1988, pp. 21–34.
[3] D. G. Aronson, *The porous medium equation*, in Some Problems of Nonlinear Diffusion, Lectures Notes in Math. 1224, Springer-Verlag, New York, 1986, pp. 1–46.
[4] D. G. Aronson, L. A. Caffarelli, and S. Kamin, *How an initially stationary interface begins to move in porous medium flow*, SIAM J. Math. Anal., 14 (1983), pp. 639–658.
[5] D. G. Aronson, L. A. Caffarelli, and J. L. Vázquez, *Interfaces with a corner point in one-dimensional porous medium flow*, Comm. Pure Appl. Math., 38 (1985), pp. 375–404.
[6] O. Baconneau and A. Lunardi, *Smooth Solutions to a Class of Free Boundary Parabolic Problems*, preprint.
[7] G. I. Barenblatt, *On self-similar motion of compressible fluids in porous media,* Prikl. Mat. Mekh., 16 (1952), pp. 679–698 (in Russian).
[8] G. I. Barenblatt, *Scaling, Self-Similarity, and Intermediate Asymptotics*, Cambridge Texts Appl. Math. 14, Cambridge University Press, Cambridge, UK, 1996.
[9] G. I. Barenblatt, M. Bertsch, A. E. Chertock, and V. M. Prostokishin, *Self-similar asymptotics for a degenerate parabolic filtration-absorption equation*, Proc. Natl. Acad. Sci. USA, 97 (2000), pp. 9844–9848.
[10] G. I. Barenblatt and J. L. Vázquez, *Nonlinear Diffusion and Image Contour Enhancement*, preprint PAM-802, Center for Pure and Applied Mathematics, University of California, Berkeley, 2003.
[11] J. Bear, *Dynamics of Fluids in Porous Media*, Dover, New York, 1972.
[12] J. G. Berryman and C. J. Holland, *Stability of the separable solution for fast diffusion,* Arch. Ration. Mech. Anal., 74 (1980), pp. 379–388.
[13] M. Bertsch and R. Dal Passo, *A numerical treatment of a superdegenerate equation with applications to the porous media equation*, Quart. Appl. Math., 48 (1990), pp. 133–152.
[14] M. Bertsch, R. Dal Passo, and M. Ughi, *Discontinuous "viscosity" solutions of a degenerate parabolic equation,* Trans. Amer. Math. Soc., 320 (1990), pp. 779–798.
[15] M. Bertsch, R. Dal Passo, and M. Ughi, *Nonuniqueness of solutions of a degenerate parabolic equation,* Ann. Mat. Pura Appl., 161 (1992), pp. 57–81.
[16] M. Bertsch and M. Ughi, *Positivity properties of viscosity solutions of a degenerate parabolic equation*, Nonlinear Anal., 14 (1990), pp. 571–592.
[17] S. Betelú, *private communication*, 2001.

[18] G. Bluman and S. Kumei, *On the remarkable nonlinear diffusion equation* $\partial/\partial x[a(u+b)^{-2}(\partial u/\partial x)] - (\partial u/\partial t) = 0$, J. Math. Phys., 21 (1980), pp. 1019–1023.

[19] L. A. Caffarelli and A. Friedman, *Regularity of the free boundary for the one-dimensional flow of gas in a porous medium*, Amer. J. Math., 101 (1979), pp. 1193–1218.

[20] L. A. Caffarelli and J. L. Vázquez, *A free-boundary problem for the heat equation arising in flame propagation*, Trans. Amer. Math. Soc., 347 (1995), pp. 411–441.

[21] L. A. Caffarelli and J. L. Vázquez, *Viscosity solutions for the porous medium equation*, in Differential Equations: La Pietra 1996 (Florence), Proc. Sympos. Pure Math. 65, M. Giaquinta et al., eds., AMS, Providence, RI, 1999, pp. 13–26.

[22] E. Chasseigne and J. L. Vázquez, *Extended theory of fast diffusion equations in optimal classes of data. Radiation from singularities*, Arch. Ration. Mech. Anal., 164 (2002), pp. 133–187.

[23] E. Chasseigne and J. L. Vázquez, *The pressure equation in the fast diffusion range*, Rev. Mat. Iberoamericana, to appear.

[24] A. Chertok, *On the stability of a class of self-similar solutions to the filtration-absorption equation*, European J. Appl. Math., 13 (2002), pp. 179–194.

[25] R. Dal Passo and S. Luckhaus, *A degenerate diffusion problem not in divergence form*, J. Differential Equations, 69 (1987), pp. 1–14.

[26] J. R. Esteban, A. Rodríguez, and J. L. Vázquez, *A nonlinear heat equation with singular diffusivity*, Comm. Partial Differential Equations, 13 (1988), pp. 985–1039.

[27] A. Friedman, *Stochastic Differential Equations and Applications*. Vol. 2, Probab. Math. Statist. 28, Academic Press, New York, London, 1976.

[28] A. Friedman *Variational Principles and Free-Boundary Problems*, 2nd ed., Robert E. Krieger, Malabar, FL, 1988.

[29] V. A. Galaktionov, S. I. Shmarev, and J. L. Vázquez, *Regularity of interfaces in diffusion processes under the influence of strong absorption*, Arch. Ration. Mech. Anal., 149 (1999), pp. 183–212.

[30] V. A. Galaktionov, S. I. Shmarev, and J. L. Vázquez, *Regularity of solutions and interfaces to degenerate parabolic equations. The intersection comparison method*, in Free Boundary Problems: Theory and Applications (Crete, 1997), Chapman & Hall/CRC Res. Notes Math. 409, Chapman & Hall/CRC, Boca Raton, FL, 1999, pp. 115–130.

[31] V. A. Galaktionov, R. Kersner, and J. L. Vázquez, *Asymptotic behaviour for an equation of superslow diffusion in a bounded domain*, Asymptot. Anal., 8 (1994), pp. 237–246.

[32] V. A. Galaktionov and J. L. Vázquez, *Asymptotic behaviour for an equation of superslow diffusion. The Cauchy problem*, Asymptot. Anal., 8 (1994), pp. 145–159.

[33] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Pure Appl. Math. 88, Academic Press, New York, London, 1980.

[34] J. R. King, *Self-similar behaviour for the equation of fast nonlinear diffusion*, Philos. Trans. Roy. Soc. London Ser. A, 343 (1993), pp. 337–375.

[35] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'tseva, *Linear and Quasilinear Equations of Parabolic Type,* Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.

[36] M. Muskat, *The Flow of Homogeneous Fluids through Porous Media*, McGraw-Hill, New York, 1937.

[37] L. A. Peletier, *The porous medium equation*, in Applications of Nonlinear Analysis in the Physical Sciences, H. Amann et al., eds., Pitman, London, 1891, pp. 229–241.

[38] P. Ya. Polubarinova-Kochina, *Theory of Groundwater Movement*, Princeton University Press, Princeton, NJ, 1962.

[39] A. Rodríguez and J. L. Vázquez, *A well posed problem in singular Fickian diffusion*, Arch. Ration. Mech. Anal., 110 (1990), pp. 141–163.

[40] A. Rodríguez and J. L. Vázquez, *Maximal solutions of singular diffusion equations with general initial data*, in Nonlinear Diffusion Equations and Their Equilibrium States, 3, N. G. Lloyd et al., eds., Birkhäuser Boston, Boston, MA, 1992, pp. 471–484.

[41] A. Rodríguez and J. L. Vázquez, *Nonuniqueness of solutions of nonlinear heat equations of fast diffusion type*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 12 (1995), pp. 173–200.

[42] A. Rodríguez and J. L. Vázquez, *Obstructions to existence in fast-diffusion equations*, J. Differential Equations, 184 (2002), pp. 348–385.

[43] M. Ughi, *A degenerate parabolic equation modelling the spread of an epidemic*, Ann. Mat. Pura Appl. (4), 143 (1986), pp. 385–400.

[44] J. L. Vázquez, *An introduction to the mathematical theory of the porous medium equation*, in Shape Optimization and Free Boundaries (Montréal, PQ, 1990), NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 380, Kluwer Academic, Dordrecht, The Netherlands, 1992, pp. 347–389.

[45] J. L. Vázquez, *Nonexistence of solutions for nonlinear heat equations of fast-diffusion type*, J. Math. Pures Appl., 71 (1992), pp. 503–526.
[46] J. L. Vázquez, *Regularity of solutions and interfaces of the porous medium equation via local estimates*, Proc. Roy. Soc. Edinburgh Sect. A, 112 (1989), pp. 1–13.
[47] J. L. Vázquez, *The free boundary problem for the heat equation with fixed gradient condition*, in Free Boundary Problems, Theory and Applications (Zakopane, 1995), Pitman Res. Notes Math. Ser. 363, Longman, Harlow, UK, 1996, pp. 277–302.
[48] J. L. Vázquez, *Positivity, Propagation Properties and Needles in Nonlinear Singular Diffusion*, in preparation.
[49] J. L. Vázquez, *One-Dimensional Nonlinear Diffusion*, Course Notes, UAM 1999, book in preparation.

SIAM J. MATH. ANAL.
Vol. 35, No. 4, pp. 1029–1041

A QUASI-LINEAR PARABOLIC SYSTEM FOR THREE-PHASE CAPILLARY FLOW IN POROUS MEDIA*

HERMANO FRID† AND VLADIMIR SHELUKHIN‡

Abstract. A parabolic system is proposed to describe three-phase capillary flows in porous media. The assumptions imposed on the phase interaction mean that the capillarity matrix is triangular. The unique solvability of an associated nondegenerate parabolic system is proved for the Cauchy problem in a class of x-periodic solutions.

Key words. porous media, three-phase capillary flows, existence, uniqueness

AMS subject classifications. Primary, 35K55; Secondary, 35B65, 76S05, 76T05

DOI. 10.1137/S0036141002402165

1. Introduction.

We consider a quasi-linear parabolic problem

$$u_t + f(u)_x = (D(u)u_x)_x, \quad u|_{t=0} = u_0(x), \quad x \in \mathbb{R}, \tag{1.1}$$

related to three-phase capillary flows in porous media, e.g., the flow of oil, gas, and water in a reservoir. Here,

$$u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad f = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}, \quad D = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix},$$

where, say, u_1 and u_2 are the oil and gas saturations, respectively, while $1 - u_1 - u_2$ is the water saturation. The meaning of the flux vector f and the capillary matrix D is explained below. We study unique global solvability. The main assumptions about matrix D are

$$D_{21} \equiv 0, \quad D_{22} \equiv D_{22}(u_2) > 0, \quad D_{11}(u_1, u_2) > 0. \tag{1.2}$$

Global existence for (1.1) in the most general setting is an open problem. Global well-posedness is proved in [8] if $D_{11} = D_{22}$ and $D_{12} = D_{21} = 0$. The restriction (1.2) is also present in the theory of H. Amann [2], but therein it is also assumed that $f_2 = f_2(u_2)$. In our case the flux vector f may be rather general.

Since we are interested in the application to the problem of the three-phase flow where u is a saturation vector, we should impose the restriction

$$u \in \Delta = \{u : u \in \mathbb{R}^2, \quad 0 \le u_i \le 1, \quad u_1 + u_2 \le 1\}. \tag{1.3}$$

We prove a global existence theorem for system (1.1) with the verification of (1.3) with the use of a slight adaptation of the principle of positively invariant regions of Chueh, Conley, and Smoller [5] (see also [11]).

*Received by the editors January 22, 2002; accepted for publication (in revised form) February 28, 2003; published electronically November 4, 2003.

http://www.siam.org/journals/sima/35-4/40216.html

†Instituto de Matemática Pura e Aplicada–IMPA, Estrada Dona Castorina, 110, Rio de Janeiro, RJ 22460-320, Brazil (hermano@impa.br). The research of this author was partially supported by CNPq grants 352871/96-2, 46.5714/00-5, and 479416/2001-0 and FAPERJ grant E-26/151.890/2000.

‡Lavrentyev Institute of Hydrodynamics, Novosibirsk, 630090, Russia (shelukhin@hydro.nsc.ru). The research of this author was partially supported by Russian Fund of Fundamental Research grant 99-01-00622, INTAS grant 01-868, and CNPq grants 46.5714/00-5 and 479416/2001-0.

A new feature in the present paper is concerned with the constitutive law of the capillary pressures $P_1(u)$ and $P_2(u)$. These functions are defined by

$$P_1 = p_1 - p_3, \quad P_2 = p_2 - p_3, \tag{1.4}$$

where p_i is the pressure in the ith phase $i \in \{1, 2, 3\}$. It is shown below how the functions P_i define the capillarity matrix D.

Very little is known about the functions P_i either theoretically or experimentally. An approximation of the real three-phase flows by two artificial two-phase flows is proposed in [13]. In [10], some engineering formulas are given by interpolation into the triangle Δ (see (1.3)) of equalities (1.4), which are assumed known at the pieces of the boundary $\partial\Delta$ where $u_1 = 0$, $u_2 = 0$, and $u_1 + u_2 = 1$, respectively, which correspond to two-phase flows. A class of capillary pressures P_1 and P_2, efficient from the computational point of view, is considered in [4].

It is proved in section 4 that the equalities in (1.2), when imposed to the constitutive equations for the three-phase flow system, are equivalent to a linear system of differential equations for P_1 and P_2 (see (4.9) below), for which we give explicit solutions in some simple cases taken as examples.

Whatever capillary pressures in (1.4) underlie a three-phase model, the governing equations (1.1) should preserve the saturation bounds (1.3) as time grows, provided that the initial saturations obey it. From the analysis in [5], it is known that matrices $\nabla f(u)$ and $D(u)$ must satisfy certain conditions in order to yield this invariance requirement. Below (see (1.6)) we formulate conditions for $D(u)$, which, together with the properties of $f(u)$ (see (1.5)), guarantee (1.3). As will become clear in section 4, physical capillarity matrices (together with their perturbations by multiples of the identity matrix) satisfy the corresponding conditions.

Three-phase flow equations without capillarity effects have been addressed in a number of papers. It is well known that the corresponding system may fail to be hyperbolic. Indeed, an example is given in [3] in which the system is elliptic in a region inside the triangle of saturations and hyperbolic elsewhere in the triangle. The asymptotic behavior of measure-valued solutions was studied in [6]. Additional information on the subject, including references, can be found in [1, 14].

We study the Cauchy problem (1.1) in a class of x-periodic functions with period 2:

$$u(t, x) = u(t, x \pm 2), \quad x \in \mathbb{R}.$$

We denote

$$\Omega = (-1, 1), \quad Q = (0, T) \times \Omega, \quad \mathbb{R}^2_+ = (0, \infty) \times \mathbb{R}.$$

Motivated by the model for three-phase flow in porous media, we assume that function $f(u)$ in (1.1) satisfies

$$\begin{cases} f_i(u) = 0 & \text{if } u_i = 0,\ i = 1, 2, \\ f_1(u) + f_2(u) = 1 & \text{if } u_1 + u_2 = 1. \end{cases} \tag{1.5}$$

We remark that, for the mathematical analysis developed in sections 2 and 3, any other constants could replace 0 and 1 in the right-hand sides of equations (1.5).

It is also assumed that the D_{ij} are smooth functions over Δ satisfying relations (1.2) and

$$\begin{cases} D_{12}(u) = 0 & \text{if } u_1 = 0,\ 0 \le u_2 \le 1, \\ D_{11}(u) - D_{12}(u) - D_{22}(u) = 0 & \text{if } u_1 + u_2 = 1,\ 0 \le u_1, u_2 \le 1. \end{cases} \tag{1.6}$$

The above relations are always verified in the model for three-phase capillary flows in porous media when $D_{21} \equiv 0$.

The main result is the following.

THEOREM 1.1. *Let functions $f_i(u)$ and $D_{ij}(u)$ and their first derivatives be Hölder continuous with the Hölder exponent $\beta \in (0,1)$. Let hypotheses* (1.2), (1.5), *and* (1.6) *about f and D be satisfied and let $u_0 \in H^{2+\beta}(\mathbb{R})$ be periodic with period* 2 *and assuming values in Δ. Then the Cauchy problem* (1.1) *has a unique solution $u(t,x) \in H^{2+\beta,1+\frac{\beta}{2}}(\mathbb{R}^2_+)$, which is periodic in x with period* 2 *and taking values in Δ.*

At this point we would like to make a remark concerning the restriction to periodic boundary conditions. Dirichlet conditions are a more natural kind of boundary conditions for applications. The well-posedness of the corresponding problems for system (1.1) is, in general, harder, as anticipated by the scalar case, since it requires a regularity analysis *up to the boundary*. In all cases, the regularity analysis in the interior is the same as that developed here, so we believe the simplification provided by the periodic conditions helps to highlight the core of the general method. The initial boundary value problem with Dirichlet conditions for (1.1) will be addressed in a forthcoming paper [7].

The remainder of this paper is organized as follows. Sections 2 and 3 are dedicated to the proof of Theorem 1.1. In section 2 we start the mathematical analysis. We obtain a priori estimates which allow the solution of the global existence question through an application of the Leray–Schauder fixed point theorem. In section 3 we complete the proof of the Theorem 1.1, discussing the final steps for the application of the above-mentioned fixed point theorem and the uniqueness question. Section 4 is devoted to a discussion of the model for capillary three-phase flows in porous media, departing from the mass conservation equations and Darcy's law. We describe the basic equations of three-phase flow in porous media, analyze the form of the capillarity matrix, and give examples. We close the section by providing details about the construction of the examples given. The reader mainly interested in the discussion of the model of three-phase capillary flows may jump directly to section 4; we call special attention to our model for capillary pressures.

2. A priori estimates. In this section we prove some a priori estimates for the periodic solutions of the Cauchy problem for (1.1) with periodic initial data subjected to hypotheses (1.2), (4.4), (1.5), and (1.6). We first consider the perturbed problem

$$u_t + f(u)_x = (D(u)u_x)_x + \varepsilon h(u), \quad u|_{t=0} = u_0(x), \tag{2.1}$$

for f, D as above and $\varepsilon > 0$ arbitrary. We prove a priori estimates independent of ε for (2.1), and the same a priori estimates for (1.1) will follow by a simple continuity argument letting $\varepsilon \to 0$. Here u_0 is a periodic function with the period 2. From now on, it is assumed that any x-periodic function has period 2. We denote the solution of (2.1) also by u without reference to ε until the study of the convergence problem when $\varepsilon \to 0$.

Observe that the triangle Δ in (1.3) can be defined as an intersection as follows:

$$\Delta = \cap_1^3 \{G_i(u) \le 0\}, \quad G_1 = -u_1, \quad G_2 = -u_2, \quad G_3 = u_1 + u_2 - 1.$$

Function h is assumed to satisfy the inequalities

$$\nabla_u G_i(u) \cdot h(u) < 0 \quad \text{over } \partial\Delta \cap \{G_i = 0\},\ i \in \{1,2,3\}. \tag{2.2}$$

For instance, we may choose $h(u) = u_* - u$, where u_* is any point in the interior of Δ.

LEMMA 2.1. *Any smooth periodic solution u of problem* (2.1) *satisfies the inclusion*

$$u(t,x) \in \Delta \quad \forall (t,x) \in Q = (0,T) \times \Omega, \quad \Omega = \{x : -1 < x < 1\}. \tag{2.3}$$

Proof. We apply a slight adaptation of the method of positively invariant regions [5] (see also [11, 12]). According to that method, inclusion (2.3) results from the following three conditions.

(i) $\nabla_u G_i(u)$ is an eigenvector of the matrices $\nabla f(u)^\top$ and $D^\top$ for any u such that $G_i(u) = 0$;

(ii) if $D^\top \nabla_u G_i(u) = \mu_i(u) \nabla_u G_i(u)$, with $\mu_i \geq 0$ and $G_i(u) = 0$, then

$$G_i''(u)(D(u)\eta, \eta) \geq 0$$

whenever $\eta \cdot \nabla_u G_i(u) = 0$;

(iii) $h(u) \cdot \nabla_u G_i(u) < 0$ for any $u \in \partial\Delta \cap \{G_i = 0\}$, $i \in \{1,2,3\}$.

Due to the choices of G_i and h, conditions (ii) and (iii) are satisfied. Let us verify condition (i) for $\nabla f(u)$ and $D(u)$. We consider only the case of function G_3, since the other two cases require fewer calculations. Over $\partial\Delta \cap \{G_3(u) = 0\}$, condition (i) is equivalent to both equalities

$$\frac{\partial f_1}{\partial u_1} + \frac{\partial f_2}{\partial u_1} = \frac{\partial f_1}{\partial u_2} + \frac{\partial f_2}{\partial u_2}$$

and

$$D_{11}(u) - D_{12}(u) - D_{22}(u) = 0.$$

The latter follows immediately from (1.6). As to the former, since, by (1.5), $f_1(u) + f_2(u) = 0$ when $u_1 + u_2 = 1$, it follows immediately by differentiating the equation $f_1(u_1, 1-u_1) + f_2(u_1, 1-u_1) = 0$ with respect to u_1. ∎

LEMMA 2.2. *There is a constant c such that*

$$\|u_{1x}\|_{L^2(Q)} + \|u_{2x}\|_{L^2(Q)} \leq c.$$

Proof. Due to $(2.1)_2$,

$$\frac{1}{2}\frac{d}{dt}\int_\Omega u_2^2 \, dx + \int_\Omega D_{22} u_{2x}^2 \, dx = \int_\Omega f_2 u_{2x} + \varepsilon h_2 u_2 \, dx.$$

Hence, one can apply Lemma 2.1, the Cauchy inequality, and the condition $D_{22} \geq \nu$ to conclude that u_{2x} is bounded in $L^2(Q)$. From $(2.1)_1$ we have

$$\frac{1}{2}\frac{d}{dt}\int_\Omega u_1^2 \, dx + \int_\Omega D_{11} u_{1x}^2 \, dx = \int_\Omega f_1 u_{1x} - D_{12} u_{1x} u_{2x} + \varepsilon h_1 u_1 \, dx.$$

By the same argument we arrive at the estimate of u_{1x} in $L^2(Q)$. ∎

LEMMA 2.3. *There are constants $c > 0$ and $\alpha \in (0,1)$ such that*

$$\|u_2\|_{H^{\alpha,\alpha/2}(\bar{Q})} \leq c.$$

Proof. Let $\tilde{\Omega}$ be a bounded open interval containing $\bar{\Omega}$, the closure of Ω. We take a test function ζ with values between 0 and 1, which is different from zero only for $x \in K_\rho$, the ball of radius ρ centered at $x_0 \in \tilde{\Omega}$. We multiply $(2.1)_2$ by

$$\zeta^2 \max\{u_2 - k, 0\} \equiv \zeta^2 u_2^{(k)}, \quad k \in \mathbb{R},$$

and integrate over the cylinder $(s,t)\times K_\rho$. We have

$$\frac{1}{2}\int_{K_\rho}\zeta^2|u_2^{(k)}|^2\,dx\Big|_s^t+\nu\int_s^t\int_{K_\rho}\zeta^2|u_{2x}^{(k)}|^2\,dxdt$$

$$\le\int_s^t\int_{K_\rho}\zeta\zeta_t|u_2^{(k)}|^2-2D_{22}u_{2x}^{(k)}u_2^{(k)}\zeta\zeta_x+2f_2\zeta\zeta_xu_2^{(k)}+f_2\zeta^2u_{2x}^{(k)}+h_2u_2^{(k)}\zeta^2\equiv I.$$

By the Young inequality

$$I\le\int_s^t\int_{K_\rho}\frac{\nu}{2}\zeta^2|u_{2x}^{(k)}|^2+|u_2^{(k)}|^2(|\zeta\zeta_t|+c|\zeta_x|^2+c\zeta^2)+c\zeta^2\chi_{A_{k,\rho}(t)}\,dxd\tau,$$

where $A_{k,\rho}(t)$ is the intersection of the support of $u_2^{(k)}$ with K_ρ, and χ_A is the characteristic function of the set A. These two inequalities imply that function u_2 belongs to the class $\mathcal{B}_2(Q,M,\gamma,r,\delta,k)$ defined in [8], and hence, by Theorem II.7.1 of [8], $u_2\in H^{\alpha,\alpha/2}(\bar Q)$ for some $\alpha\in(0,1)$. □

LEMMA 2.4. *There is a constant $c>0$ such that*

$$\sup_{0\le t\le T}\int_\Omega u_{2x}^2\,dx+\int_Q u_{2xx}^2+u_{2x}^4+u_{2t}^2\,dxdt\le c.$$

Proof. Let $\zeta(t,x)$ be a test function as above. We multiply $(2.1)_2$ by $(u_{2x}\zeta^2)_x$ and integrate over $(0,t)\times K_\rho$ to arrive at the inequality

$$\frac{1}{2}\int_{K_\rho}u_{2x}^2\zeta^2\,dx\Big|_0^t+\nu\int_0^t\int_{K_\rho}u_{2xx}^2\zeta^2\,dxdt\le\int_0^t\int_{K_\rho}u_{2x}^2\zeta\zeta_t-\frac{\partial}{\partial u_2}D_{22}u_{2x}^2u_{2xx}\zeta^2$$
$$-2\frac{\partial}{\partial u_2}D_{22}u_{2x}^3\zeta\zeta_x-2D_{22}u_{2xx}u_{2x}\zeta\zeta_x+\frac{\partial f_2}{\partial u_1}u_{1x}u_{2xx}\zeta^2+\frac{\partial f_2}{\partial u_2}u_{2x}u_{2xx}\zeta^2$$
$$+2\frac{\partial f_2}{\partial u_1}u_{1x}u_{2x}\zeta\zeta_x+2\frac{\partial f_2}{\partial u_2}u_{2x}^2\zeta\zeta_x+\varepsilon h_2(u_{2x}\zeta^2)_x\,dxdt\equiv J.$$

By Lemmas 2.1 and 2.2 and the Young inequality,

$$J\le\frac{\nu}{2}\int_0^t\int_{K_\rho}u_{2xx}^2\zeta^2+c_*u_{2x}^4\zeta^2+c_*\,dxdt.$$

We estimate the last integral, using the inequality (see Lemma II.5.4 of [8])

$$\int_{K_\rho}v_x^4\zeta^2\,dx\le16\mathrm{osc}^2\{v,K_\rho\}\int_{K_\rho}2v_{xx}^2\zeta^2+v_x^2\zeta_x^2\,dx. \tag{2.4}$$

By Lemma 2.3,

$$\mathrm{osc}^2\{u_2,K_\rho\}\le c\rho^\alpha.$$

Now the assertion of the lemma follows if we take ρ such that $32c_*\rho^\alpha<\nu/4$. □

LEMMA 2.5. *There are constants $c>0$ and $\alpha\in(0,1)$ such that $\|u_1\|_{H^{\alpha,\alpha/2}(\bar Q)}\le c$.*

Proof. We consider $(2.1)_1$ as a quasi-linear equation with respect to u_1:

$$u_{1t}-a(t,x,u_1,u_{1x})_x+b(t,x,u_1,u_{1x})=0 \tag{2.5}$$

with

$$
\begin{aligned}
&a(t,x,u_1,p) = D_{11}(u_1,u_2(t,x))p,\\
&b(t,x,u_1,p) = \frac{\partial f_1}{\partial u_1}(u_1,u_2(t,x))p + \frac{\partial f_1}{\partial u_2}(u_1,u_2(t,x))u_{2x}(t,x)\\
&\quad - \left(\frac{\partial D_{12}}{\partial u_1}(u_1,u_2(t,x))p + \frac{\partial D_{12}}{\partial u_2}(u_1,u_2(t,x))u_{2x}(t,x)\right)u_{2x}(t,x)\\
&\qquad - D_{12}(u_1,u_2(t,x))u_{2xx}(t,x) - \varepsilon h_1(u_1,u_2(t,x)).
\end{aligned}
$$

By the above lemmas, we have

$$
ap \geq \nu p^2, \quad |a| \leq \mu p, \quad \mu = \text{const} > 0,
$$

$$
|b(t,x,u_1,p)| \leq p^2 + \varphi(t,x), \quad \left(\int_0^T \left(\int_\Omega \varphi^q \, dx\right)^{r/q} dt\right)^{1/r} \equiv \|\varphi\|_{q,r,Q} \leq \mu
$$

when $q = 2, r = 2$. Moreover, the constants $q = 2$ and $r = 2$ satisfy the restrictions

$$
\frac{1}{r} + \frac{1}{2q} = 1 - \kappa_1, \quad q \in [1,\infty], \quad r \in \left[\frac{1}{1-\kappa_1}, \frac{2}{1-2\kappa_1}\right], \quad 0 < \kappa_1 < \frac{1}{2}, \tag{2.6}
$$

when $\kappa_1 = 1/4$. Thus, functions a and b verify the conditions of [8], Chapter 5, implying the Hölder continuity of u_1 which is a solution of the quasi-linear parabolic equation (2.5). □

LEMMA 2.6. *There is a constant $c > 0$ such that*

$$
\sup_{0\leq t\leq T} \int_\Omega u_{1x}^2 \, dx + \int_Q u_{1xx}^2 + u_{1x}^4 + u_{1t}^2 \, dxdt \leq c.
$$

Proof. By the same argument as in Lemma 2.4,

$$
\begin{aligned}
&\frac{1}{2}\int_{K_\rho} u_{1x}^2\zeta^2 \, dx\Big|_0^t + \nu\int_0^t\int_{K_\rho} u_{1xx}^2\zeta^2 \, dxdt\\
&\qquad \leq \int_0^t\int_{K_\rho} \frac{\nu}{2}u_{1xx}^2\zeta^2 + c(u_{1x}^2 + u_{2x}^2 + u_{2x}^4 + u_{1x}^4\zeta^2 + u_{2xx}^2) \, dxdt.
\end{aligned}
$$

Applying inequality (2.4) for $v = u_1$, we arrive at the conclusion of the lemma. □

LEMMA 2.7. *There are constants $c > 0$ and $\alpha \in (0,1)$ such that $\|u_{2x}\|_{H^{\alpha,\alpha/2}(\bar{Q})} \leq c$.*

Proof. Consider $(2.1)_2$ as an equation for u_2 in the domain $Q_2 = (0,T)\times(-2,2)$:

$$
u_{2t} - \frac{\partial}{\partial x}a(t,x,u_2,u_{2x}) + b(t,x,u_2,u_{2x}) = 0,
$$

with

$$
\begin{aligned}
&a(t,x,u_2,p) = D_{22}(u_2)p,\\
&b(t,x,u_2,p) = \frac{\partial f_2}{\partial u_1}(u_1(t,x),u_2)u_{1x}(t,x) + \frac{\partial f_2}{\partial u_2}(u_1(t,x),u_2)p - \varepsilon h_2(u_1(t,x),u_2).
\end{aligned}
$$

By the above lemmas, we have

$$\nu \le a_p \le \mu, \quad |a| + |a_u| \le \mu(p+1), \quad a_x = 0, \quad |b| \le \mu p^2 + \varphi(t,x), \quad \|\varphi\|_{2q,2r,Q_2} \le \mu$$

if $q = r = 2$. Clearly, these constants q and r satisfy the restrictions (3.6) with $\kappa_1 = 1/4$. These properties of functions a and b imply the estimate of the lemma for the subdomain $Q \subseteq Q_2$, according to the theory of quasi-linear parabolic equations (cf. Theorem V.3.1 of [8]). □

LEMMA 2.8. *There are constants $c > 0$ and $\alpha \in (0,1)$ such that*

$$\|u_{1x}\|_{H^{\alpha,\alpha/2}(\bar{Q})} \le c.$$

Proof. Let $\zeta(x)$ be a smooth function such that $\zeta(x) = 1$ if $|x| < 1$ and $\zeta(x) = 0$ if $|x| \ge 3/2$. Then the function $w = \zeta u_2$ solves the initial boundary-value problem

$$w_t - D_{22}(u_1,u_2)w_{xx} = F, \quad w|_{|x|=2} = 0, \quad w|_{t=0} = \zeta u_{02}, \tag{2.7}$$

$$F \equiv -D_{22}\zeta_{xx} - D_{22}\zeta_x u_{2x} - \left(\frac{\partial f_2}{\partial u_1}u_{1x} + \frac{\partial f_2}{\partial u_2}u_{2x}\right)\zeta + \frac{\partial D_{22}}{\partial u_2}u_{2x}^2\zeta + \varepsilon h_2\zeta.$$

By the above estimates, there exists a constant c such that

$$\|F\|_{L^4(Q_2)} \le c, \quad Q_2 = (0,T)\times(-2,2).$$

Hence (see, e.g., [8]) u_{2xx} verifies the estimate $\|u_{2xx}\|_{L^4(Q)} \le c$, and function b from (2.5) satisfies the inequality

$$|b(t,x,u_1,p)| \le p^2 + \varphi(t,x), \quad \|\varphi\|_{4,4,Q} \le \mu.$$

Now the estimate of the lemma can be obtained by the same argument as in Lemma 2.7. □

Following [8], we denote by $|v|_Q^{(\alpha)}$ and $|v_0|_\Omega^{(\alpha)}$ the norms of v and v_0 in $H^{\alpha,\alpha/2}(\bar{Q})$ and $H^\alpha(\bar{\Omega})$, respectively.

LEMMA 2.9. *If $u_0 \in H^{2+\beta}(\bar{\Omega})$, $0 < \beta < 1$, then there exists a constant c such that for any solution u from $H^{2+\beta,1+\beta/2}(\bar{Q})$ of (2.1) the estimate $|u|_Q^{(2+\beta)} \le c$ holds, where $c > 0$ depends on T, $|u_0|_\Omega^{(2+\beta)}$, the norms of the functions $f(u)$ and $D(u)$, and their first derivatives in $C(\Delta)$.*

Proof. We know from the above lemmas that there are constants $c > 0$ and $\alpha \in (0,1)$ such that $|u_1, u_{1x}|_Q^{(\alpha)} \le c$. If $\gamma = \min\{\alpha,\beta\}$, it follows from linear equation (2.7) that $|u_2|_Q^{(2+\gamma)} \le c$ (see, e.g., [8]). Let ζ be as in the proof of Lemma 2.8. The function $z = u_1\zeta(x)$ solves the linear problem

$$z_t - D_{11}(u)z_{xx} = G, \quad z|_{|x|=2} = 0, \quad z|_{t=0} = \zeta(x)u_0, \tag{2.8}$$

$$G = D_{11}(u_1\zeta_x)_x - \frac{\partial f_1}{\partial x}\zeta + \frac{\partial D_{11}}{\partial x}u_{1x}\zeta + (D_{12}u_{2x})_x\zeta + \varepsilon h_1\zeta.$$

Hence, by the same argument, $|u_1|_Q^{(2+\gamma)} \le c$. To increase γ up to β, one should return to problem (2.7), which now ensures that $|u_2|_Q^{(2+\beta)} \le c$, and then pass to problem (2.8) to obtain that $|u_1|_Q^{(2+\beta)} \le c$. □

3. Existence and uniqueness. To prove the solvability of problem (4.1), we apply a fixed point argument in the form of the Leray–Schauder principle as in [8]. Let B be a Banach space of x-periodic vector-functions $\mathbf{u}(t,x) \in \mathbb{R}^2$ such that

$$\|\mathbf{u}\|_B := |\mathbf{u}|_Q^{(\beta)} + |\mathbf{u}_x|_Q^{(\beta)} < \infty.$$

Given $\mathbf{a} \equiv (a_1, a_2) \in B$ and $\lambda \in [0,1]$, we define $\mathbf{u} = (u,v)$ to be a solution to the linear problem

$$v_t + \lambda \left[\frac{\partial f_2(\mathbf{a})}{\partial x} - (D_{22}(a_2)v_x)_x - \varepsilon h_2(\mathbf{a}) \right] = (1-\lambda)\nu v_{xx},$$

$$u_t + \lambda \left[\frac{\partial f_1(\mathbf{a})}{\partial x} - (D_{11}(\mathbf{a})u_x)_x - (D_{12}(\mathbf{a})v_x)_x - \varepsilon h_1(\mathbf{a}) \right] = (1-\lambda)\nu u_{xx},$$

$$\mathbf{u}|_{t=0} = (u_{01}(x), u_{02}(x)).$$

Due to uniqueness of the linear Cauchy problem (cf., e.g., [8]), u is an x-periodic vector-function. Thus, the operator $\mathbf{a} \mapsto \mathbf{u} \equiv A_\lambda(\mathbf{a})$ is well-defined, and its fixed points are solutions to problem (4.1) when $\lambda = 1$. By repeating the arguments of the above lemmas, one arrives at the a priori estimates for the fixed points $\mathbf{u}_\lambda$ of the operator A_λ:

$$\mathbf{u}_\lambda \in \Delta, \quad |\mathbf{u}_\lambda, \mathbf{u}_{\lambda x}|_Q^{(\beta)} \le M, \quad |\mathbf{u}_\lambda|_Q^{(2+\beta)} \le M_1, \tag{3.1}$$

where the constants M, M_1 are independent of λ. We restrict A_λ to the set

$$\mathcal{U} = \{\mathbf{u} \in B : \mathbf{u} \in \Delta', \quad |\mathbf{u}_\lambda, \mathbf{u}_{\lambda x}|_Q^{(\beta)} \le M', \quad \mathbf{u}|_{t=0} = \mathbf{u}_0(x)\},$$

where $\operatorname{int} \Delta' \supset \bar{\Delta}$ and $M' > M$. Clearly, $\mathcal{U}$ is a bounded convex set in B, and all the fixed points $\mathbf{u}_\lambda$ of A_λ are strictly inside $\mathcal{U}$.

As in [8], one can prove that the other conditions of the Leray–Schauder theorem are also verified. Namely, we prove that

(i) the set $A_\lambda(\mathcal{U})$ is compact in B for each $\lambda \in [0,1]$;
(ii) the map $\mathbf{a} \mapsto A_\lambda(\mathbf{a})$, $\lambda \in [0,1]$, is continuous on $\mathcal{U}$ uniformly in $(\mathbf{a}, \lambda) \in \mathcal{U} \times [0,1]$;
(iii) the map $\lambda \mapsto A_\lambda(\mathbf{a})$, $\mathbf{a} \in \mathcal{U}$, is continuous on $[0,1]$ uniformly in $\mathbf{a} \in \mathcal{U}$;
(iv) the operator A_0 has a unique fixed point inside $\mathcal{U}$, and the mapping $\mathbf{a} \mapsto \mathbf{a} - A_0(\mathbf{a})$ has an inverse near this fixed point.

Hence, the Cauchy problem (2.1) has at least one x-periodic solution in the Hölder space $H^{2+\beta,1+\beta/2}(\bar{Q})$. Uniqueness can be established in the same manner as in [8]. Thus, we have proved the following.

THEOREM 3.1. *Let functions $f_i(u)$ and $D_{ij}(u)$, their first derivatives, and function $h(u)$ be Hölder continuous with the Hölder exponent $\beta \in (0,1)$. Let hypotheses* (1.2), (1.5), *and* (1.6) *hold, and assume that the x-periodic (with the period 2) initial datum $u_0(x)$ satisfies $u_0 \in H^{2+\beta}(\bar{\Omega})$, $u_0(x) \in \Delta$. Then the Cauchy problem* (2.1) *has a unique x-periodic solution $u(t,x) \in H^{2+\beta,1+\beta/2}(\bar{Q})$.*

Proof of Theorem 1.1. The assertion of Theorem 1.1 about existence follows as a consequence of Theorem 3.1. Indeed, since the estimate of Lemma 2.9 does not depend on ε, there is a sequence $\varepsilon_k \downarrow 0$ such that the corresponding sequence $u_k(t,x)$ of solutions of problem (2.1) converges to a function $u(t,x) \in H^{2+\beta,1+\beta/2}(\bar{Q})$ in the $|\cdot|_Q^{(2+\gamma)}$-norm for any $\gamma < \beta$. Clearly, u solves problem (1.1). Uniqueness can be proved in a straightforward manner as in Theorem 3.1. ∎

4. Basic equations of three-phase flow. We recall here some basic facts about multiphase flow in a porous medium (cf., e.g., [1]). We consider a one-dimensional horizontal flow of three fluid phases in a porous medium. These phases are oil, gas, and water with saturations u_1, u_2, and u_3, respectively. The balance of masses is governed by the mass conservation equations

$$\frac{\partial}{\partial t}(mu_i\rho_i) + \frac{\partial}{\partial x}(\rho_i v_i) = 0, \tag{4.1}$$

where m denotes the porosity of the porous medium, ρ_i is the density, and v_i is the seepage velocity of the ith phase. The functions u_i satisfy the volume-balance equation

$$u_1 + u_2 + u_3 = 1. \tag{4.2}$$

The theory of multiphase flows in porous media is based on the following form of Darcy's law:

$$v_i = -k\lambda_i p_{ix}, \quad \lambda_i = \lambda_i(u_1, u_2),\ i = 1, 2, 3, \tag{4.3}$$

where k stands for the absolute permeability, λ_i is the mobility of the ith phase, and p_i is the pressure of the ith phase. The phase mobilities $\lambda_i(u)$ are known to satisfy (see, e.g., [1])

$$\begin{cases} \lambda_i(u) > 0 \quad \text{if } u_i > 0,\ i \in \{1, 2, 3\}, \\ \lambda_i|_{u_i=0} = 0, \quad i \in \{1, 2, 3\}. \end{cases} \tag{4.4}$$

The capillary pressures are defined as the pressure differences (cf., e.g., [9, 1]), and we assume here that they are functions of the saturations u_1, u_2, that is,

$$P_1(u_1, u_2) = p_1 - p_3, \quad P_2(u_1, u_2) = p_2 - p_3, \tag{4.5}$$

with

$$\frac{\partial P_1}{\partial u_1} \geq 0, \quad \frac{\partial P_2}{\partial u_2} \geq 0. \tag{4.6}$$

Denote

$$\lambda = \sum_1^3 \lambda_i, \quad f_i = \frac{\lambda_i}{\lambda}, \qquad i = 1, 2, 3. \tag{4.7}$$

Assume, for simplicity, $k = m = \rho_i \equiv 1$, $i = 1, 2, 3$. For

$$v = \sum_1^3 v_i,$$

we find from (4.1) and (4.2) that $v_x = 0$, so v depends only on t. Thus, we also assume for simplicity that $v \equiv 1$ as well.

Eliminating the pressure derivative p_{3x}, we have from (4.3)

$$v_1 = f_1(1 + \lambda_2 P_{2x} - (\lambda_2 + \lambda_3)P_{1x}), \quad v_2 = f_2(1 + \lambda_1 P_{1x} - (\lambda_1 + \lambda_3)P_{2x}).$$

When we substitute these velocities into the first two equations in (4.1) we obtain a system like (1.1), where u and f are two-dimensional vectors with components u_1, u_2 and f_1, f_2, respectively, and the 2×2-matrix D is given by

$$\begin{cases} D_{11} = \frac{\lambda_1(\lambda_2+\lambda_3)}{\lambda}\frac{\partial P_1}{\partial u_1} - \frac{\lambda_1\lambda_2}{\lambda}\frac{\partial P_2}{\partial u_1},\ D_{12} = -\frac{\lambda_1\lambda_2}{\lambda}\frac{\partial P_2}{\partial u_2} + \frac{\lambda_1(\lambda_2+\lambda_3)}{\lambda}\frac{\partial P_1}{\partial u_2}, \\ \\ D_{21} = \frac{\lambda_2(\lambda_1+\lambda_3)}{\lambda}\frac{\partial P_2}{\partial u_1} - \frac{\lambda_1\lambda_2}{\lambda}\frac{\partial P_1}{\partial u_1}, \quad D_{22} = -\frac{\lambda_1\lambda_2}{\lambda}\frac{\partial P_1}{\partial u_2} + \frac{\lambda_2(\lambda_1+\lambda_3)}{\lambda}\frac{\partial P_2}{\partial u_2}. \end{cases} \tag{4.8}$$

We first observe that the physical capillarity matrix given by (4.8) satisfies (1.6) and also $D_{11} \geq 0$ and $D_{22} \geq 0$, whenever we have $D_{21} = 0$, and (4.6) holds.

Imposing hypotheses (1.2) on the above capillarity matrix D means, in particular, that we are restricted to the case when the first phase (oil) is not responsible for the amount of diffusion in the equation for the second phase (gas).

On the other hand, the restrictions $D_{21} = 0$ and $D_{22} = D_{22}(u_2)$ also provide a model to define the capillary pressures P_1, P_2 inside Δ. In fact, they are equivalent to the following linear hyperbolic system:

$$A\frac{\partial P_1}{\partial u_1} = \frac{\partial P_2}{\partial u_1}, \quad \frac{\partial P_2}{\partial u_2} = A\frac{\partial P_1}{\partial u_2} + \frac{\lambda D_{22}}{\lambda_2(\lambda_1+\lambda_3)}, \quad A = \frac{\lambda_1}{\lambda_1+\lambda_3}. \tag{4.9}$$

Thus, (4.9) provides a recipe for defining P_1, P_2 inside Δ, when combined with any given model for defining the mobilities inside Δ, such as that in [13].

As a first example, we can easily see that equations (4.9) are valid if

$$\lambda_i = ku_i, \quad P_1 = \alpha\xi, \quad P_2 = \frac{\alpha\xi^2}{2} + \frac{\beta u_2}{k}, \quad \xi \equiv \frac{u_1}{1-u_2}, \tag{4.10}$$

where k, α, β are constants. The corresponding capillarity matrix is

$$\begin{cases} D^*_{11} = k\alpha\xi(1-\xi), \quad D^*_{12} = k\alpha\xi^2(1-\xi) - \beta u_1 u_2, \\ \\ D^*_{21} = 0, \quad D^*_{22} = \beta u_2(1-u_2), \end{cases} \tag{4.11}$$

so the conditions (1.2) are satisfied inside Δ.

Other examples are provided by

$$\lambda_i = k_i u_i, \quad k_1 = k_3, \quad P_1 = \alpha\xi, \quad P_2 = \frac{\alpha\xi^2}{2} + \frac{\beta}{2}\left(\frac{1}{k_3} - \frac{1}{k_2}\right)u_2^2 + \beta\frac{u_2}{k_2} \tag{4.12}$$

and

$$\lambda_i = k_i u_i, \quad k_1 \neq k_3, \quad P_1 = \alpha\xi - \frac{\beta k_0}{2}u_2^2, \quad k_0 = \frac{k_3 - k_1}{k_1 k_3},$$

$$P_2 = \frac{\beta}{2}\left(\frac{1}{k_3} - \frac{1}{k_2}\right)u_2^2 + \frac{\beta}{k_2}u_2 + \frac{\alpha k_1\xi}{k_1 - k_3} - \frac{\alpha k_1 k_3}{(k_1-k_3)^2}\ln\left((k_1-k_3)\xi + k_3\right). \tag{4.13}$$

Observe that the functions P_i given by (4.13) satisfy inequalities (4.6), provided α, β are chosen properly. In both cases, it is not difficult to check that equations (4.9) as well as hypotheses (1.2) are satisfied inside Δ, with D_{21} and D_{22} obtained from (4.11).

Furthermore, a more general class of solutions of (4.9) with mobilities defined as in (4.13) is provided by

$$P_1 = \alpha\xi^m - \beta k_0 \frac{u_2^2}{2}, \quad P_2 = \alpha\varphi(\xi; k_1, k_3, m) + \frac{\beta}{2}u_2^2\left(\frac{1}{k_3} - \frac{1}{k_2}\right) + \frac{\beta}{k_2}u_2, \tag{4.14}$$

where ξ is defined as above and the function φ is defined from the equation

$$\frac{\partial\varphi}{\partial\xi} = \frac{mk_1}{k_1 - k_3}\frac{(k_1 - k_3)\xi^m}{(k_1 - k_3)\xi + k_3},$$

$$m > 1, \quad k_0 = \frac{k_3 - k_1}{k_1 k_3}, \quad \alpha = \text{const}, \quad \beta = \text{const}.$$

The derivation of formulas (4.10), (4.12), (4.13), and (4.14) is given at the end of this section.

If $u \in \Delta$, the variable ξ takes values in the interval $[0, 1]$, and ξ is a smooth function of u everywhere in Δ except the point $(u_1, u_2) = (0, 1)$. From the above examples and the general formulas (4.8) we see that, in the model for three-phase capillary flows, equations (1.1) will, in general, constitute a degenerate parabolic system with bounded coefficients. The set of degeneration coincides with the boundary of Δ.

The well-posedness theory developed in sections 2 and 3, nevertheless, can be applied to a suitable regularization of these physical systems. For instance, in all the above examples we may easily regularize the system by replacing matrix $D(u)$ with a matrix $D^\nu(u)$ defined as

$$D^\nu(u) = \rho^\nu(u_2)D(u) + \nu I, \tag{4.15}$$

where $\nu > 0$ is as small as we wish, I represents the identity matrix, and ρ^ν is a nonnegative smooth function satisfying

$$\begin{cases} \rho^\nu(u_2) = 0 & \text{if } |u_2 - 1| < \nu/2, \\ \rho^\nu(u_2) = 1 & \text{if } |u_2 - 1| > \nu. \end{cases}$$

We can easily verify that $D^\nu(u)$ satisfies the hypotheses of Theorem 1.1.

To close this section we give the details of the derivation of formulas (4.10), (4.12), (4.13), and (4.14). We consider the linear system (4.9) and describe a class of solutions which includes (4.10), (4.12), (4.13), and (4.14) when

$$\lambda_i = k_i u_i, \quad D_{22} = \beta u_2(1 - u_2).$$

First, we explain how the variable $\xi = \frac{u_1}{1-u_2}$ appears. Eliminating function P_2 from (4.9) by cross-differentiation, we get

$$\frac{u_1}{1 - u_2}\frac{\partial P_1}{\partial u_1} - \frac{\partial P_1}{\partial u_2} = \frac{D_{22}(u_2)k_0}{1 - u_2}. \tag{4.16}$$

Clearly, this equation has a particular solution $P_1^*(u_2)$ depending only on the variable u_2. On the other hand, one can verify that any smooth function $\Phi(\xi)$ solves the homogeneous equation (4.16). Thus, the function $\Phi(\xi) + P_1^*(u_2)$ is a solution of (4.16).

Now we look for a solution of (4.9) having the form

$$P_i = q_i(\xi) + p_i(u_2). \tag{4.17}$$

Since $u_3 = 1 - u_1 - u_2$, we calculate

$$\frac{\partial \xi}{\partial u_1} = \frac{1}{1-u_2}, \quad \frac{\partial \xi}{\partial u_2} = \frac{\xi}{1-u_2}, \quad A = \frac{k_1 \xi}{(k_1 - k_3)\xi + k_3},$$

$$\frac{\partial P_1}{\partial u_1} = \frac{q_1'}{1-u_2}, \quad \frac{\partial P_1}{\partial u_2} = \frac{\xi q_1'}{1-u_2} + p_1',$$

$$\frac{\partial P_2}{\partial u_1} = \frac{q_2'}{1-u_2}, \quad \frac{\partial P_2}{\partial u_2} = \frac{\xi q_2'}{1-u_2} + p_2',$$

$$\frac{\lambda D_{22}}{\lambda_2(\lambda_1 + \lambda_3)} = \frac{\beta(1-u_2)}{k_2} + \frac{\beta u_2}{(k_1 - k_3)\xi + k_3}.$$

With these formulas at hand, system (4.9) reads as

$$A(\xi) q_1'(\xi) = q_2'(\xi), \quad p_2'(u_2) = \frac{k_1 \xi p_1'(u_2) + \beta u_2}{(k_1 - k_3)\xi + k_3} + \frac{\beta(1-u_2)}{k_2}. \tag{4.18}$$

If one chooses

$$p_1'(u_2) = -k_0 \beta u_2, \tag{4.19}$$

$(4.18)_2$ reads as

$$p_2'(u_2) = \beta u_2 \left(\frac{1}{k_3} - \frac{1}{k_2} \right) + \frac{\beta}{k_2}. \tag{4.20}$$

Let us take

$$q_1(\xi) = \alpha \xi^m, \quad m \geq 1.$$

Then

$$q_2'(\xi) = \alpha m \xi^{m-1} A(\xi). \tag{4.21}$$

Now, to obtain a class of solutions of type (4.17), we just integrate equations (4.19), (4.20), and (4.21).

REFERENCES

[1] M.B. Allen, J.B. Behie, and J.A. Trangenstein, *Multiphase Flows in Porous Media: Mechanics, Mathematics and Numerics*, Lecture Notes in Engineering 34, Springer-Verlag, New York, Heidelberg, Berlin, 1988.

[2] H. Amann, *Dynamic theory of quasi-linear parabolic systems* III. *Global existence*, Math. Z., 202 (1989), pp. 219–250.

[3] J.B. Bell, J.A. Trangenstein, and G.R. Shubin, *Conservation laws of mixed type describing three-phase flows in porous media*, SIAM J. Appl. Math, 46 (1986), pp. 1000–1017.

[4] Z. Chen and R.E. Ewing, *Comparison of various formulations of three-phase flow in porous media*, J. Comput. Phys., 132 (1997), pp. 362–373.
[5] K. Chueh, C. Conley, and J. Smoller, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 373–392.
[6] H. Frid, *Existence and asymptotic behavior of measure-valued solutions for three-phase flow in porous media*, J. Math. Anal. Appl., 196 (1995), pp. 614–627.
[7] H. Frid and V. Shelukhin, *A quasilinear parabolic system for three-phase capillary flow in porous media* II: *Dirichlet boundary conditions*, in preparation.
[8] O.A. Ladyženskaja, V.A. Solonnikov, and N.N. Ural'ceva, *Linear and Quasi-linear Equations of Parabolic Type*, AMS, Providence, RI, 1968.
[9] D.W. Peaceman, *Fundamentals of Numerical Reservoir Simulation*, Elsevier, Amsterdam, Oxford, New York, 1977.
[10] V.I. Pen'kovskii, *Capillary pressure, gravitational and dynamical phases distribution in the system water-oil-gas-rocks*, Appl. Mech. Tech. Phys., 37, 6 (1996), pp. 85–90.
[11] D. Serre, *Systèmes de Lois de Conservation* II, Diderot Editeur, Arts et Sciences, Paris, New York, Amsterdam, 1996.
[12] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, Heidelberg, Berlin, 1983.
[13] H.L. Stone, *Probability model for estimating three-phase relative permeability*, J. Petroleum Techn., (February 1970), pp. 214–218.
[14] A.N. Varchenko and A.F. Zazovskii, *Three-phase filtration of immiscible fluids*, Itogi Nauki i Tekhniki, Seriya Koleksnie i spetsial'nie Razdely Mekhaniki, 4 (1991), pp. 98–154 (in Russian).

SIAM J. MATH. ANAL.
Vol. 35, No. 4, pp. 1042–1080

WEAK INTERACTION BETWEEN SOLITARY WAVES OF THE GENERALIZED KdV EQUATIONS*

TETSU MIZUMACHI†

Abstract. In this paper we study the large time behavior of two decoupled solitary waves of the generalized KdV equations $u_t + (u_{xx} + f(u))_x = 0$, where $f(u) = |u|^{p-1}u/p$ $(3 \le p < 5)$. We prove that if the speeds of the solitary waves are sufficiently close at the initial time, the leading wave becomes larger and the trailing wave becomes smaller, and the distance between two solitary waves becomes larger as $t \to \infty$.

Key words. generalized Korteweg–de Vries equations, interaction of solitary waves, asymptotic behavior

AMS subject classifications. 35Q53, 35B40, 35B35

DOI. 10.1137/S003614100240871X

1. Introduction.

In the present paper, we study the large time behavior of solutions of the generalized KdV equation (GKdV)

$$\begin{cases} u_t + f(u)_x + u_{xxx} = 0 \quad \text{for } x \in \mathbb{R},\ t > 0, \\ u(x,0) = u_0(x) \quad \text{for } x \in \mathbb{R}, \end{cases} \tag{1.1}$$

where $f(u) = |u|^{p-1}u/p$. The equation was derived by Korteweg and de Vries in [26] as a model for long waves propagating in a canal in the case where $f(u) = u^2$. The Cauchy problem of (1.1) has been studied by many authors. See [7, 8, 10, 16, 21, 22, 23] and the references therein.

Let φ_c be a positive solution of

$$\begin{cases} \varphi'' - c\varphi + f(\varphi) = 0 \quad \text{for } y \in \mathbb{R}, \\ \lim_{y \to \pm\infty} \varphi(y) = 0. \end{cases} \tag{1.2}$$

Equation (1.1) has solitary wave solutions with finite energy, which are written as

$$u(x,t) = \varphi_c(x - ct - \gamma),$$

where c is a positive number and γ is a real number. The positive solution of (1.2) with its maximum at $y = 0$ satisfies

$$\begin{aligned} \varphi_c(y) &= 2^{-\frac{2}{p-1}} \alpha(c) \left(\operatorname{sech}\left((p-1)\sqrt{c}y/2\right)\right)^{\frac{2}{p-1}} \\ &= \alpha(c) e^{-\sqrt{c}|y|}(1 + O(e^{-(p-1)\sqrt{c}|y|})), \end{aligned} \tag{1.3}$$

where $\alpha(c) = (2p(p+1)c)^{1/(p-1)}$.

The orbital stability of the solitary wave solutions has been studied by Benjamin [2]; Bona [3]; Bona, Souganidis, and Strauss [5]; Grillakis, Shatah, and Strauss [17]; and Weinstein [44] (see also Bona and Soyeur [6]). They proved that the solitary wave

*Received by the editors June 2, 2002; accepted for publication (in revised form) February 28, 2003; published electronically November 4, 2003. This research was supported by Grant-in-Aid for Scientific Research 12740095.

http://www.siam.org/journals/sima/35-4/40871.html

†Department of Mathematical Sciences, Faculty of Science, Yokohama City University, 22-2 Seto, 236-0027, Japan (mizumati@yokohama-cu.ac.jp).

solutions are stable in $H^1(\mathbb{R})$ if $1 < p < 5$ and unstable if $p > 5$. That is to say, if the solution u of (1.1) is initially close to a solitary wave $\varphi_c(x)$, it remains close to the set $\{\varphi_c(x - ct + \gamma) \,|\, \gamma \in \mathbb{R}\}$ for all the time when $1 < p < 5$. Recently Martel and Merle [29, 31] have proved that solitary wave solutions are unstable if $p = 5$.

On the other hand, Pego and Weinstein [36] proved the asymptotic stability of solitary waves in an exponentially weighted space. Mizumachi [33] extended the result to the algebraically weighted space, which shows that the solitary wave is asymptotically stable under the presence of small solitary waves which lie far behind the main wave. Moreover, [33] proved that if $3 < p < 5$ and the solution is initially close to a solitary wave in some weighted space, it decouples into a sum of the solitary wave with slightly displaced parameters and a small dispersive wave which is asymptotically free in $L^2(\mathbb{R})$.

If $f(u) = u^2$ or $f(u) = \pm u^3$, inverse scattering theory is available. Inverse scattering theory informs us that the solution of (1.1) with well-localized initial data resolves into a train of solitary waves moving to the right, and dispersive radiation, which moves to the left (see [1, 12, 38] and the references therein). In the integrable case, Maddocks and Sachs [27] proved the stability of N-soliton solutions in $H^N(\mathbb{R})$. Although inverse scattering theory does not apply to (1.1) with more general p, this type of asymptotic resolution appears to extend to equations with more general nonlinearities.

Pereleman [34] studied the large time asymptotics of 2-pulse solutions of nonlinear Schrödinger equations in the case where the two pulses are well separated and move in opposite directions with large relative velocities. In this case, the interaction of solitary waves is rather weak. Recently Martel, Merle, and Tsai [32] proved the asymptotic stability of multipulse solutions of (1.1) in H^1, based on energy arguments. They deal with the case where the solitary waves are well separated and the larger solitary wave goes ahead of the smaller one.

On the other hand, Ei, Fujii, and Kunihiro [14] formally analyzed the large time behavior of multisoliton solutions, in the case where $p = 2$ and the solitons are well separated and of almost the same speed. In that case, the interaction of solitary waves is stronger and plays an important role. In fact, it makes solitary waves repulsive. In the present paper, we rigorously show that, for $3 \le p < 5$, the motion of solitary waves is described by a 4-dimensional system of ordinary differential equations which almost conserves energy, and that the solitary waves are repulsive if they are well separated, of almost the same speed, and of the same sign.

Before we state our result more precisely, we introduce several notations. We use the notation $\|\cdot\|_p$ for the $L^p(\mathbb{R})$-norm and $\|\cdot\| = \|\cdot\|_2$ and $\|\cdot\|_{m,s}$ for the norms defined by $\|v\|_{m,s} = \|\langle x\rangle^s (1-\partial^2)^{m/2} v\|_2$. Let $H^1_a(\mathbb{R}) = \{v \in H^1_{loc}(\mathbb{R}) \,|\, e^{ax} v \in H^1(\mathbb{R})\}$ be equipped with the norm $\|v\|_{H^1_a} = \|e^{ax} v\|_{1,0}$.

Now we introduce our main results. Since (1.1) has a two-parameter family of solitary wave solutions $\{\varphi_c(\cdot + h) \,|\, c > 0, h \in \mathbb{R}\}$, the linearized operator around the solitary wave has an eigenvalue 0 with multiplicity two. In the following theorem, we assume that the linearized operator has no other eigenvalues (see (2.6) in section 2), which holds generically for $p \in (1, 5)$.

THEOREM 1.1. *Assume that* $3 \le p < 5$ *and that* (2.6) *holds. Let* I *be a compact subset of* $(0, \infty)$ *and let* $c_{1,0}$, $c_{2,0} \in I$. *Let* $d_0 = \min_{i=1,2} \sqrt{c_{i,0}}$, $0 < a_1 < a < a_2 \le d_0/100$, *and let*

$$u_0(x) = \varphi_{c_{1,0}}(x - x_{1,0}) + \varphi_{c_{2,0}}(x - x_{2,0}) + v_0(x - x_{1,0}),$$

$$\|v_0\|_{1,1} + \|v_0\|_{H^1_{a_1}} + \|v_0\|_{H^1_{a_2}} = \varepsilon_0,$$
$$|c_{2,0} - c_{1,0}| + e^{-\frac{d_0}{2}(x_{2,0}-x_{1,0})} = \varepsilon_1,$$
$$c_{2,0} - c_{1,0} \geq -\varepsilon_1/2.$$

Then there exist positive numbers C and $\bar{\varepsilon}_1$ depending on a_1 and I such that if $\varepsilon_1 \in (0, \bar{\varepsilon}_1)$ and $\varepsilon_0 \leq C\varepsilon_1^2$, there exist $c_{2,\infty} > c_{1,\infty} > 0$, $x_{i,\infty} \in \mathbb{R}$ $(i = 1, 2)$, satisfying $c_{2,\infty} - c_{1,\infty} = O(\varepsilon_1)$ and

(1.4)
$$\left\| e^{a(x-c_{1,\infty}t-x_{1,\infty})} \left(u(t,\cdot) - \sum_{i=1,2} \varphi_{c_{i,\infty}}(\cdot - c_{i,\infty}t - x_{i,\infty}) \right) \right\|_{1,0} = O(e^{-C_1\varepsilon_1^{1-\eta}t}),$$

(1.5)
$$\left\| u(t,\cdot) - \sum_{i=1,2} \varphi_{c_{i,\infty}}(\cdot - c_{i,\infty}t - x_{i,\infty}) \right\|_{\infty} = O(t^{-\frac{1}{3}})$$

as $t \to \infty$, where C_1 and η are positive constants with $\eta = O(\varepsilon_0)$. Furthermore if $3 < p < 5$, there exist $v_\infty \in L^2(\mathbb{R})$ and a positive number $\delta(\varepsilon_0)$ satisfying $\lim_{\varepsilon_0 \to 0} \delta(\varepsilon_0) = 0$ and

(1.6)
$$\left\| u(t,\cdot) - \sum_{i=1,2} \varphi_{c_{i,\infty}}(\cdot - c_{i,\infty}t - x_{i,\infty}) \right\|_{1,0} \leq \delta(\varepsilon_0),$$

(1.7)
$$\left\| u(t,\cdot) - \sum_{i=1,2} \varphi_{c_{i,\infty}}(\cdot - c_{i,\infty}t - x_{i,\infty}) - e^{-t\partial_x^3} v_\infty \right\| = o(1)$$

as $t \to \infty$.

Remark 1.1. Let us decompose the solution of (1.1) into solitary wave parts and a dispersive wave part: $u(x,t) = \varphi_{c_1(t)}(x - x_1(t)) + \varphi_{c_2}(x - x_2(t)) + v$. Then we have

$$\sum_{i=1,2} (|c_i(0) - c_{i,0}| + |x_i(0) - x_{i,0}|) = O(\varepsilon_0)$$

(see Lemma 3.1 for details). The repulsive force between solitary waves is approximately given by

$$f(\textstyle\sum_{i=1,2} \varphi_{c_i(t)}(x - x_i(t))) - \sum_{i=1,2} f(\varphi_{c_i(t)}(x - x_i(t))) \sim e^{-d(t)(x_2(t)-x_1(t))},$$

where $d(t) = \min_{i=1,2} \sqrt{c_i(t)}$. We assume $c_{2,0} - c_{1,0} \geq -\varepsilon_1/2$ and assume the smallness of $\varepsilon_0\varepsilon_1^{-1}$ so that the repulsive forces between the solitary waves is sufficiently strong at the initial time compared with relative velocity. In other words, we assume that $x_{2,0} - x_{1,0}$ is sufficiently but not arbitrary large in the case where $c_{1,0} > c_{2,0}$. Our result does not cover the case where $c_{1,0} > c_{2,0}$ and $x_{2,0} - x_{1,0}$ is extremely large.

To prove the result, we make use of an exponentially weighted space as in [36]. Since the linearized operator around the multipulse satisfies the spectral gap condition in the exponentially weighted space, we apply a dynamical systems point of view as in [13]. However, we cannot directly apply local-manifold theory to our problem. The difficult point of the problem is to show that the dispersive part of the solution

remains small in H^1. Since the orbital stability of multipulse solutions of (1.1) remains unknown if the equation is nonintegrable and the relative velocity between solitary waves is small, we use the scattering results to obtain an H^1-estimate. This idea was first used by [9, 41, 42] to show the asymptotic stability of solitary wave solutions to nonlinear Schrödinger equations. We show that the interaction of the dispersive part and the solitary waves becomes small as $t \to \infty$, and we use the scattering result (1.1) due to Hayashi and Naumkin [19, 20], which gave the large time asymptotics of $H^{1,1}$-small solutions of (1.1) with $p \geq 3$, to obtain H^1-estimate of the dispersive part of the solution. For the other results on nonlinear scattering of solutions to (1.1), see [11, 22, 24, 25, 37, 39, 40] and the references therein.

A similar result is expected to hold for the KdV equation ($p = 2$). However, we cannot prove that result, because the dispersive part of the solutions of (1.1) is not asymptotically free if $p \leq 3$ and it cannot be estimated in the same way if $p < 3$.

Our plan of the present paper is as follows. In section 2, we investigate the spectrum of the linearized operator and obtain decay properties of the linearized equation. In section 3, we decompose the solution into solitary wave parts and a dispersive wave part and obtain the system of ordinary differential equations of phases and speeds of solitary waves. In section 4, we obtain a priori estimates of the system obtained in section 3 and prove Theorem 1.1.

Finally, let us introduce some notation, which shall be used later. For an operator A, we denote by $\sigma(A)$ the spectrum of A and by $\rho(A)$ the resolvent set of A. For any Banach spaces X, Y, we denote by $\mathcal{L}(X, Y)$ the space of bounded linear operators from X to Y. We abbreviate $\mathcal{L}(X, X)$ as $\mathcal{L}(X)$.

We define D^α as

$$\begin{aligned} D^\alpha f &\equiv \mathcal{F}^{-1} \xi^\alpha e^{-i\pi \frac{1+\alpha}{2}} \mathcal{F} f \\ &= \frac{2\pi}{\Gamma(1-\alpha)} \int_0^\infty (f(x+y) - f(x)) \frac{dy}{y^{\alpha+1}} \end{aligned}$$

for $\alpha \in (0, 1)$, where

$$\mathcal{F} f(\xi) = (2\pi)^{-\frac{1}{2}} \int e^{-ix\xi} f(x) dx \quad \text{and} \quad \mathcal{F}^{-1} g(x) = (2\pi)^{-\frac{1}{2}} \int e^{ix\xi} g(\xi) d\xi.$$

Let $\langle f, g \rangle = \int f(x) \overline{g(x)} dx$ and $\langle t \rangle = \sqrt{1+t^2}$. Various constants will be simply denoted by C and C_i ($i \in \mathbb{N}$) in the course of the calculations.

2. Spectral analysis of the linearized equation.

2.1. Spectral properties of the linearized equation around solitary wave solutions. Let $L = -\partial_y^2 + c - f'(\varphi_c)$. The essential spectrum of the operator $\partial_y L$ in $L^2(\mathbb{R})$ consists of $i\mathbb{R}$, and it holds that $\partial_y L \partial_y \varphi_c = 0$, $\partial_y L \partial_c \varphi_c = -\partial_y \varphi_c$. So $\lambda = 0$ is always an eigenvalue of $\partial_y L$ imbedded in the essential spectrum. Set

$$\tilde{\xi}_1(y, c) = \partial_y \varphi_c(y), \quad \tilde{\xi}_2(y, c) = \partial_c \varphi_c(y), \tag{2.1}$$

$$\tilde{\eta}_1(y, c) = \theta_1(c) \int_{-\infty}^{y} \partial_c \varphi_c + \theta_2(c) \varphi_c(y), \quad \tilde{\eta}_2(y, c) = \theta_3(c) \varphi_c(y), \tag{2.2}$$

where

$$\theta_3(c) = -\theta_1(c) = 2 \left(\frac{d}{dc} \|\varphi_c\|^2 \right)^{-1}, \quad \theta_2(c) = 2 \left(\frac{d}{dc} \int_{\mathbb{R}} \varphi_c \right)^2 \left(\frac{d}{dc} \|\varphi_c\|^2 \right)^{-2}.$$

The functions $\tilde{\xi}_1(y,c)$, $\tilde{\xi}_2(y,c)$, and $\tilde{\eta}_2(y,c)$ decay exponentially as $|y| \to \infty$. The function $\tilde{\eta}_1(y)$ also decays exponentially as $y \to -\infty$ but is merely bounded as $y \to \infty$. Let

$$Au = e^{ay}\partial_y L e^{-ay}, \quad \xi_i(y,c) = e^{ay}\tilde{\xi}_i(y,c), \quad \eta_i(y,c) = e^{-ay}\tilde{\eta}_i(y,c) \quad \text{for } i = 1,2.$$

The spectrum of $\partial_y L$ in $L^2_a = \{v \,|\, e^{ay}v \in L^2\}$ is equivalent to the spectrum of A in L^2. Since $L\partial_y\varphi_c = 0$ and $L\partial_c\varphi_c = -\varphi_c$,

$$A\xi_1(y,c) = 0, \quad A\xi_2(y,c) = -\xi_1(y,c), \tag{2.3}$$

$$A^*\eta_1(y,c) = -\eta_2(y,c), \quad A^*\eta_2(y,c) = 0. \tag{2.4}$$

By (2.1), (2.2), and the definitions of $\xi_i(\cdot,c)$ and $\eta_j(\cdot,c)$ $(i,j = 1,2)$,

$$\langle \xi_i(\cdot,\, c),\, \eta_j(\cdot,\, c)\rangle = \delta_{ij} \tag{2.5}$$

for $i,j = 1,2$. The essential spectrum of A consists of

$$S(a) := \big\{-(i\tau - a)^3 + c(i\tau - a) \,\big|\, \tau \in \mathbb{R}\big\}.$$

The complement of $S(a)$ in $\mathbb{C}$ consists of two disjoint open components. We denote by $\Omega(a)$ the connected component that includes the right-hand side of $S(a)$. Let us recall some results due to Pego and Weinstein.

PROPOSITION 2.1 (see [35, 36]).

1. *Assume $1 < p < 5$ and $0 < a < \sqrt{c/3}$. Then $\lambda = 0$ is an eigenvalue for A with algebraic multiplicity two. In addition, if $\partial_y L$ has no eigenvalue in L^2 other than 0, there exists $0 < b < a(c - a^2)$ such that*

$$\sigma(A) \subset \{0\} \cup \{\lambda \in \mathbb{C} \,|\, \mathrm{Re}\lambda < -b\}.$$

2. *The set of values of p with $1 < p \le 5$ such that the operator $\partial_y L$ has some nonzero eigenvalues is a finite set which does not include $p = 2, 3$.*

Throughout the paper we assume the following:

$$\text{The operator } \partial_y L \text{ has no nonzero eigenvalue in } L^2(\mathbb{R}). \tag{2.6}$$

Thus

$$P_c u = \sum_{i=1}^{2} \langle u,\, \eta_i(\cdot,c)\rangle \xi_i(\cdot,c) \text{ and } Q_c u = (1 - P_c)u \tag{2.7}$$

are spectral projections associated with A.

2.2. Properties of the linearized operator around 2-pulse solutions. Let $\tau_h u = u(\cdot - h)$ and let

$$\begin{aligned}
&A_0(c)u = -(\partial_y - a)^3 u + c(\partial_y - a)u, && R_0(\lambda) = (\lambda - A_0)^{-1},\\
&V_1 u = -(\partial_y - a)(f'(\varphi_{c_1})u), && V_2 u = -(\partial_y - a)(f'(\tau_h\varphi_{c_2})u),\\
&A_1 = A_0(c_1) + V_1(c_1), && A_2 = A_0(c_2) + V_2(c_2),\\
&P_1(c) = P_c, \quad P_2(c) = \tau_h P_c \tau_{-h}, && Q_i(c) = 1 - P_i(c) \quad \text{for } i = 1,2.
\end{aligned}$$

Let χ be a smooth nonnegative function satisfying $0 \le \chi \le 1$, with

$$\chi(y) = \begin{cases} 0 \text{ if } y \ge 2/3, \\ 1 \text{ if } y \le 1/3 \end{cases}$$

and $\chi_1(y,h) = \chi(y/h)$, $\chi_2(y,h) = 1 - \chi_1(y,h)$. Let $\mathbf{c} = (c_1, c_2)$ and let c_1, c_2, and h be positive numbers. We define the operator $\mathcal{A}_{\mathbf{c},h}$ by

$$\mathcal{A}_{\mathbf{c},h}(a)u = -(\partial_y - a)^3 u + \sum_{i=1,2} c_i(\partial_y - a)(\chi_i u) + \sum_{i=1,2} V_i u.$$

For simplicity, we abbreviate $\mathcal{A}_{\mathbf{c},h}(a)$ as $\mathcal{A}_{\mathbf{c},h}$ if no confusion arises. We investigate the spectral properties of the operator $\mathcal{A}_{\mathbf{c},h}$ for large h.

LEMMA 2.2. *Assume* (2.6). *Let I be a compact subset of $(0,\infty)$ and let c_1, $c_2 \in I$. Let ε_* be a sufficiently small number and let h_* be a sufficiently large number. Then there exist positive numbers b' and ρ such that if $\varepsilon := |c_2 - c_1| \le \varepsilon_*$ and $h \ge h_*$,*

$$\left\{ \lambda \in \mathbb{C} \,|\, \mathrm{Re}\lambda > -b', |\lambda| \ge \frac{\rho}{2} \right\} \subset \rho(\mathcal{A}_{\mathbf{c},h}).$$

The number ρ can be chosen as a function of ε and h satisfying $\lim_{\varepsilon \to 0, h \to \infty} \rho = 0$. Furthermore, the operator $\mathcal{P}_{\mathbf{c},h}$ defined by

$$\mathcal{P}_{\mathbf{c},h} = \frac{1}{2\pi i} \int_{|\lambda| = \rho} (\lambda - \mathcal{A}_{\mathbf{c},h})^{-1} d\lambda$$

is the spectral projection associated with $\mathcal{A}_{\mathbf{c},h}$, and the range of $\mathcal{P}_{\mathbf{c},h}$ is 4-dimensional.

Proof. First, we remark that there exists $b' > 0$ such that

$$\sup_{\mathrm{Re}\lambda > -b'} \|Q_i(c_i)(\lambda - A_i)^{-1}\|_{\mathcal{L}(H^{-3}(\mathbb{R}), L^2(\mathbb{R}))} < \infty$$

for $i = 1, 2$, that

$$(\lambda - A_1)^{-1} P_1(c_1) u = \sum_{j=1}^{2} \frac{1}{\lambda} \langle u, \eta_j(\cdot, c_1) \rangle \xi_j(\cdot, c_1) - \frac{1}{\lambda^2} \langle u, \eta_2(\cdot, c_1) \rangle \xi_1(\cdot, c_1),$$

$$(\lambda - A_2)^{-1} P_2(c_2) u = \sum_{j=1}^{2} \frac{1}{\lambda} \langle u, \tau_h \eta_j(\cdot, c_2) \rangle \tau_h \xi_j(\cdot, c_2) - \frac{1}{\lambda^2} \langle u, \tau_h \eta_2(\cdot, c_2) \rangle \tau_h \xi_1(\cdot, c_2),$$

and that

$$\sup_{y \in \mathbb{R}} \|[A_i, \chi_i]\|_{\mathcal{L}(L^2, H^{-2})} = O(1/h). \tag{2.8}$$

Since $[A_i, \chi_i] = 0$ for $y \le h/3$ and $y \ge 2h/3$, it follows from (1.3) and (2.7) that

$$\|P_i[A_i, \chi_i]\|_{\mathcal{L}(L^2)} = O\left(e^{-\frac{ah}{3}}\right).$$

Combining the above, we have

$$\begin{aligned}
&\left((\lambda-A_1)^{-1}\chi_1+(\lambda-A_2)^{-1}\chi_2\right)(\lambda-\mathcal{A}_{\mathbf{c},h})\\
&=1+\sum_{i=1,2}(\lambda-A_i)^{-1}([A_i,\chi_i]-\chi_iV_{3-i})\\
&\quad+(c_1-c_2)(\lambda-A_1)^{-1}\chi_1(\partial_y-a)(\chi_2u)+(c_2-c_1)(\lambda-A_2)^{-1}\chi_2(\partial_y-a)(\chi_1u)\\
&=1+\sum_{i=1}^{2}(\lambda-A_i)^{-1}P_i[A_i,\chi_i]+\sum_{i=1}^{2}(\lambda-A_i)^{-1}Q_i[A_i,\chi_i]\\
&\quad+(\lambda-A_1)^{-1}(P_1+Q_1)\left(\chi_1V_2+(c_1-c_2)\chi_1(\partial_y-a)(\chi_2\cdot)\right)\\
&\quad+(\lambda-A_2)^{-1}(P_2+Q_2)\left(\chi_2V_1+(c_2-c_1)\chi_2(\partial_y-a)(\chi_1\cdot)\right)\\
&=:1+r(\lambda)
\end{aligned}$$

and

$$\|r(\lambda)\|_{\mathcal{L}(L^2)}=O\left(|c_2-c_1|+\frac{1}{h}+(|c_2-c_1|+e^{-\frac{ah}{3}})\left(\frac{1}{|\lambda|}+\frac{1}{|\lambda|^2}\right)\right)$$

for $\lambda\in\{\lambda\in\mathbb{C}\setminus\{0\}\,|\,\mathrm{Re}\lambda>-b'\}$. Therefore, for any $\rho>0$ there exist $\varepsilon,h_*>0$ such that if $|c_2-c_1|\le\varepsilon_*$, $h\ge h_*$, $|\lambda|\ge\rho/2$, and $\mathrm{Re}\lambda>-b'$, the operator $1+r(\lambda)$ is invertible and $\lambda\in\rho(\mathcal{A}_{\mathbf{c},h})$.

Next, we show that the range of the operator $\mathcal{P}_{\mathbf{c},h}$ is 4-dimensional for large h. By a simple computation,

$$\begin{aligned}
&\|\mathcal{P}_{\mathbf{c},h}-P_1(c_1)\chi_1-P_2(c_2)\chi_2\|_{\mathcal{L}(L^2)}\\
&\le\frac{1}{2\pi}\left\|\oint_{|\lambda|=\frac{b'}{2}}\left((\lambda-\mathcal{A}_{\mathbf{c},h})^{-1}-(\lambda-A_1)^{-1}\chi_1-(\lambda-A_2)^{-1}\chi_2\right)d\lambda\right\|_{\mathcal{L}(L^2)}\\
&\le\frac{1}{2\pi}\left\|\oint_{|\lambda|=\frac{b'}{2}}r(\lambda)(1+r(\lambda))^{-1}\left((\lambda-A_1)^{-1}\chi_1+(\lambda-A_2)^{-1}\chi_2\right)d\lambda\right\|_{\mathcal{L}(L^2)}\\
&=O\left(\|r(\lambda)\|_{\mathcal{L}(L^2)}\right).
\end{aligned}\tag{2.9}$$

Let

$$\xi_i(y,c,h,a)=\begin{cases}e^{ay}\tilde{\xi}_i(y,c)\text{ for }i=1,2,\\ e^{ay}\tilde{\xi}_{i-2}(y-h,c)\text{ for }i=3,4,\end{cases}$$

$$\eta_i(y,c,h,a)=\begin{cases}e^{-ay}\tilde{\eta}_i(y,c)\text{ for }i=1,2,\\ e^{-ay}\tilde{\eta}_{i-2}(y-h,c)\text{ for }i=3,4.\end{cases}$$

We abbreviate $\xi_i(y,c,h,a)$ and $\eta_i(y,c,h,a)$ as $\xi_i(y,c,h)$ and $\eta_i(y,c,h)$, respectively, if there is no confusion. Let c_{ij} be real numbers such that

$$\hat{P}u=\sum_{1\le i,j\le4}c_{ij}\langle u,\chi_{[\frac{i+1}{2}]}\eta_i(\cdot,c_{[\frac{i+1}{2}]},h)\rangle\xi_j(\cdot,c_{[\frac{j+1}{2}]},h)$$

is a projection onto the linear subspace

$$\mathrm{span}\left\{\xi_i(\cdot,c_{[\frac{i+1}{2}]},h)\;\middle|\;1\le i\le4\right\}.$$

Then it follows that $c_{ij} = \delta_{ij} + O(e^{-dh/3})$, where $d = \min_{i=1,2} \sqrt{c_i}$. Hence we have

$$\begin{aligned}
&\|\mathcal{P}_{\mathbf{c},h} - \hat{P}\|_{\mathcal{L}(L^2)} \\
&\le \|\mathcal{P}_{\mathbf{c},h} - P_1(c_1)\chi_1 - P_2(c_2)\chi_2\|_{\mathcal{L}(L^2)} + \|P_1(c_1)\chi_1 + P_2(c_2)\chi_2 - \hat{P}\|_{\mathcal{L}(L^2)} \\
&= O(|c_1 - c_2| + 1/h).
\end{aligned}$$

Thus the range of $\hat{P}$ and the range of $\mathcal{P}_{\mathbf{c},h}$ are isomorphic, which yields that the $\mathcal{P}_{\mathbf{c},h}$ is 4-dimensional. ∎

Next, we investigate (generalized) eigenfunctions of $\mathcal{A}_{\mathbf{c},h}$ around $\lambda = 0$. Let $\bar{\eta}_j(\cdot, \mathbf{c}, h) = \mathcal{P}^*_{\mathbf{c},h}\eta_j(\cdot, c_{[\frac{j+1}{2}]}, h)$ for $1 \le j \le 4$.

LEMMA 2.3. *Let I be a compact subset of $(0, \infty)$. Then there exist positive constants ε_*, h_*, and C such that for any $h \ge h_*$ and c_1, $c_2 \in I$ with $|c_1 - c_2| \le \varepsilon_*$, it holds that*

$$\begin{aligned}
&\|\bar{\eta}_j(\cdot, \mathbf{c}, h) - \eta_j(\cdot, c_1, h)\|_{H^1(\mathbb{R})} \le Ce^{-\frac{d+a}{3}h} \quad \textit{for } j = 1,\ 2, \\
&\|\bar{\eta}_j(\cdot, \mathbf{c}, h) - \eta_j(\cdot, c_2, h)\|_{H^1(\mathbb{R})} \le Ce^{-\frac{d+2a}{3}h} \quad \textit{for } j = 3,\ 4, \\
&\|\bar{\eta}_j\|_{H^3(\mathbb{R})} + \left\|\frac{\partial \bar{\eta}_j}{\partial c_i}\right\|_{H^3(\mathbb{R})} + \left\|\frac{\partial \bar{\eta}_j}{\partial h}\right\|_{H^3(\mathbb{R})} \le C \quad \textit{for } i = 1, 2 \textit{ and } j = 1, 2, \\
&\|\bar{\eta}_j\|_{H^3(\mathbb{R})} + \left\|\frac{\partial \bar{\eta}_j}{\partial c_i}\right\|_{H^3(\mathbb{R})} + \left\|\frac{\partial \bar{\eta}_j}{\partial h}\right\|_{H^3(\mathbb{R})} \le Ce^{-ah} \quad \textit{for } i = 1, 2 \textit{ and } j = 3, 4,
\end{aligned}$$

where $d = \min_{i=1,2} \sqrt{c_i}$.

Proof. We give the proof for $j = 1$:

$$\begin{aligned}
&\|\bar{\eta}_1(\cdot, \mathbf{c}, h) - \eta_1(\cdot, c_1)\|_{H^1(\mathbb{R})} \\
=&\left\|\frac{1}{2\pi i}\oint_{|\lambda| = \frac{b'}{2}} \left((\lambda - \mathcal{A}^*_{\mathbf{c},h})^{-1} - (\lambda - A_1^*)^{-1}\right)\eta_1 d\lambda\right\|_{H^1(\mathbb{R})} \\
\le&\frac{1}{2\pi}\oint_{|\lambda| = \frac{b'}{2}} \left\|(\lambda - \mathcal{A}^*_{\mathbf{c},h})^{-1}V_2^*(\lambda - A_1^*)^{-1}\eta_1\right\|_{H^1(\mathbb{R})} d\lambda \\
+&\frac{1}{2\pi}\oint_{|\lambda| = \frac{b'}{2}} \left\|(\lambda - \mathcal{A}^*_{\mathbf{c},h})^{-1}(c_2 - c_1)\chi_2(\partial_y + a)(\lambda - A_1^*)^{-1}\eta_1\right\|_{H^1(\mathbb{R})} d\lambda \\
\le&C\oint_{|\lambda| = \frac{b'}{2}} \left(\left\|V_2^*\left(\lambda^{-1}\eta_1 - \lambda^{-2}\eta_2\right)\right\| + \left\|(c_2 - c_1)\chi_2(\partial_y + a)\left(\lambda^{-1}\eta_1 - \lambda^{-2}\eta_2\right)\right\|\right) d\lambda \\
\le&Ce^{-\frac{a+d}{3}h}.
\end{aligned}$$

Here we use

$$(\lambda - A_1^*)^{-1}\eta_1(\cdot, c_1) = \lambda^{-1}\eta_1(\cdot, c_1) - \lambda^{-2}\eta_2(\cdot, c_1).$$

The latter part of the lemma can be obtained by differentiating

$$\bar{\eta}_1(\cdot, \mathbf{c}, h) = \frac{1}{2\pi i}\oint_{|\lambda| = \frac{b'}{2}} (\lambda - \mathcal{A}^*_{\mathbf{c},h})^{-1}\eta_1(\cdot, c_1) d\lambda$$

with respect to c_i $(i = 1, 2)$ and h and estimating the right-hand side. ∎

Let $(M_\rho f)(x) := e^{\rho x} f(x)$. The following lemma shows that $\bar{\eta}_i(x, \mathbf{c}, h)$ $(1 \le i \le 4)$ decay like η_i as $x \to \pm\infty$.

LEMMA 2.4. *Let c_1, c_2, and h satisfy the assumptions in Lemma 2.3, and let ρ be a number satisfying $-\min_{i=1,2}\sqrt{c_i} < \rho < a$. Then*

$$\|M_\rho \bar{\eta}_i(\cdot, \mathbf{c}, h)\|_{H^3(\mathbb{R})} \le C \quad \text{for } i = 1, 2,$$
$$\|M_\rho \bar{\eta}_i(\cdot, \mathbf{c}, h)\|_{H^3(\mathbb{R})} \le Ce^{(\rho-a)h} \quad \text{for } i = 3, 4,$$

where C is a positive constant depending only on ρ.

Proof. Since $\mathcal{A}_{\mathbf{c},h}(\rho + a) = M_\rho \mathcal{A}_{\mathbf{c},h}(a) M_{-\rho}$,

$$\begin{aligned} M_\rho \bar{\eta}_i(\cdot, \mathbf{c}, h, a) &= \frac{1}{2\pi i} \oint_{|\lambda| = \frac{b'}{2}} M_\rho (\lambda - \mathcal{A}_{\mathbf{c},h}(a)^*)^{-1} \eta_i(\cdot, \mathbf{c}, h, a) d\lambda \\ &= \frac{1}{2\pi i} \oint_{|\lambda| = \frac{b'}{2}} (\lambda - \mathcal{A}_{\mathbf{c},h}(a-\rho)^*)^{-1} \eta_i(\cdot, \mathbf{c}, h, a-\rho) d\lambda \\ &= \bar{\eta}_i(\cdot, \mathbf{c}, h, a-\rho). \end{aligned}$$

Hence it follows from Lemma 2.3 that

$$\|M_\rho \bar{\eta}_i(\cdot, \mathbf{c}, h)\|_{H^3(\mathbb{R})} \le \begin{cases} C & \text{for } i = 1, 2, \\ Ce^{(\rho-a)h} & \text{for } i = 3, 4, \end{cases}$$

where $-\min_{i=1,2}\sqrt{c_i} < \rho < a$. ∎

2.3. Decay properties of the linearized equation. In the case where the speeds of the solitary waves are almost the same, the properties of the linearized equation are similar to those of linearized equation around the multibump function. In this subsection, we investigate the decay properties of solutions of linearized equations in an exponentially weighted space.

Let ε_* and h_* be numbers as in Lemmas 2.2 and 2.3, and let $\mathbf{c} = (c_1, c_2)$ and c_1, c_2, and h be positive numbers satisfying $|c_2 - c_1| \le \varepsilon_*$ and $h \ge h_*$. Let $\bar{\xi}_i(\cdot, \mathbf{c}, h)$ $(1 \le i \le 4)$ be (generalized) eigenfunctions of $\mathcal{A}_{\mathbf{c},h}$ satisfying

$$\left\langle \bar{\xi}_i(\cdot, \mathbf{c}, h), \bar{\eta}_j(\cdot, \mathbf{c}, h) \right\rangle = \delta_{ij}.$$

We introduce the subspace of $L^2(\mathbb{R})$ defined by

$$X(\mathbf{c}, h) = \left\{ u \in L^2(\mathbb{R}) \,|\, \langle u, \bar{\eta}_i(\cdot, \mathbf{c}, h) \rangle = 0 \text{ for } 1 \le i \le 4 \right\}.$$

Let $\mathbf{c_0} = (c_{1,0}, c_{2,0})$ and $c_{1,0}$, $c_{2,0}$, and h_0 be positive number with $|c_{2,0} - c_{1,0}| \le \varepsilon_*$ and $h_0 \ge h_*$. We define the operator $\Pi(\mathbf{c}, h; \mathbf{c_0}, h_0)$ from $X(\mathbf{c_0}, h_0)$ to $X(\mathbf{c}, h)$ by

$$\Pi(\mathbf{c}, h; \mathbf{c_0}, h_0) u = w(1; \mathbf{c}, h; \mathbf{c_0}, h_0),$$

where $w(\theta; \mathbf{c}, h; \mathbf{c_0}, h_0)$ is the solution of

$$\begin{cases} \dfrac{dw}{d\theta} = \displaystyle\sum_{i=1,2} \sum_{1 \le j \le 4} (c_{i,0} - c_i) \left\langle w(\theta), \frac{\partial \bar{\eta}_j}{\partial c_i}(\mathbf{c}(\theta), h(\theta)) \right\rangle \bar{\xi}_j(\mathbf{c}(\theta), h(\theta)) \\ \qquad\qquad + (h_0 - h) \displaystyle\sum_{1 \le j \le 4} \left\langle w(\theta), \frac{\partial \bar{\eta}_j}{\partial h}(\mathbf{c}(\theta), h(\theta)) \right\rangle \bar{\xi}_j(\mathbf{c}(\theta), h(\theta)), \\ w(0) = u \in X(\mathbf{c_0}, h_0), \end{cases}$$

$\mathbf{c}(\theta) = \theta \mathbf{c} + (1 - \theta)\mathbf{c_0}$, and $h(\theta) = \theta h + (1 - \theta) h_0$.

Set

$$\mathcal{K}(h_0, \mathbf{c_0}, \rho, \delta) = \left\{(c_1, c_2, h) \;\middle|\; |h - h_0| \le \rho,\ \ |c_i - c_{i,0}| \le \delta \text{ for } i = 1, 2\right\}.$$

The construction of $\Pi(\mathbf{c}, h; \mathbf{c_0}, h_0)$ yields that $\Pi(\mathbf{c}, h; \mathbf{c_0}, h_0)$ is an isomorphism satisfying the following.

LEMMA 2.5. *There exist positive numbers h_* and δ such that for every $0 < \rho < h_*/2$ and $|c_{2,0} - c_{1,0}| \le \delta$*

$$\sup_{h_0 \ge h_*} \sup_{\mathcal{K}(h_0, \mathbf{c_0}, \rho, \delta)} \left(\|\Pi(\mathbf{c}, h; \mathbf{c_0}, h_0)\|_{\mathcal{L}(L^2)} + \|\Pi(\mathbf{c}, h; \mathbf{c_0}, h_0)^{-1}\|_{\mathcal{L}(L^2)}\right) < \infty,$$

$$\sup_{h_0 \ge h_*} \sup_{\mathcal{K}(h_0, \mathbf{c_0}, \rho, \delta)} \left(\sum_{i=1,2} \left\| \frac{\partial \Pi}{\partial c_i}(\mathbf{c}, h; \mathbf{c_0}, h_0) \right\|_{\mathcal{L}(L^2)} + \left\| \frac{\partial \Pi}{\partial h}(\mathbf{c}, h; \mathbf{c_0}, h_0) \right\|_{\mathcal{L}(L^2)} \right) < \infty.$$

Next, we obtain the local decay estimate of the solutions of the linearized equation. Let

$$\bar{\mathcal{A}}(t) = \Pi(\mathbf{c}(t), h(t); \mathbf{c_0}, h_0) \mathcal{A}_{\mathbf{c}(t), h(t)} \Pi(\mathbf{c}(t), h(t); \mathbf{c_0}, h_0)^{-1},$$

and let $U(t, s)f$ denote the solution of

$$\begin{cases} \partial_t w = \bar{\mathcal{A}}(t) w \quad \text{for } t \ge s \text{ and } y \in \mathbb{R}, \\ w(s, \cdot) = f \in X(\mathbf{c_0}, h_0). \end{cases}$$

LEMMA 2.6. *Assume* (2.6). *Let c_0, h_0, ε, δ, and ρ be positive numbers with $0 < \delta < c_0$, and let $0 < a_1 \le a_2 < \sqrt{c_0/6}$ and $a \in [a_1, a_2]$. Suppose that $c_i(t)$ $(i = 1, 2)$ and $h(t)$ are C^1-functions satisfying*

$$\begin{aligned} &h(t) \in [h_0, h_0 + \rho], \\ &c_i(t) \in [c_0 - \delta, c_0 + \delta], \\ &|c_2(t) - c_1(t)| + |\dot{c}_1(t)| + |\dot{c}_2(t)| + |\dot{h}(t)| \le \varepsilon \end{aligned} \tag{2.10}$$

for every $t \ge 0$. Then there exists a positive number h_* satisfying the following. Suppose $h_0 \ge h_*$. Then for every ρ there exists $\varepsilon_* > 0$ such that if $0 < \varepsilon \le \varepsilon_*$,

$$\|U(t, s)f\| \le M e^{-b(t-s)} \|f\|, \tag{2.11}$$

$$\|U(t, s)f\|_{1,0} \le M (t - s)^{-\frac{1}{2}} e^{-b(t-s)} \|f\|, \tag{2.12}$$

$$\|U(t, s)f\|_{1,0} \le M e^{-b(t-s)} \|f\|_{1,0} \tag{2.13}$$

for any $t \ge s \ge 0$ and $f \in H^1(\mathbb{R})$. The constants M and b are positive, chosen uniformly with respect to $a \in [a_1, a_2]$ and $h_0 \ge h_*$, and depend only on ρ and h_*.

Proof. The proof follows along the lines of [13]. First, we estimate the operator norm of $e^{\bar{\mathcal{A}}(s)t}$ for each $s \ge 0$. Let $c \in [c_0 - \delta, c_0 + \delta]$ and let $0 < \alpha < a < (c_0 - \delta/3)^{\frac{1}{2}}$. By Lemma 4.3 in [36], there exists a $C > 0$ such that for $\lambda \in \overline{\Omega(\alpha)}$

$$\left\| \partial_y^n (\lambda - A_{0,c})^{-1} \right\|_{\mathcal{L}(L^2(\mathbb{R}))} \le C |\lambda|^{\frac{n-2}{3}} \quad \text{for } n = 0, 1. \tag{2.14}$$

The constant C can be chosen uniformly with respect to $c \in [c_0 - \delta, c_0 + \delta]$. By (2.6) and (2.14), there exist C_0, $C_1 > 0$ such that for $\lambda \in \overline{\Omega(\alpha)}$ with $|\lambda| \geq C_0$,

$$
\begin{aligned}
&\left\|\partial_y^n(\lambda - \mathcal{A}_{\mathbf{c},h})^{-1}\right\|_{\mathcal{L}(L^2)} \\
&= \left\| \partial_y^n (\lambda - A_{0,c})^{-1} \left\{ 1 - \sum_{i=1,2} \left((c_i - c)(\partial_y - a)\chi_i + V_i\right)(\lambda - A_{0,c})^{-1} \right\}^{-1} \right\|_{\mathcal{L}(L^2)} \\
&\leq C|\lambda|^{\frac{n-2}{3}} \quad \text{for } n = 0, 1,
\end{aligned} \tag{2.15}
$$

where C can be chosen uniformly with respect to $c_i \in [c_0 - \delta, c_0 + \delta]$ and $h \geq h_*$.

Let b' be a positive number as in Lemma 2.2 and let $\lambda \in \overline{\Omega(\alpha)}$ with $\mathrm{Re}\lambda = -\bar{b} > -b'$. We compute

$$
\begin{aligned}
&\left(Q_1(\lambda - A_1)^{-1}\chi_1 + Q_2(\lambda - A_2)^{-1}\chi_2\right)(\lambda - \mathcal{A}_{\mathbf{c},h}) \\
&= Q_1\chi_1 + Q_2\chi_2 + \tilde{r}(\lambda),
\end{aligned}
$$

where

$$
\tilde{r}(\lambda) = \sum_{i=1,2} Q_i(\lambda - A_i)^{-1}\left\{[A_i, \chi_i] - \chi_i V_{3-i} + (-1)^i(c_2 - c_1)(\partial_y - a)(\chi_{3-i}\cdot)\right\}.
$$

Let

$$
\mathcal{Q}_{\mathbf{c},h} = 1 - \mathcal{P}_{\mathbf{c},h} \quad \text{and} \quad T(\mathbf{c}, h, \lambda) = \mathcal{Q}_{\mathbf{c},h}(Q_1\chi_1 + Q_2\chi_2 + \tilde{r}(\lambda))\mathcal{Q}_{\mathbf{c},h}.
$$

By (2.8) and (2.9), we have

$$
\lim_{h\to\infty, c_2 \to c_1} \|\tilde{r}(\lambda)\|_{\mathcal{L}(L^2)} = 0, \qquad \lim_{h\to\infty, c_2 \to c_1} \|\mathcal{Q}_{\mathbf{c},h} - Q_1(c_1)\chi_1 - Q_2(c_2)\chi_2\|_{\mathcal{L}(L^2)} = 0
$$

uniformly with respect to λ and c_1. Hence, for sufficiently large h_* and sufficiently small ε_*, $T(\mathbf{c}, h, \lambda)^{-1} : X(\mathbf{c}, h) \to X(\mathbf{c}, h)$ exists and is uniformly bounded with respect to λ, c_1, c_2, and h. Since

$$
(\lambda - \mathcal{A}_{\mathbf{c},h})^{-1}\big|_{X_{\mathbf{c},h}} = T(\mathbf{c}, h, \lambda)^{-1}\mathcal{Q}_{\mathbf{c},h} \sum_{i=1,2} Q_i(\lambda - A_i)^{-1}\chi_i\mathcal{Q}_{\mathbf{c},h},
$$

$$
\sup_{\substack{|\lambda| \leq C_0, \mathrm{Re}\lambda = -\bar{b}, \\ c_i \in [c_0 - \delta, c_0 + \delta], h \geq h_*}} \left\|(\lambda - \mathcal{A}_{\mathbf{c},h})^{-1}\right\|_{X_{\mathbf{c},h}} < \infty. \tag{2.16}
$$

Let $\Gamma = \{\lambda \in S(\alpha) \,|\, |\lambda| \geq C_0\} \cup \{\lambda \in \mathbb{C} \,|\, \mathrm{Re}\lambda = -\bar{b}, |\lambda| \leq C_0\}$. From (2.15)–(2.16) and Lemmas 2.3 and 2.5, it follows that

$$
\sup_{s \geq 0, \lambda \in \Gamma} \|\partial_y^n(\lambda - \bar{\mathcal{A}}(s))^{-1}\|_{\mathcal{L}(X(\mathbf{c},h))} \leq C|\lambda|^{\frac{n-2}{3}} \text{ for } n = 0, 1.
$$

Now, by using the inversion of the Laplace formula as in [36, pp. 325–328], we have, for a $C > 0$,

$$
\begin{aligned}
&\|\mathcal{Q}_{\mathbf{c}(s),h(s)}e^{\bar{\mathcal{A}}(s)t}f\| \leq Ce^{-bt}\|f\|, \\
&\|\mathcal{Q}_{\mathbf{c}(s),h(s)}e^{\bar{\mathcal{A}}(s)t}f\|_{1,0} \leq Ce^{-bt}\|f\|_{1,0}, \\
&\|\mathcal{Q}_{\mathbf{c}(s),h(s)}e^{\bar{\mathcal{A}}(s)t}f\|_{1,0} \leq Ct^{-\frac{1}{2}}e^{-bt}\|f\|
\end{aligned}
$$

for every t, $s \geq 0$ and $f \in H^1(\mathbb{R})$.

By Lemma 2.5 and (2.10), there exists a $C > 0$ such that

$$\|\bar{\mathcal{A}}(t_2)f - \bar{\mathcal{A}}(t_1)f\| \le C\varepsilon|t_2 - t_1|\|f\|_{1,0}.$$

Combining the above with the proof of Theorem 7.4.2 in [18], we have an $\varepsilon_* > 0$ such that (2.11)–(2.12) hold for b with $0 < b < \bar{b}$ if $\varepsilon < \varepsilon_*$. We see that (2.13) follows from (2.12) and the fact that $U(t,s)$ is a semigroup on H^1. □

3. Decomposition of the 2-pulse solutions. To decompose the solution into solitary wave parts and a dispersive part, set

$$\begin{aligned} u(x,t) &= \varphi_{c_1(t)}(x - x_1(t)) + \varphi_{c_2(t)}(x - x_2(t)) + v(y,t), \\ y &= x - x_1(t), \qquad h(t) = x_2(t) - x_1(t). \end{aligned} \tag{3.1}$$

Substituting (3.1) into (1.1), we have

$$\partial_t v + \partial_y \mathcal{L} v + \tilde{l} + \partial_y \tilde{\mathcal{N}} = 0,$$

where

$$\begin{aligned} &\mathcal{L} = \partial_y^2 - c_1 + f'(\varphi_{c_1}) + f'(\tau_h \varphi_{c_2}), \\ &\tilde{l} = (c_1 - \dot{x}_1)\partial_y v + \dot{c}_1 \partial_c \varphi_{c_1} + (c_1 - \dot{x}_1)\partial_y \varphi_{c_1} + \dot{c}_2 \tau_h \partial_c \varphi_{c_2} + (c_2 - \dot{x}_2)\tau_h \partial_y \varphi_{c_2}, \\ &\tilde{\mathcal{N}} = \tilde{\mathcal{N}}_1 + \tilde{\mathcal{N}}_2 + \tilde{\mathcal{N}}_3, \\ &\tilde{\mathcal{N}}_1 = f(\varphi_{c_1} + \tau_h \varphi_{c_2} + v) - f(\varphi_{c_1} + \tau_h \varphi_{c_2}) - f'(\varphi_{c_1} + \tau_h \varphi_{c_2})v, \\ &\tilde{\mathcal{N}}_2 = f(\varphi_{c_1} + \tau_h \varphi_{c_2}) - f(\varphi_{c_1}) - f(\tau_h \varphi_{c_2}), \\ &\tilde{\mathcal{N}}_3 = \left(f'(\varphi_{c_1} + \tau_h \varphi_{c_2}) - f'(\varphi_{c_1}) - f'(\tau_h \varphi_{c_2})\right) v. \end{aligned}$$

Let $w(y,t) = e^{ay} v(y,t)$. Then

$$\partial_t w - \mathcal{A}_{\mathbf{c}(t),h(t)} w + l + (\partial_y - a)\mathcal{N} = 0, \tag{3.2}$$

where $l = e^{ay}\tilde{l} + (c_2 - c_1)(\partial_y - a)(\chi_2 w)$ and $\mathcal{N} = e^{ay}\tilde{\mathcal{N}}$. To fix the components of (3.1), we assume the following condition:

$$w(\cdot,t) \in \text{Range}\left(\mathcal{Q}_{\mathbf{c}(t),h(t)}\right). \tag{3.3}$$

This requirement corresponds to

$$\langle w(\cdot,t),\, \bar{\eta}_i(\cdot,\, \mathbf{c}(t), h(t))\rangle = 0 \quad \text{for } i = 1,\, 2,\, 3,\, 4, \tag{3.4}$$

where $\mathbf{c}(t) = (c_1(t), c_2(t))$, which can be satisfied locally in time.

LEMMA 3.1. *Let $1 < p < 5$, $0 < a < d_0/3$, and $0 \le t_0 \le 1$. Assume that (2.6) holds and that $u_0 \in H^1_a(\mathbb{R}) \cap H^1(\mathbb{R})$. Let I be a compact subset of $(0,\infty)$. Then there exist positive numbers ε, δ_0, δ_1, and h_* such that if $c_{1,0}$, $c_{2,0} \in I$, $|c_{2,0} - c_{1,0}| < \varepsilon$, $h_0 = x_{2,0} - x_{1,0} \ge h_*$ and the solution u of (1.1) satisfies*

$$\sup_{0\le t\le t_0} \left\|u(t, \cdot + x_{1,0}) - \varphi_{c_{1,0}}(\cdot - c_{1,0}t) - \varphi_{c_{2,0}}(\cdot - c_{2,0}t - h_0)\right\|_{H^1_a} < \delta_0,$$

there exists a unique function $(x_1(t), c_1(t), x_2(t), c_2(t)) \in C^1([0,t_0];\mathbb{R}^4)$ satisfying (3.4) and

$$\sup_{0\le t\le t_0} \sum_{i=1,2} \left(|x_i(t) - x_{i,0} - c_{i,0}t| + |c_i(t) - c_{i,0}|\right) < \delta_1.$$

The number δ_0 may be chosen as the decreasing function of t_0.

Proof. Let

$$\begin{aligned}&F_i[u, x_1, c_1, x_2, c_2]\\ &= \int_{\mathbb{R}} e^{ay} \left\{u(t,x) - \varphi_{c_1(t)}(x - x_1(t)) - \varphi_{c_2(t)}(x - x_2(t))\right\} \bar{\eta}_i(x - x_1(t), \mathbf{c}, h) dx\end{aligned}$$

be the functional defined on $C([0,t_0]; H^1_a(\mathbb{R}))$ for $1 \le i \le 4$. Let

$$\Phi_0 = (\varphi_{c_{1,0}}(\cdot - c_{1,0}t - x_{1,0}) + \varphi_{c_{2,0}}(\cdot - c_{2,0}t - x_{2,0}), x_{1,0} + c_{1,0}t, c_{1,0}, x_{2,0} + c_{2,0}t, c_{2,0}).$$

Then it follows that $F[\Phi_0] = 0$, and in view of Lemma 2.3, the Fréchet derivatives of ${}^t(F_1, F_2, F_3, F_4)$ at Φ_0 with respect to x_1, c_1, x_2, and c_2 are given by

$$\begin{aligned}\frac{\partial(F_1, F_2, F_3, F_4)}{\partial(x_1, c_1, x_2, c_2)}[\Phi_0] &= -\left(\left\langle \xi_i(\cdot, c_{[\frac{i+1}{2}]}, h), \bar{\eta}_j(\cdot, \mathbf{c}, h)\right\rangle\right)_{\substack{i=1,2,3,4\downarrow\\ j=1,2,3,4\rightarrow}}\\ &= -\left(\langle \xi_i(\cdot, c_1, h), \eta_j(\cdot, c_1, h)\rangle\right)_{\substack{i=1,2,3,4\downarrow\\ j=1,2,3,4\rightarrow}} + O(|c_2 - c_1| + e^{-\frac{d-2a}{3}h}).\end{aligned}$$

By (2.1), (2.2), (2.5), and (1.3),

$$(\langle \xi_i(c,h), \eta_j(c,h)\rangle)_{\substack{i=1,2,3,4\rightarrow\\ j=1,2,3,4\downarrow}} = B(c) + O(he^{-\sqrt{c}h}), \tag{3.5}$$

where

$$B(c) = \begin{pmatrix} 1 & 0 & 0 & \gamma(c)\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix}, \qquad \gamma(c) = \left(\frac{d}{dc}\int_{\mathbb{R}} \varphi_c\right)^2 \theta_1(c).$$

Applying the implicit function theorem, we can choose x_i and c_i $(i = 1, 2)$ satisfying (3.4). We remark that if a value of δ_0 works for some t_0, it also works for smaller values of t_0 because the functionals F_i $(1 \le i \le 4)$ depend only on u, c_i, and x_i $(i = 1, 2)$. For the details of the proof, we refer the readers to [36]. □

Remark 3.1. Since

$$u \in C([0,\infty); H^1(\mathbb{R})) \text{ and } e^{ax}u \in C([0,\infty); H^1(\mathbb{R}))$$

(see [16, 21]), there exist $T > 0$ and $(x_1(t), c_1(t), x_2(t), c_2(t)) \in C^1([0,T]; \mathbb{R}^4)$ satisfying (3.4) for $0 \le t \le T$.

Remark 3.2. Since $H^1_a \ni u \mapsto (x_1, c_1, x_2, c_2) \in C([0,t_0]; \mathbb{R}^4)$ is a C^1-mapping, we have

$$|x_1(0) - x_{1,0}| + |c_1(0) - c_{1,0}| \le C\|v_0\|_{H^1_a}, \tag{3.6}$$

$$|x_2(0) - x_{2,0}| + |c_2(0) - c_{2,0}| \le Ce^{-ah_0}\|v_0\|_{H^1_a} \tag{3.7}$$

in some neighborhood of Φ_0. Hence it follows that $h(0) = h_0 + O(\|v_0\|_{H^1_a})$,

$$\begin{aligned}\|v(0)\|_{1,0} &\le \|v_0\|_{1,0} + \sum_{i=1,2} \left\|\varphi_{c_{i,0}}(\cdot - x_{i,0}) - \varphi_{c_i(0)}(\cdot - x_i(0))\right\|_{1,0}\\ &\le \|v_0\|_{1,0} + C\sum_{i=1,2}(|x_{i,0} - x_i(0)| + |c_{i,0} - c_i(0)|) \qquad (3.8)\\ &\le \|v_0\|_{1,0} + C\|v_0\|_{H^1_a},\end{aligned}$$

and

$$
\begin{aligned}
\|w(0)\|_{1,0} &\le e^{a(x_{1,0}-x_1(0))}\|v_0\|_{H^1_a} \\
&\quad + \sum_{i=1,2} \left\| e^{a(\cdot - x_1(0))}\varphi_{c_{i,0}}(\cdot - x_{i,0}) - \varphi_{c_i(0)}(\cdot - x_i(0)) \right\|_{1,0} \\
&\le C\left(\|v_0\|_{H^1_a} + |x_{1,0} - x_1(0)| + |c_{1,0} - c_1(0)|\right) \\
&\quad + Ce^{ah_0}\left(|x_{2,0} - x_2(0)| + |c_{2,0} - c_2(0)|\right) \\
&\le C\|v_0\|_{H^1_a}.
\end{aligned} \tag{3.9}
$$

Finally, we derive the system of ordinary differential equations of modulating speeds and phase of solitary waves. Let $\alpha = \alpha(c_1)$, $\theta_i = \theta_i(c_1)$ $(i = 1, 2, 3)$. Let a and b be positive numbers and

$$
M_1(a,b) = \left\langle f'(\varphi_a)e^{\sqrt{b}y}, \partial_c\varphi_a \right\rangle \quad \text{and} \quad M_2(a,b) = \left\langle f'(\varphi_a)e^{\sqrt{b}y}, \partial_y\varphi_a \right\rangle,
$$

and $M_i = M_i(c_1, c_1)$ $(i = 1, 2)$. We remark that $M_2(a,b) < 0$ because $\partial_y\varphi_a$ is an odd function and $e^{\sqrt{b}y}$ is a positive increasing function.

LEMMA 3.2. *Let $3 \le p < 5$ and let (2.6) be satisfied. Let I be a compact subset of $(0,\infty)$. There exist positive numbers ε and h_* such that if $c_i(t) \in I$ $(i = 1, 2)$, $|c_2(t) - c_1(t)| < \varepsilon$, $h(t) \ge h_*$, $0 < a < d(t)/3$, and (3.4) holds for $t \in [0, T]$,*

$$
\begin{pmatrix} c_1 - \dot{x}_1 \\ \dot{c}_1 \\ c_2 - \dot{x}_2 \\ \dot{c}_2 \end{pmatrix} = \alpha \begin{pmatrix} (\theta_1 M_1 + \theta_2 M_2)e^{-\sqrt{c_2}h} - 2\theta_2 M_2 e^{-\sqrt{c_1}h} \\ \theta_3 M_2 e^{-\sqrt{c_2}h} \\ (\theta_1 M_1 - \theta_2 M_2)e^{-\sqrt{c_1}h} \\ -\theta_3 M_2 e^{-\sqrt{c_1}h} \end{pmatrix} + \vec{\mathcal{R}} + \begin{pmatrix} \mathcal{R}_1 \\ \mathcal{R}_2 \\ \mathcal{R}_3 \\ \mathcal{R}_4 \end{pmatrix}
$$

for $0 \le t \le T$, where $\mathcal{R}$ is a vector with $|\vec{\mathcal{R}}| = O\left((\|w\|_{1,0} + |g| + e^{-(d-2a)h/3})e^{-dh}\right)$ and $\mathcal{R}_i$ $(1 \le i \le 4)$ satisfy

$$
\begin{aligned}
|\mathcal{R}_1| + |\mathcal{R}_2| &\le C(|g| + e^{-dh} + \|w\|_{1,0})\|w\|_{1,0} + Ce^{-\frac{4d-2a}{3}h}, \\
|\mathcal{R}_3| + |\mathcal{R}_4| &\le Ce^{-ah}(|g| + e^{-dh} + \|w\|_{1,0})\|w\|_{1,0} + Ce^{-\frac{4d-a}{3}h},
\end{aligned}
$$

where $d(t) = \min_{i=1,2}\sqrt{c_i(t)}$.

Proof. To prove the lemma, we translate (3.4) into a system of ordinary differential equations. Differentiating (3.4), we obtain

$$
\begin{aligned}
\frac{d}{dt}\langle w(\cdot,t),\, \bar{\eta}_j(\cdot,\, \mathbf{c}(t), h(t))\rangle &= \sum_{i=1,2} \dot{c}_i \langle w,\, \partial_{c_i}\bar{\eta}_j(\cdot,\, \mathbf{c}(t), h(t))\rangle \\
&\quad + \dot{h}\langle w,\, \partial_h\bar{\eta}_j(\cdot,\, \mathbf{c}(t), h(t))\rangle + \langle \partial_t w,\, \bar{\eta}_j(\cdot,\, \mathbf{c}(t), h(t))\rangle \\
&= 0.
\end{aligned} \tag{3.10}
$$

By (3.2) and (3.4),

$$
\langle \partial_t w,\, \bar{\eta}_j(\cdot,\, \mathbf{c}, h)\rangle = -\langle l,\, \bar{\eta}_j(\cdot,\, \mathbf{c}, h)\rangle + \langle \mathcal{N},\, (\partial_y + a)\bar{\eta}_j(\cdot,\, \mathbf{c}, h)\rangle
$$

and

$$\begin{aligned}
&\langle l, \bar{\eta}_j(\cdot,\, \mathbf{c}, h)\rangle \\
&= g\,\langle(\partial_y - a)\chi_2 w, \bar{\eta}_j(\cdot,\, \mathbf{c}, h)\rangle \\
&\quad + (c_1 - \dot{x}_1)\,\{\langle((\partial_y - a)w, \bar{\eta}_j(\cdot,\, \mathbf{c}, h)\rangle + \langle\xi_1, \bar{\eta}_j(\cdot,\, \mathbf{c}, h)\rangle\} \\
&\quad + \dot{c}_1\,\langle\xi_2, \bar{\eta}_j(\cdot,\, \mathbf{c}, h)\rangle + (c_2 - \dot{x}_2)\,\langle\xi_3, \bar{\eta}_j(\cdot,\, \mathbf{c}, h)\rangle + \dot{c}_2\,\langle\xi_4, \bar{\eta}_j(\cdot,\, \mathbf{c}, h)\rangle\,,
\end{aligned} \tag{3.11}$$

where $g(t) = c_2(t) - c_1(t)$, $\xi_j = \xi_j(\cdot, c_1)$ for $j = 1, 2$ and $\xi_j = \xi_j(\cdot, c_2, h)$ for $j = 3, 4$. Combining (3.10) and (3.11) yields

$$\mathcal{B}\begin{pmatrix} c_1 - \dot{x}_1 \\ \dot{c}_1 \\ c_2 - \dot{x}_2 \\ \dot{c}_2 \end{pmatrix} = (\langle\mathcal{N}_2,\, (\partial_y + a)\eta_j\rangle)_{j=1,2,3,4\downarrow} + \begin{pmatrix} \widetilde{\mathcal{R}}_1 \\ \widetilde{\mathcal{R}}_2 \\ \widetilde{\mathcal{R}}_3 \\ \widetilde{\mathcal{R}}_4 \end{pmatrix}, \tag{3.12}$$

where $\mathcal{N}_i = e^{ay}\tilde{\mathcal{N}}_i$ for $i = 1, 2, 3$, $\eta_j = \eta_j(\cdot, c_1)$ for $j = 1, 2$, $\eta_j = \eta_j(\cdot, c_2, h)$ for $j = 3, 4$, $\mathcal{B} = (b_{j1}, b_{j2}, b_{j3}, b_{j4})_{j=1,2,3,4\downarrow}$,

$$\begin{aligned}
&b_{j1} = \langle\xi_1, \bar{\eta}_j\rangle + \langle(\partial - a)w, \bar{\eta}_j\rangle - \langle w, \partial_h\bar{\eta}_j\rangle, \quad b_{j2} = \langle\xi_2, \bar{\eta}_j\rangle - \langle w, \partial_{c_1}\bar{\eta}_j\rangle, \\
&b_{j3} = \langle\xi_3, \bar{\eta}_j\rangle + \langle w, \partial_h\bar{\eta}_j\rangle, \quad b_{j4} = \langle\xi_4, \bar{\eta}_j\rangle - \langle w, \partial_{c_2}\bar{\eta}_j\rangle,
\end{aligned} \tag{3.13}$$

and

$$\widetilde{\mathcal{R}}_j = \sum_{i=1,3}\langle\mathcal{N}_i, (\partial_y + a)\bar{\eta}_j\rangle + \langle\mathcal{N}_2, (\partial_y + a)(\bar{\eta}_j - \eta_j)\rangle + g\langle w, \chi_2(\partial_y + a)\bar{\eta}_j + \partial_h\bar{\eta}_j\rangle.$$

By Lemma 2.3 and (3.5),

$$\begin{aligned}
&b_{jk} = \delta_{jk} + O(e^{-\frac{d-2a}{3}h} + \|w\|) \quad \text{for } j = 1, 2,\ 1 \le k \le 4 \text{ and } (j, k) \ne (1, 4), \\
&b_{jk} = \delta_{jk} + O(e^{-\frac{d-a}{3}h} + e^{-ah}\|w\|) \quad \text{for } j = 3, 4 \text{ and } 1 \le k \le 4, \\
&b_{41} = \gamma(c_1) + O(|g| + e^{-\frac{d-2a}{3}h} + \|w\|).
\end{aligned}$$

Since $\varphi_c(x) = c^{\frac{1}{p-1}}\varphi_1(c^{\frac{1}{2}}x)$ and $\varphi_1 \in L^\infty$,

$$\|\mathcal{N}_1\| = \left\|\int_0^1 (1-\theta)f''(\varphi_{c_1} + \tau_h\varphi_{c_2} + \theta v)d\theta v w\right\| \tag{3.14}$$

$$\le C(\|\varphi_{c_1}\|_\infty + \|\varphi_{c_2}\|_\infty + \|v\|_\infty)^{p-2}\|v\|_\infty\|w\| \tag{3.15}$$

$$\le C(c_1^{\frac{1}{p-1}} + c_2^{\frac{1}{p-1}} + \|v\|_\infty)^{p-2}\|v\|_\infty\|w\|. \tag{3.16}$$

By (1.3),

$$\begin{aligned}
\|e^{ay}\varphi_{c_1}\tau_h\varphi_{c_2}\|_\infty &\le C(c_1c_2)^{\frac{1}{p-1}}\|e^{ay}e^{-\sqrt{c_1}|y| - \sqrt{c_2}|y-h|}\|_\infty \\
&\le C(c_1c_2)^{\frac{1}{p-1}}e^{(a-d)h},
\end{aligned}$$

and $\|\varphi_{c_1}\tau_h\varphi_{c_2}\|_\infty \le C(c_1c_2)^{\frac{1}{p-1}}e^{-dh}$, where C is a constant independent of c_1 and c_2.

Hence it follows that

$$\begin{aligned}\|\mathcal{N}_2\| &= \left\| e^{ay}\int_0^1\int_0^1 f''(\theta_1\varphi_{c_1}+\theta_2\tau_h\varphi_{c_2})d\theta_1 d\theta_2\varphi_{c_1}\tau_h\varphi_{c_2}\right\| \\ &\le C(\|\varphi_{c_1}\|_\infty^{p-3}\|\varphi_{c_1}\|+\|\varphi_{c_2}\|_\infty^{p-3}\|\varphi_{c_2}\|)\|e^{ay}\varphi_{c_1}\tau_h\varphi_{c_2}\|_\infty \\ &\le C(c_1+c_2)^{\frac{3}{4}+\frac{1}{p-1}}e^{(a-d)h},\end{aligned} \tag{3.17}$$

$$\begin{aligned}\|\mathcal{N}_3\| &= \left\| \int_0^1\int_0^1 f'''(\theta_1\varphi_{c_1}+\theta_2\tau_h\varphi_{c_2})d\theta_1 d\theta_2\varphi_{c_1}\tau_h\varphi_{c_2}w\right\| \\ &\le C(\|\varphi_{c_1}\|_\infty+\|\varphi_{c_2}\|_\infty)^{p-3}\|\varphi_{c_1}\tau_h\varphi_{c_2}\|_\infty\|w\| \\ &\le C(c_1+c_2)e^{-dh}\|w\|.\end{aligned} \tag{3.18}$$

By Lemma 2.4,

$$\begin{aligned}|\langle\mathcal{N}_1,(\partial_y+a)\bar\eta_j\rangle| &\le \|e^{ay}\mathcal{N}_1\|\|(\partial_y+2a)e^{-ay}\bar\eta_j\| \\ &\le C(c_1^{\frac{1}{p-1}}+c_2^{\frac{1}{p-1}}+\|v\|_\infty)^{p-2}\|w\|_{1,0}^2 \quad \text{for } j=1,\,2, \\ |\langle\mathcal{N}_1,(\partial_y+a)\bar\eta_j\rangle| &\le \|e^{ay}\mathcal{N}_1\|\|(\partial_y+2a)e^{-ay}\bar\eta_j\| \\ &\le Ce^{-ah}(c_1^{\frac{1}{p-1}}+c_2^{\frac{1}{p-1}}+\|v\|_\infty)^{p-2}\|w\|_{1,0}^2 \quad \text{for } j=3,\,4.\end{aligned} \tag{3.19}$$

Using Lemma 2.3 and (3.17)–(3.19), we have

$$\begin{aligned}|\widetilde{\mathcal{R}}_1|+|\widetilde{\mathcal{R}}_2| &\le C(|g|+e^{-dh}+\|w\|_{1,0})\|w\|_{1,0}+Ce^{-\frac{4d-2a}{3}h}, \\ |\widetilde{\mathcal{R}}_3|+|\widetilde{\mathcal{R}}_4| &\le Ce^{-ah}(|g|+e^{-dh}+\|w\|_{1,0})\|w\|_{1,0}+Ce^{-\frac{4d-a}{3}h},\end{aligned} \tag{3.20}$$

where C is a continuous function of c_1, $c_2\in(0,\infty)$ and $\|v\|_\infty$. Hence C is bounded as long as c_1 and c_2 remain in some compact interval of $(0,\infty)$ and $\|v\|_\infty$ remains bounded.

For $f(u)=|u|^{p-1}u/p$ $(p\ge 3)$,

$$\begin{aligned}f(a+b)-f(a)-f(b)-f'(a)b &= \int_0^1\{f'(s_1a+b)-f'(s_1a)\}ds_1a-f'(a)b \\ &= \int_0^1\int_0^1\{f''(s_1a+s_2b)-f''(s_1a)\}ds_1ds_2ab \\ &= \int_0^1\int_0^1(1-s_2)f'''(s_1a+s_2b)ds_1ds_2ab^2.\end{aligned}$$

Using the above, we have $\langle\tilde{\mathcal{N}}_2,\partial_y\tilde\eta_i\rangle = I_i+\langle\bar r_i,\partial_y\tilde\eta_i\rangle$ $(1\le i\le 4)$, where

$$\begin{aligned}I_1 &= \langle f'(\varphi_{c_1})\tau_h\varphi_{c_2},\theta_1(c_1)\partial_c\varphi_{c_1}+\theta_2(c_1)\partial_y\varphi_{c_1}\rangle, \\ I_2 &= \langle f'(\varphi_{c_1})\tau_h\varphi_{c_2},\theta_3(c_1)\partial_y\varphi_{c_1}\rangle, \\ I_3 &= \langle f'(\tau_h\varphi_{c_2})\varphi_{c_1},\theta_1(c_2)\tau_h\partial_c\varphi_{c_2}+\theta_2(c_2)\tau_h\partial_y\varphi_{c_2}\rangle, \\ I_4 &= \langle f'(\tau_h\varphi_{c_2})\varphi_{c_1},\theta_3(c_2)\tau_h\partial_y\varphi_{c_2}\rangle,\end{aligned}$$

$$\begin{aligned}\bar r_1=\bar r_2 &= \int_0^1\int_0^1(1-s_2)f'''(s_1\varphi_{c_1}+s_2\tau_h\varphi_{c_2})ds_1ds_2\varphi_{c_1}(\tau_h\varphi_{c_2})^2, \\ \bar r_3=\bar r_4 &= \int_0^1\int_0^1(1-s_1)f'''(s_1\varphi_{c_1}+s_2\tau_h\varphi_{c_2})ds_1ds_2(\varphi_{c_1})^2\tau_h\varphi_{c_2},\end{aligned}$$

Noting that $\partial_y \tilde{\eta}_i(y,c) \sim e^{-\sqrt{c}|y|}$ for $i = 1,2$ and that $\partial_y \tilde{\eta}_i(y,c,h) \sim e^{-\sqrt{c}|y-h|}$ for $i = 3,4$, we have $|\langle \bar{r}_i, \partial_y \tilde{\eta}_i \rangle| \le Che^{-2dh}$ for $1 \le i \le 4$.

By (1.3),

$$\begin{aligned}
I_1 &= \alpha(c_2) \left\langle f'(\varphi_{c_1}) \left(e^{-\sqrt{c_2}|y-h|} + O(e^{-(p-1)\sqrt{c_2}|y-h|}) \right), \theta_1(c_1)\partial_c \varphi_{c_1} + \theta_2(c_1)\partial_y \varphi_{c_1} \right\rangle \\
&= \alpha(c_2) \left\langle f'(\varphi_{c_1}) e^{\sqrt{c_2}y}, \theta_1(c_1)\partial_c \varphi_{c_1} + \theta_2(c_1)\partial_y \varphi_{c_1} \right\rangle e^{-\sqrt{c_2}h} + O\left(e^{-(p-1)dh} \right) \\
&= \alpha(c_2) \left(\theta_1(c_1) M_1(c_1,c_2) + \theta_2(c_1) M_2(c_1,c_2) \right) e^{-\sqrt{c_2}h} + O\left(e^{-(p-1)dh} \right), \\
I_2 &= \alpha(c_2) \left\langle f'(\varphi_{c_1}) \left(e^{-\sqrt{c_2}|y-h|} + O(e^{-(p-1)\sqrt{c_2}|y-h|}) \right), \theta_3(c_1)\partial_y \varphi_{c_1} \right\rangle \\
&= \alpha(c_2) \left\langle f'(\varphi_{c_1}) e^{\sqrt{c_2}y}, \theta_3(c_1)\partial_y \varphi_{c_1} \right\rangle e^{-\sqrt{c_2}h} + O\left(e^{-(p-1)dh} \right) \\
&= \alpha(c_2)\theta_3(c_1) M_2(c_1,c_2) e^{-\sqrt{c_2}h} + O\left(e^{-(p-1)dh} \right).
\end{aligned}$$

Noting that φ_{c_2} and $\partial_c \varphi_{c_2}$ are even functions and $\partial_y \varphi_{c_2}$ is an odd function, we compute

$$\begin{aligned}
I_3 &= \alpha(c_1) \left\langle f'(\tau_h \varphi_{c_2}) e^{-\sqrt{c_1}|y|}, \theta_1(c_2)\tau_h \partial_c \varphi_{c_2} + \theta_2(c_2)\tau_h \partial_y \varphi_{c_2} \right\rangle + O\left(e^{-(p-1)dh} \right) \\
&= \alpha(c_1) \left\langle f'(\varphi_{c_2}) e^{-\sqrt{c_1}|y-h|}, \theta_1(c_2)\partial_c \varphi_{c_2} - \theta_2(c_2)\partial_y \varphi_{c_2} \right\rangle + O\left(e^{-(p-1)dh} \right) \\
&= \alpha(c_1) \left(\theta_1(c_2) M_1(c_2,c_1) - \theta_2(c_2) M_2(c_2,c_1) \right) e^{-\sqrt{c_1}h} + O\left(e^{-(p-1)dh} \right), \\
I_4 &= \alpha(c_1) \left\langle f'(\tau_h \varphi_{c_2}) e^{-\sqrt{c_1}|y|}, \theta_3(c_2)\tau_h \partial_y \varphi_{c_2} \right\rangle + O\left(e^{-(p-1)dh} \right) \\
&= \alpha(c_1) \left\langle f'(\varphi_{c_2}) e^{-\sqrt{c_1}|y+h|}, \theta_3(c_2)\partial_y \varphi_{c_2} \right\rangle + O\left(e^{-(p-1)dh} \right) \\
&= -\alpha(c_1) \left\langle f'(\varphi_{c_2}) e^{-\sqrt{c_1}|y-h|}, \theta_3(c_2)\partial_y \varphi_{c_2} \right\rangle + O\left(e^{-(p-1)dh} \right) \\
&= -\alpha(c_1)\theta_3(c_2) M_2(c_2,c_1) e^{-\sqrt{c_1}h} + O\left(e^{-(p-1)dh} \right).
\end{aligned}$$

Since $M_i(a,b)$ $(i = 1,2)$, $\alpha(c)$, and $\theta_i(c)$ $(i = 1,2,3)$ are locally Lipschitz continuous,

$$\begin{aligned}
&|M_i(c_1,c_2) - M_i| + |M_i(c_2,c_1) - M_i| \le C|g| \quad \text{for } i = 1,2, \\
&|\alpha(c_2) - \alpha(c_1)| \le C|g|, \quad |\theta_i(c_2) - \theta_i(c_1)| \le C|g| \quad \text{for } i = 1,2,3,
\end{aligned}$$

where C can be chosen uniformly with respect to c_1, $c_2 \in I$ with $|c_2 - c_1| \le \varepsilon$. Combining the above, we have

(3.21)

$$\left(\langle \tilde{\mathcal{N}}_2, \partial_y \tilde{\eta}_j \rangle \right)_{j=1,2,3,4\downarrow} = \begin{pmatrix} \alpha \left(\theta_1 M_1 + \theta_2 M_2 \right) e^{-\sqrt{c_2}h} \\ \alpha\theta_3 M_2 e^{-\sqrt{c_2}h} \\ \alpha \left(\theta_1 M_1 - \theta_2 M_2 \right) e^{-\sqrt{c_1}h} \\ -\alpha\theta_3 M_2 e^{-\sqrt{c_1}h} \end{pmatrix} + O(|g|e^{-dh} + he^{-2dh}).$$

Now, let $\mathcal{B}^{-1} = (\beta_{jk})_{\substack{j=1,2,3,4\downarrow \\ k=1,2,3,4\rightarrow}}$. Then by (3.13),

$$\begin{aligned}
&\beta_{jj} = 1 + O(|g| + e^{-\frac{d-2a}{3}h} + \|w\|) \quad \text{for } 1 \le j \le 4,\\
&\beta_{jk} = O(e^{-\frac{d-2a}{3}h} + \|w\|) \quad \text{for } j = 1,2,\ 1 \le k \le 4 \text{ and } (j,k) \ne (1,4),\\
&\beta_{jk} = O(e^{-\frac{d-2a}{3}h} + e^{-ah}\|w\|) \quad \text{for } j = 3,4 \text{ and } 1 \le k \le 4,\\
&\beta_{14} = -\gamma(c_1) + O(|g| + e^{-\frac{d-2a}{3}h} + \|w\|).
\end{aligned}$$

Combining (3.12), (3.13), and (3.21), we have

$$\begin{aligned}
\begin{pmatrix} c_1 - \dot{x}_1 \\ \dot{c}_1 \\ c_2 - \dot{x}_2 \\ \dot{c}_2 \end{pmatrix} &= \mathcal{B}^{-1}\left(\langle \tilde{\mathcal{N}}_2, \partial_y \tilde{\eta}_i\rangle + \widetilde{\mathcal{R}}_j\right)_{j=1,2,3,4\downarrow}\\
&= \alpha \begin{pmatrix} (\theta_1 M_1 + \theta_2 M_2)e^{-\sqrt{c_2}h} - 2\theta_2 M_2 e^{-\sqrt{c_1}h} \\ \theta_3 M_2 e^{-\sqrt{c_2}h} \\ (\theta_1 M_1 - \theta_2 M_2)e^{-\sqrt{c_1}h} \\ -\theta_3 M_2 e^{-\sqrt{c_1}h} \end{pmatrix} + \vec{\mathcal{R}} + \begin{pmatrix} \mathcal{R}_1 \\ \mathcal{R}_2 \\ \mathcal{R}_3 \\ \mathcal{R}_4 \end{pmatrix},
\end{aligned}$$

where $(\mathcal{R}_j)_{j=1,2,3,4\downarrow} = \mathcal{B}^{-1}(\widetilde{\mathcal{R}}_j)_{j=1,2,3,4\downarrow}$ and

$$\begin{aligned}
\vec{\mathcal{R}} =& \mathcal{B}^{-1}\left(\langle \tilde{\mathcal{N}}_2, \partial_y \tilde{\eta}_i\rangle\right)_{j=1,2,3,4\downarrow} - \begin{pmatrix} 1 & 0 & 0 & -\gamma \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha\left(\theta_1 M_1 + \theta_2 M_2\right) e^{-\sqrt{c_2}h} \\ \alpha\theta_3 M_2 e^{-\sqrt{c_2}h} \\ \alpha\left(\theta_1 M_1 - \theta_2 M_2\right) e^{-\sqrt{c_1}h} \\ -\alpha\theta_3 M_2 e^{-\sqrt{c_1}h} \end{pmatrix}\\
=& O\left(\left(\|w\|_{1,0} + |g| + e^{-\frac{d-2a}{3}h}\right) e^{-dh}\right).
\end{aligned}$$

Thus we complete the proof. □

4. Proof of Theorem 1.1.

4.1. Decay property of w. In this subsection, we will obtain the decay estimate of w. To start with, we investigate the motion of solitary waves assuming the decay of $\|w(t)\|_{1,0}$ and the smallness of $\|v(t)\|_\infty$. For this purpose, we construct an energy inequality with respect to $g(t) = c_2(t) - c_1(t)$ and $h(t) = x_2(t) - x_1(t)$.

LEMMA 4.1. *Let $T > 0$, $\delta \in (0,1)$, and $0 < a < (2c_1(0))^{1/2}/50$. Assume that*

$$\bar{c} = \inf_{0\le t\le T} \min_{i=1,2} c_i(t) \ge \frac{49}{50} c_1(0) \tag{4.1}$$

and that there exist positive constants A and ε_2 such that

$$\|v(t)\|_\infty \le \varepsilon_2, \tag{4.2}$$

$$\|w(t)\|_{1,0} \le A(e^{-\delta b t}\|v_0\|_{H^1_a} + e^{(a-d(t))h}) \tag{4.3}$$

for $0 \le t \le T$. Let

$$E(t) = \frac{1}{2} g(t)^2 - \alpha\theta_3 M_2 \left(\frac{1}{\sqrt{c_1(t)}} e^{-\sqrt{c_1(t)}h(t)} + \frac{1}{\sqrt{c_2(t)}} e^{-\sqrt{c_2(t)}h(t)}\right),$$

where $\alpha = \alpha(c_1(t))$, $\theta_3 = \theta_3(c_1(t))$ *and* $M_2 = M_2(c_1(t), c_1(t))$. *Then there exists* $A_1 > 0$ *such that*

$$\left|\frac{dE}{dt}\right| \le A_1(E^{\frac{7}{6}+\frac{2}{3}\gamma} + Ee^{-\delta bt}\|v_0\|_{H_a^1} + E^{\frac{1}{2}}e^{-2\delta bt}\|v_0\|_{H_a^1}^2) \tag{4.4}$$

for $\gamma = 1 - a/\sqrt{\bar{c}} \ge 34/35$.

Remark 4.1. By the definitions of α, θ_3, and M_2, we have $\alpha\theta_3 M_2 < 0$. The function $E(t)$ can be interpreted as a sum of kinetic energy $g(t)^2/2$ and potential energy $-\alpha\theta_3 M_2(\frac{1}{\sqrt{c_1}}e^{-\sqrt{c_1}h} + \frac{1}{\sqrt{c_2}}e^{-\sqrt{c_2}h})$, where g is the relative velocity and h is the distance between the centers of two pulses. In the following, we will show that the energy $E(t)$ is *almost conserved* for a while and that the pulses are repulsive.

Proof of Lemma 4.1. By Lemma 3.2,

$$\dot{g} = -\alpha\theta_3 M_2(e^{-\sqrt{c_1}h} + e^{-\sqrt{c_2}h}) + O(R_2), \tag{4.5}$$

$$\dot{h} = g + \alpha(\theta_1 M_1 + \theta_2 M_2)(e^{-\sqrt{c_2}h} - e^{-\sqrt{c_1}h}) + O(R_2), \tag{4.6}$$

where

$$\begin{aligned} R_2 &= |\vec{\mathcal{R}}| + \sum_{i=1}^{4} |\mathcal{R}_i| \\ &= O\left((|g| + \|w\|_{1,0})(\|w\|_{1,0} + e^{-dh}) + e^{-\frac{4d-2a}{3}h}\right) \\ &= O\left(|g|(e^{-\delta bt}\|v_0\|_{H_a^1} + e^{(a-d)h}) + e^{-2\delta bt}\|v_0\|_{H_a^1}^2 + e^{-\frac{4d-2a}{3}h}\right). \end{aligned} \tag{4.7}$$

Here we use (3.20) and (4.3). Hence by the definition of $E(t)$, we have $g = O(E^{\frac{1}{2}})$ and

$$R_2 = O(E^{\frac{2}{3}(1+\gamma)} + E^{\frac{1}{2}}e^{-\delta bt}\|v_0\|_{H_a^1} + e^{-2\delta bt}\|v_0\|_{H_a^1}^2). \tag{4.8}$$

Thus Lemma 3.2, (4.5), and (4.6) imply

$$\begin{aligned} \frac{dE}{dt} &= g\dot{g} + \alpha\theta_3 M_2(e^{-\sqrt{c_1}h} + e^{-\sqrt{c_2}h})\dot{h} + \frac{\alpha\theta_3 M_2}{2}\dot{c}_1 e^{-\sqrt{c_1}h}\left(c_1^{-\frac{3}{2}} + {c_1}^{-1}h\right) \\ &\quad + \frac{\alpha\theta_3 M_2}{2}\dot{c}_2 e^{-\sqrt{c_2}h}\left(c_2^{-\frac{3}{2}} + {c_2}^{-1}h\right) \\ &\quad - \frac{d}{dt}(\alpha\theta_3 M_2)\left(\frac{1}{\sqrt{c_1}}e^{-\sqrt{c_1}h} + \frac{1}{\sqrt{c_2}}e^{-\sqrt{c_2}h}\right) \\ &= O((|g| + he^{-dh})|R_2| + he^{-2dh}). \end{aligned}$$

Using the Cauchy–Schwarz inequality, we obtain (4.4). □

The following lemmas will be used to show the repulsiveness of solitary waves. Let $\beta = \sqrt{E(0)}$, a_i $(i = 1, 2)$, and a be positive numbers with $0 < a_1 < a_2 \le \sqrt{2c_1(0)}/50$ and $a \in [a_1, a_2]$. Set

$$\begin{aligned} T_0 &= \sup\{\tau \mid (4.2) \text{ and } (4.3) \text{ hold for } 0 \le t \le \tau\}, \\ T_1 &= \sup\left\{\tau \le T_0 \;\middle|\; \min_{i=1,2} c_i(t) \ge \tfrac{49}{50}c_1(0), \max_{i=1,2} c_i(t) \le \tfrac{51}{50}c_1(0) \text{ hold for } 0 \le t \le \tau\right\}, \\ T_2 &= \sup\left\{\tau \le T_1 \;\middle|\; \tfrac{8}{9}E(0) \le E(t) \le \tfrac{9}{8}E(0) \text{ holds for } 0 \le t \le \tau\right\}, \end{aligned}$$

where ε_2, δ, and A are positive constants to be fixed later.

LEMMA 4.2. *Let a, v_0, $x_{i,0}$, and $c_{i,0}$ $(i=1,2)$ satisfy the assumptions in Theorem 1.1. If ε_0 and ε_1 are sufficiently small, there exist positive numbers C_i $(i=1,2,3)$ such that, for $t \in [0, T_2]$,*

$$g(t) \le 2\beta, \tag{4.9}$$

$$h(t) \ge h(0) + \frac{9}{10}\beta t + C_1\varepsilon_0 \log\beta - C_2, \tag{4.10}$$

$$h(t) \le h(0) + 3\beta t + C_3. \tag{4.11}$$

Proof. The definitions of T_2 and $E(t)$ and the fact that $\alpha\theta_3 M_2 < 0$ imply (4.9).

Let $g_0 = c_{2,0} - c_{1,0}$. By (3.6) and (3.7), we have

$$\begin{aligned} &|g(0) - g_0| + |h(0) - h_0| = O(\varepsilon_0), \\ &e^{-d(0)h(0)} = e^{-d_0 h_0(1+O(\varepsilon_0))} \ge e^{-d_0 h_0(1+\eta)}, \quad \varepsilon_1^{\frac{1}{1-\eta}} \le \beta \le \varepsilon_1^{\frac{1}{1+\eta}}, \end{aligned} \tag{4.12}$$

where η is a nonnegative number with $\eta = O(\varepsilon_0)$.

First, we consider the case where $g(0) > -\frac{1}{2}\beta$. If β is sufficiently small, we get

$$\begin{aligned} \dot g &\ge \bar d\left(E(t) - \frac{1}{2}g(t)^2\right) + O(R_2) \\ &\ge \bar d\left(\frac{8}{9}E(0) - \frac{1}{2}g(t)^2\right) + O(\beta^{\frac{4}{3}(1+\gamma)} + \beta\|v_0\|_{H_a^1} + \|v_0\|_{H_a^1}^2) \\ &\ge \frac{\bar d}{2}(E(0) - g(t)^2), \end{aligned} \tag{4.13}$$

where $\bar d = \min_{0\le t\le T_2} d(t)$.

By (4.13),

$$g(t) \ge \beta\left(1 + \frac{2A_2 e^{-\beta\bar d t}}{1 - A_2 e^{-\beta\bar d t}}\right), \tag{4.14}$$

where $A_2 = \frac{g(0)-\beta}{g(0)+\beta}$. We remark that $-3 \le A_2 \le (\sqrt{2}-1)^2$ because $-\frac{\beta}{2} \le g(0) \le \sqrt{2}\beta$. Let κ be a positive constant such that

$$\min_{-3\le a\le(\sqrt{2}-1)^2} \frac{1 + ae^{-\bar d\kappa}}{1 - ae^{-\bar d\kappa}} > 9/10$$

and let $t_* = \kappa\beta^{-1}$.

Suppose $0 \le t \le t_*$. By (4.5) and (4.8),

$$\begin{aligned} h(t) &= h(0) + \int_0^t \left\{\left(1 + \alpha(\theta_1 M_1 + \theta_2 M_2)\frac{e^{-\sqrt{c_2}h} - e^{-\sqrt{c_1}h}}{c_2 - c_1}\right) g + O(R_2)\right\} ds \\ &= h(0) + \int_0^t \left(1 + O(he^{-dh})\right) g ds \\ &\quad + O\left(\beta^{\frac{4}{3}(1+\gamma)}t + \int_0^t (\beta\|v_0\|_{H_a^1} e^{-\delta bs} + \|v_0\|_{H_a^1}^2 e^{-2\delta bs}) ds\right) \\ &= h(0) + \int_0^t \left(1 + O(he^{-dh})\right) g ds + O\left(\beta^{\frac{4}{3}(1+\gamma)}t_* + \beta^2\right). \end{aligned} \tag{4.15}$$

By (4.14),

$$\int_0^t g(s)ds \geq \beta t + \frac{2}{\bar{d}} \log \left(\frac{1 - A_2 e^{-\beta \bar{d} t}}{1 - A_2} \right) \geq \beta t - \frac{4}{\bar{d}} \log 2. \tag{4.16}$$

Thus if $T_2 \leq t_*$, combining (4.9), (4.15), and (4.16) yields (4.10) and (4.11) for small β.

Suppose that $T_2 \geq t_*$. Then by (4.14),

$$g(t) \geq \frac{9}{10}\beta \quad \text{for } t_* \leq t \leq T_2. \tag{4.17}$$

Let $z(t) = e^{-\sqrt{c_1(0)}h(t)}$. Since $d(t)^2 \geq \frac{49}{50}c_1(0)$ and $a < \frac{\sqrt{2}}{50}\sqrt{c_1(0)}$, we get

$$\frac{4d - 2a}{3} \geq \nu_1 \sqrt{c_1(0)}, \quad d - a \geq \nu_2 \sqrt{c_1(0)}, \quad d \geq \nu_3 \sqrt{c_1(0)}, \tag{4.18}$$

where $\nu_1 = \frac{69}{75}\sqrt{2} > 1.3$, $\nu_2 = \frac{17}{25}\sqrt{2} > 0.96$, and $\nu_3 = \frac{7}{10}\sqrt{2}$. Substituting (4.7) and (4.18) into (4.6), we have

$$\begin{aligned} \dot{h} &= g\left(1 + O(e^{(a-d)h} + e^{-\delta b t}\|v_0\|_{H_a^1})\right) \\ &\quad + O(e^{-\frac{4d-2a}{3}h} + e^{-2\delta b t}\|v_0\|_{H_a^1}^2) \\ &= g(1 + O(e^{-\delta b t}\|v_0\|_{H_a^1} + z^{\nu_2})) + O(e^{-2\delta b t}\|v_0\|_{H_a^1}^2 + z^{\nu_1}). \end{aligned} \tag{4.19}$$

If β is sufficiently small, it follows from (4.17) and (4.19) that

$$\dot{z} = -\sqrt{c_1(0)}\dot{h}z \leq -\sqrt{c_1(0)}z\left(\frac{4}{5}\beta + O(z^{\nu_1} + e^{-2\delta b t}\|v_0\|_{H_a^1}^2)\right).$$

Since (4.10) holds for $t = t_*$ and $C_1 = 0$, it follows from the above that $z(t_*) \leq Ce^{-\sqrt{c_1(0)}h(0)} = O(\beta^2)$ and that $z(t)$ is a monotonically decreasing function satisfying

$$z(t) \leq C\beta^2 e^{-\frac{3}{5}\sqrt{c_1(0)}\beta(t - t_*)} \quad \text{for } t \in [t_*, T_2]. \tag{4.20}$$

By (4.17), (4.19), and (4.20), we obtain

$$\begin{aligned} h(t) &\geq h(t_*) + \frac{9}{10}\beta t + O\left(\int_{t_*}^t (e^{-\delta b s}\|v_0\|_{H_a^1} + z^{\nu_1} + \beta z^{\nu_2})ds\right) \\ &= h(t_*) + \frac{9}{10}\beta + O(\beta^{2\nu_1 - 1}) \end{aligned}$$

and

$$|h(t_*) - h(0)| \leq \int_0^{t_*} \left|\dot{h}(s)\right| ds = O(1).$$

Thus we have (4.10). Using (4.9), (4.19), and (4.20), we can show (4.11) in the same way.

Next, we consider the case where $g(0) \leq -\frac{\beta}{2}$. The assumptions of Theorem 1.1 and (4.12) imply that $g_0 \geq -\frac{\varepsilon_1}{2}$ and

$$\begin{aligned} g_0 &= g(0) + O(\|v_0\|_{H_a^1}) \\ &\leq -\frac{\beta}{2} + O(\beta^{2(1-\eta)}) < 0. \end{aligned}$$

Hence it follows from the above and (4.12) that $e^{-\frac{d_0 h_0}{2}} \geq \frac{\varepsilon_1}{2}$ and that $e^{-d(0)h(0)} \geq C\beta^{\frac{2(1+\eta)}{1-\eta}}$ for a $C > 0$.

Let $x(t) = g(t) + \sqrt{2E(t)}$. Then $x(t) \geq 0$ for $t \geq 0$, $x(0) < (\sqrt{2} - \frac{1}{2})\beta$ and

$$x(0) \geq C\frac{e^{-d(0)h(0)}}{\sqrt{2}\beta - g(0)} \geq A_3\beta^{\frac{1+3\eta}{1-\eta}} \tag{4.21}$$

for positive numbers C and A_3. Let

$$\bar{t}_1 = \sup\left\{0 \leq \tau \leq T_2 \;\middle|\; x(t) \geq A_3\beta^{\frac{1+3\eta}{1-\eta}}/2 \text{ for } 0 \leq t \leq \tau\right\}.$$

Making use of Lemma 4.1, (4.8), and the definition of T_2, we have

$$\begin{aligned}\dot{x} \geq& \bar{d}\left(E(t) - \frac{1}{2}g(t)^2\right) + O(R_2)\\ \geq& \frac{\bar{d}}{2}x\left(\frac{8}{3}\beta - x\right) + O(R_2)\\ \geq& \frac{\bar{d}}{2}x\left(\frac{5}{2}\beta - x\right)\end{aligned} \tag{4.22}$$

for a $t \in [0, \bar{t}_1]$. Hence it follows that

$$x(t) \geq \frac{5\beta A_4 e^{\frac{5}{4}\bar{d}\beta t}}{2(1 + A_4 e^{\frac{5}{4}\bar{d}\beta t})}, \tag{4.23}$$

where $A_4 = \frac{2x(0)}{5\beta - 2x(0)}$ and $\bar{t}_1 = T_2$. By (4.23) and the definition of T_2,

$$g(t) \geq x(t) - \frac{3}{2}\beta \geq \beta\left\{1 - \frac{5}{2(1 + A_4 e^{\frac{5}{4}\bar{d}\beta t})}\right\} \tag{4.24}$$

for $t \in [0, T_2]$. Let κ be a positive number with $24\beta^{\frac{5}{4}}\kappa\bar{d} = A_4$ and let $t_* = -\kappa\beta^{-1}\log\beta$. Then (4.24) implies that $g(t) \geq \frac{9}{10}\beta$ for $t \geq t_*$. By (4.21) and the definition of A_4, we have $\kappa = O(\eta)$ and

$$\begin{aligned}|h(t) - h(0)| \leq& \int_0^{t_*} |\dot{h}(s)|ds\\ \leq& \int_0^{t_*} (1 + O(he^{-dh}))|g(s)|ds + O(\beta^{\frac{4}{3}(1+\gamma)}t_*) = O(\eta\log\beta)\end{aligned}$$

for $t \in [0, t_*]$, and $z(t_*) \leq Ce^{-\sqrt{c_1(0)}h(t_*)} = O(\beta^{2-O(\eta)})$. Now, we can show (4.10) and (4.11) in the same manner as in the case where $g(0) > -\frac{\beta}{2}$. □

LEMMA 4.3. *Let a, v_0, $x_{i,0}$, and $c_{i,0}$ $(i = 1, 2)$ satisfy the assumptions in Lemma 4.2. Suppose $T_0 > t_*$. Then it holds that*

$$h(t) \geq h(t_*) + \frac{3}{5}\beta(t - t_*), \tag{4.25}$$

$$h(t) \leq h(t_*) + 4\beta(t - t_*), \tag{4.26}$$

$$g(t) \in \left[\frac{4}{5}\beta, 3\beta\right], \tag{4.27}$$

$$\frac{49}{50}c_1(0) \leq \min_{i=1,2} c_i(t) \leq \max_{i=1,2} c_i(t) \leq \tfrac{51}{50}c_1(0) \tag{4.28}$$

for any $t_* \le t \le T_0$.

Proof. To begin with, we show that $T_2 > t_*$ for small ε_1. Suppose the contrary. Then by Lemma 3.2, (4.8), the definition of t_*, and the fact that $e^{-\sqrt{c_i}h} = O(\beta^2)$,

$$\begin{aligned}|c_i(t) - c_i(0)| &\le C\int_0^t (e^{-dh} + R_2)ds \\ &\le C\int_0^t (\beta^2 + \beta\|v_0\|_{H_a^1}e^{-\delta bs} + \|v_0\|_{H_a^1}^2 e^{-2\delta bs})ds \\ &\le C(\kappa\beta|\log\beta| + \beta\|v_0\|_{H_a^1} + \|v_0\|_{H_a^1}^2) = O(\beta\log\beta)\end{aligned}$$

for $0 \le t \le T_2$. Thus we have

$$|c_i(T_2) - c_1(0)| \le |c_i(T_2) - c_i(0)| + |c_i(0) - c_1(0)| < \frac{1}{50}c_1(0)$$

for small β. By (4.4) and the definition T_2,

$$\left|\frac{dE}{dt}\right| \le C\left(\beta^{\frac{4}{3}\gamma+\frac{7}{3}} + \beta^2 e^{-\delta bt}\|v_0\|_{H_a^1} + \beta e^{-2\delta bt}\|v_0\|_{H_a^1}^2\right).$$

Hence

$$\begin{aligned}|E(t) - E(0)| &\le C\int_0^t (\beta^{\frac{4}{3}\gamma+\frac{7}{3}} + \beta^2 e^{-\delta bs}\|v_0\|_{H_a^1} + \beta e^{-2\delta bs}\|v_0\|_{H_a^1}^2)ds \\ &\le C(\kappa\beta^{\frac{4}{3}(1+\gamma)}|\log\beta| + \beta^2\|v_0\|_{H_a^1} + \beta\|v_0\|_{H_a^1}^2)\end{aligned}$$

for $0 \le t \le T_2$. Thus we have $|E(t) - E(0)| < \frac{1}{9}E(0)$ for small β, which is a contradiction.

Let

$$T_3 = \sup\{\tau \le T_0 \,|\, (4.27) \text{ and } (4.28) \text{ hold for } t \in [t_*, \tau]\}.$$

From the proof of Lemma 4.2, we see that $\frac{9}{10}\beta \le g(t_*) \le 2\beta$ and

$$z(t) \le C_1\beta^{2-C_2\eta}e^{-\frac{3}{5}\sqrt{c_1(0)}\beta(t-t_*)}$$

for $t \in [t_*, T_2]$, where C_1 and C_2 are positive constants. Using (4.5), (4.7), (4.18), and (4.20), we have

$$\begin{aligned}|g(t) - g(t_*)| &\le C\left(\int_{t_*}^t (e^{-dh} + R_2)ds\right) \\ &\le C\beta\left(\int_{t_*}^t \left(e^{-\delta bs}\|v_0\|_{H_a^1} + e^{(a-d)h}\right)ds\right) \\ &\quad + C\left(\int_{t_*}^t \left(e^{-2\delta bs}\|v_0\|_{H_a^1}^2 + e^{-dh}\right)ds\right) \\ &\le C(\beta\|v_0\|_{H_a^1} + \|v_0\|_{H_a^1}^2) + C\left(\beta\int_{t_*}^t z(s)^{\nu_2}ds + \int_{t_*}^t z(s)^{\nu_3}ds\right) \\ &= O(\beta^{2\nu_3 - 1 - O(\eta)})\end{aligned} \tag{4.29}$$

for $t_* \le t \le T_3$. Similarly, we have

$$|c_i(t) - c_i(0)| = O(\beta^{2\nu_2 - 1 - O(\eta)})$$

for $t_* \le t \le T_3$. Hence it follows from the standard contradiction argument that $T_3 = T_0$ if β is sufficiently small. Now we have (4.25) and (4.26), following the argument in the proof of Lemma 4.2. □

LEMMA 4.4. *Let a, v_0, $x_{i,0}$ and $c_{i,0}$ $(i = 1, 2)$ satisfy the assumptions in Lemma 4.2. If $T_0 \le t_*$, $T_0 = T_2$.*

Proof. The proof of this lemma is omitted, since it is similar to that of Lemma 4.3. □

Second, we will prove that $T_0 = \infty$. To begin with, we find the decay estimate of w, assuming the smallness of $\|v\|_\infty$. Let

$$T_4 = \sup\left\{\tau \,|\, \sup_{0\le t\le\tau} \|v(t)\|_\infty \le \varepsilon_2\right\}.$$

LEMMA 4.5. *Assume* (2.6). *Let v_0, $x_{i,0}$, and $c_{i,0}$ $(i = 1, 2)$ satisfy the assumptions in Theorem* 1.1. *Then for any $0 < \delta < 1$, there exist positive constants A and $\bar\varepsilon_i$ $(i = 0, 1, 2)$ such that if $\varepsilon_i \le \bar\varepsilon_i$ for $i = 0, 1, 2$ and $a \in [a_1, a_2]$,*

$$\|w(t)\|_{1,0} \le A(e^{-\delta b t}\|v_0\|_{H^1_a} + e^{(a-d(t))h(t)}) \quad \text{for } 0 \le t \le T_4.$$

Proof. To prove the lemma, we will show that $T_0 = T_4$ for small ε_1. By the continuity of $\|w(t)\|_{1,0}$ (see [16, 21]), we have $T_0 \ge 1$ for an appropriate A.

Since $h(t)$ grows up as t becomes large, we divide $[0, T_4]$ into a series of intervals $[t_{j-1}, t_j]$ $(j = 1, 2, \dots)$ to apply Lemma 2.6 as follows. Let $t_0 = 0$ and let ρ and $0 < t_1 < t_2 < \cdots \le T_4$ be positive numbers such that $h(t_j) = h(0) + j\rho$ and $h(t)$ is monotonically increasing for every $t \ge t_1$. We remark that Lemmas 4.2 and 4.3 imply $t_k - t_{k-1} \ge c\rho\beta^{-1}$ $(k \in \mathbb{N})$ for some $c > 0$. Let

$$\begin{aligned}
&\Pi_j(\mathbf{c}, h) = \Pi(\mathbf{c}, h, ; \mathbf{c}(t_j), h(t_j)),\\
&\bar{\mathcal{A}}_j(t) = \Pi_j(\mathbf{c}(t), h(t))\mathcal{A}_{\mathbf{c}(t),h(t)}\Pi_j(\mathbf{c}(t), h(t))^{-1},
\end{aligned}$$

and let $U_j(t, s)f$ be the solution of

$$\begin{cases} \partial_t w = \bar{\mathcal{A}}_j(t)w \quad \text{for } t_j \le s \le t \le t_{j+1},\\ w(s, \cdot) = f \in X(\mathbf{c}(t_j), h(t_j)). \end{cases}$$

Then Lemma 2.5 implies that there exists $B_1 > 0$ such that

(4.30)

$$\sup_{j\in\mathbb{N}\cup\{0\}} \sup_{t\in[t_j,t_{j+1}]} \left(\|\Pi_j(\mathbf{c}(t), h(t))\|_{\mathcal{L}(L^2)} + \left\|\Pi_j(\mathbf{c}(t), h(t))^{-1}\right\|_{\mathcal{L}(L^2)}\right) \le B_1,$$

$$\sup_{j\in\mathbb{N}\cup\{0\}} \sup_{t\in[t_j,t_{j+1}]} \left(\sum_{i=1,2} \left\|\frac{\partial\Pi_j}{\partial c_i}(\mathbf{c}(t), h(t))\right\|_{\mathcal{L}(L^2)} + \left\|\frac{\partial\Pi_j}{\partial h}(\mathbf{c}(t), h(t))\right\|_{\mathcal{L}(L^2)}\right) < \infty$$

for $t \in [t_j, t_{j+1}]$. By Lemma 2.6, we see that

$$\begin{aligned}
&\|U_j(t, s)f\| \le Me^{-b(t-s)}\|f\|,\\
&\|U_j(t, s)f\|_{1,0} \le M(t - s)^{-\frac{1}{2}}e^{-b(t-s)}\|f\|,\\
&\|U_j(t, s)f\|_{1,0} \le Me^{-b(t-s)}\|f\|_{1,0}
\end{aligned}$$

for any $t_{j-1} \le s \le t \le t_j$ and $f \in H^1(\mathbb{R})$, where M and b are positive constants independent of j.

Making use of (3.3), we can rewrite (3.2) as

$$\mathcal{Q}_{\mathbf{c}(t),h(t)}\partial_t w - \mathcal{A}_{\mathbf{c}(t),h(t)}w + \mathcal{Q}_{\mathbf{c}(t),h(t)}\left(l + (\partial_y - a)\mathcal{N}\right) = 0.$$

Hence we have

$$\partial_t w = \mathcal{A}_{\mathbf{c}(t),h(t)}w + \left(\frac{d}{dt}\mathcal{Q}_{\mathbf{c}(t),h(t)}\right)w - \mathcal{Q}_{\mathbf{c}(t),h(t)}\left(l + (\partial_y - a)\mathcal{N}\right).$$

Let $w_j(t) = \Pi_j(\mathbf{c}(t), h(t))w$. Then for $j \in \mathbb{N} \cup 0$,

$$\begin{cases} \partial_t w_j = \bar{\mathcal{A}}_j(t)w_j + k_{1,j} + k_{2,j} + k_{3,j}, \\ w_j(t_j) = w(t_j), \end{cases}$$

where

$$\begin{aligned} k_{1,j} =& \sum_{i=1,2} \dot{c}_i(t)\left(\Pi_j(\mathbf{c}(t),h(t))\frac{\partial \mathcal{Q}_{\mathbf{c}(t),h(t)}}{\partial c_i} + \frac{\partial \Pi_j}{\partial c_i}(\mathbf{c}(t),h(t))\right)\Pi_j(\mathbf{c}(t),h(t))^{-1}w_j \\ &+ \dot{h}\left(\Pi_j(\mathbf{c}(t),h(t))\frac{\partial \mathcal{Q}_{\mathbf{c}(t),h(t)}}{\partial h} + \frac{\partial \Pi_j}{\partial h}(\mathbf{c}(t),h(t))\right)\Pi_j(\mathbf{c}(t),h(t))^{-1}w_j, \\ k_{2,j} =& -\Pi_j(\mathbf{c}(t),h(t))\mathcal{Q}_{\mathbf{c}(t),h(t)}l, \\ k_{3,j} =& -\Pi_j(\mathbf{c}(t),h(t))\mathcal{Q}_{\mathbf{c}(t),h(t)}(\partial_y - a)\mathcal{N}. \end{aligned}$$

Using the variation of constants formula, we have

$$\begin{aligned} \|w_j(t)\|_{1,0} =& \left\| U_j(t,t_j)w(t_j) + \int_{t_j}^t U_j(t,s)(k_{1,j}(s) + k_{2,j}(s) + k_{3,j}(s))ds \right\|_{1,0} \\ \le& Me^{-b(t-t_j)}\|w(t_j)\|_{1,0} \\ &+ M\int_{t_j}^t (t-s)^{-\frac{1}{2}}e^{-b(t-s)}(\|k_{1,j}\| + \|k_{2,j}\| + \|k_{3,j}\|)ds. \end{aligned} \tag{4.31}$$

Lemmas 4.2–4.4 imply $|g(t)| \le 3\beta$ and $\sup_{t\in[0,T_0]} e^{-dt} \le C\beta^{2\nu_3}$ for a $C > 0$. Hence by Lemma 3.2,

$$\begin{aligned} |\dot{c}_1| + |\dot{x}_1 - c_1| &\le C(e^{-dh} + |\vec{\mathcal{R}}| + |\mathcal{R}_1| + |\mathcal{R}_2|) \\ &\le C(\beta + \|w\|_{1,0})\|w\|_{1,0} + Ce^{-dh}, \end{aligned}$$

$$\begin{aligned} |\dot{c}_2| + |\dot{x}_2 - c_2| &\le C(e^{-dh} + |\vec{\mathcal{R}}| + |\mathcal{R}_3| + |\mathcal{R}_4|) \\ &\le Ce^{-ah}(\beta + \|w\|_{1,0})\|w\|_{1,0} + Ce^{-dh}, \end{aligned}$$

and

$$|\dot{h}| \le |\dot{x}_1 - c_1| + |\dot{x}_2 - c_2| + |g|.$$

In view of (4.3) and (4.18), we have

$$\|w(t)\|_{1,0} \le A(\|v_0\|_{H^1_a} + e^{(a-d)h}) \le CA\beta \quad \text{for } 0 \le t \le T_0.$$

Hence it follows that

$$\begin{aligned}\|k_{1,j}\| &\le C(|\dot c_1| + |\dot c_2| + |\dot h|)\|w_j\| \\ &\le C\left(\beta + (\beta + \|w\|_{1,0})\|w\|_{1,0}\right)\|w_j\| \\ &\le C\left(1 + (1+A)A\beta\right)\beta\|w_j\|,\end{aligned}$$

$$\begin{aligned}\|k_{2,j}\| &\le C\left\{(1+\|w\|_{1,0})(|\dot x_1 - c_1| + |\dot c_1|) + e^{ah}(|\dot c_2| + |\dot x_2 - c_2|) + |g|\|w\|_{1,0}\right\} \\ &\le C(\beta + \|w\|_{1,0})\|w\|_{1,0} + Ce^{(a-d)h} \\ &\le C(1+A)(1+A\beta)\beta\|w_j\|_{1,0} + Ce^{(a-d)h},\end{aligned}$$

$$\begin{aligned}\|k_{3,j}\| &\le C(\|\mathcal{N}_1\|_{1,0} + \|\mathcal{N}_2\|_{1,0} + \|\mathcal{N}_3\|_{1,0}) \\ &\le C\left((1 + \|v\|_\infty^{p-2})\|v\|_\infty\|w\|_{1,0} + e^{(a-d)h} + e^{-dh}\|w\|_{1,0}\right) \\ &\le C(\varepsilon_2 + \beta)\|w_j\|_{1,0} + Ce^{(a-d)h}\end{aligned}$$

for $t \in [t_j, t_{j+1}]$. We remark that C does not depend on j in the above estimates. Substituting the above into (4.31), we have

$$\begin{aligned}\|w_j(t)\|_{1,0} \le{}& Me^{-b(t-t_j)}\|w(t_j)\|_{1,0} + \widetilde{C}(\varepsilon_2,\beta,A)\int_0^t (t-s)^{-\frac12}e^{-b(t-s)}\|w_j(s)\|_{1,0}ds \\ &+ C\int_0^t (t-s)^{-\frac12}e^{-b(t-s)}e^{(a-d(s))h(s)}ds,\end{aligned}$$

where $\widetilde{C}(\varepsilon_2,\beta,A)$ is a positive number satisfying $\lim_{\varepsilon_2,\beta\to 0}\widetilde{C} = 0$. Now we choose ε_2, β, and δ such that

$$\widetilde{C}(\varepsilon_2,\beta,A)\int_0^\infty s^{-\frac12}e^{-(1-\delta)s}ds \le \frac12, \quad 2MB_1e^{-(1-\delta)\frac{bc\rho}{\beta}} \le \frac12.$$

Since

$$e^{\delta bs}e^{(a-d(s))h(s)} \le e^{\delta bt}e^{(a-d(t))h(t)} \text{ for } t \ge s \ge 0, \tag{4.32}$$

$$\begin{aligned}e^{\delta b(t-t_j)}\|w_j(t)\|_{1,0} \le{}& Me^{-(1-\delta)b(t-t_j)}\|w_j(t_j)\|_{1,0} + Ce^{\delta bt}e^{(a-d(t))h(t)} \\ &+ \widetilde{C}(\varepsilon_2,\beta,A)\int_0^t (t-s)^{-\frac12}e^{-(1-\delta)b(t-s)}e^{\delta bs}\|w_j(s)\|_{1,0}ds \\ \le{}& Me^{-(1-\delta)b(t-t_j)}\|w_j(t_j)\|_{1,0} + Ce^{\delta b(t-t_j)}e^{(a-d(t))h(t)} \\ &+ \frac12\sup_{t_j\le s\le t} e^{\delta b(s-t_j)}\|w_j(s)\|_{1,0}.\end{aligned}$$

Thus we have

$$\|w_j(t)\|_{1,0} \le 2Me^{-b(t-t_j)}\|w(t_j)\|_{1,0} + B_2e^{(a-d(t))h(t)} \tag{4.33}$$

for $t \in [t_j, t_{j+1}]$, where B_2 is a constant which does not depend on j.

By (4.30), (4.32), and (4.33), we have

$$\begin{aligned}
&\|w(t_j)\|_{1,0}\\
&\le B_1\|w_{j-1}(t_j)\|_{1,0}\\
&\le B_1(2Me^{-b(t_j-t_{j-1})}\|w(t_{j-1})\|_{1,0}+B_2e^{(a-d(t_j))h(t_j)})\\
&\le (2MB_1)^je^{-bt_j}\|w(0)\|_{H^1_a}+\sum_{k=1}^{j}(2MB_1)^{j-k}B_1B_2e^{-b(t_j-t_k)}e^{(a-d(t_k))h(t_k)}\\
&\le (2MB_1)^je^{-bt_j}\|w(0)\|_{H^1_a}+\sum_{k=1}^{j}(2MB_1)^{j-k}B_1B_2e^{-(1-\delta)b(t_j-t_k)}e^{(a-d(t_j))h(t_j)}\\
&\le 2^{-j}e^{-\delta bt_j}\|w(0)\|_{H^1_a}+2e^{(a-d(t_j))h(t_j)}.
\end{aligned}$$

Combining the above with (4.33) and Remark 3.2, we have

$$\begin{aligned}
&\|w(t)\|_{1,0}\\
&\le B_1\|w_j(t)\|_{1,0}\\
&\le B_1(2Me^{-b(t-t_j)}\|w(t_j)\|_{1,0}+B_2e^{(a-d(t))h(t)})\\
&\le 2^{-j+1}MB_1e^{-\delta bt}\|w(0)\|+4MB_1e^{-b(t-t_j)}e^{(a-d(t_j)h(t_j)}+B_1B_2e^{(a-d(t))h(t)}\\
&\le 2^{-j+1}C_1MB_1e^{-\delta bt}\|v_0\|_{H^1_a}+(4M+B_2)B_1e^{(a-d(t))h(t)}.
\end{aligned}$$

Letting $A=((C_1+4)M+B_2)B_1$, we obtain $T_0=T_4$. Thus we have proved the lemma. □

COROLLARY 4.6. *Let $0<\delta<1$ and let v_0, $x_{i,0}$, and $c_{i,0}$ $(i=1,2)$ satisfy the assumptions in Theorem* 1.1. *Then if ε_i $(i=0,1,2)$ are sufficiently small,*

$$\|\partial_x w(t)\|_\infty \le C\left(t^{-\frac14}e^{-\delta bt}\|v_0\|_{H^1_a}+e^{(a-d(t))h(t)}\right) \quad \text{for } 0\le t\le T_4,$$

where C is a positive constant.

Proof. Since $e^{tA_0(c)}f=e^{(a^3-ac)t}e^{3at\partial_y^2}e^{t\left(-\partial_y^3+(c-3a^2)\partial_y\right)}$ and

$$\sup_{t>0}\left(t^{\frac14}\|e^{3at\partial_y^2}f\|_\infty+t^{\frac34}\|\partial_y e^{3at\partial_y^2}f\|_\infty\right)\le C\|f\|,$$

we have

$$\begin{aligned}
&\|e^{tA_0(c)}f\|_\infty\le Ct^{-\frac14}e^{(a^3-ac)t}\|f\|,\\
&\|\partial_y e^{tA_0(c)}f\|_\infty\le Ct^{-\frac34}e^{(a^3-ac)t}\|f\|,
\end{aligned}\tag{4.34}$$

where C is positive constants.

Using the variation of constants formula, (3.2) can be rewritten as

$$w(t)=e^{tA_0(c_1(0))}w(0)-\int_0^t e^{(t-s)A_0(c_1(0))}F(s)ds,$$

where

$$F(t)=(c_1(0)-c_1(t))(\partial_y-a)w+(\partial_y-a)\left\{\left(f'(\varphi_{c_1})+f'(\tau_h\varphi_{c_2})\right)w\right\}+e^{ay}\tilde{l}+(\partial_y-a)\mathcal{N}.$$

Then by Lemma 4.5, (3.14)–(3.18), and (4.34)

$$\begin{aligned}\|\partial_y w(t)\|_\infty &\le \left\|e^{tA_0(c_1(0))}\partial_y w(0)\right\|_\infty + \left\|\int_0^t \partial_y e^{(t-s)A_0(c_1(0))}F(s)ds\right\|_\infty \\ &\le Ct^{-\frac{1}{4}}e^{(a^3-ac)t}\|v_0\|_{H^1_a} + C\int_0^t (t-s)^{-\frac{3}{4}}e^{(a^3-ac)(t-s)}\,\|F(s)\|\,ds \\ &\le C\left(t^{-\frac{1}{4}}e^{-\delta bt}\|v_0\|_{H^1_a} + e^{(a-d)h}\right).\end{aligned}$$

Thus we complete the proof. □

4.2. L^∞-estimate of v. In this subsection, we show that the L^∞-norm of v remains small for every $t \ge 0$ and that the solution tends to two solitary waves plus a dispersive wave as $t \to \infty$.

For this purpose, we use the L^∞-estimate obtained by [19] and [20]. Let $\mathfrak{L}$, J, and I be operators defined by

$$\mathfrak{L}\phi = \partial_t\phi + \partial_x^3\phi, \quad J\phi = x\phi - 3t\partial_x^2\phi, \quad I\phi = x\phi + 3t\int_{-\infty}^x \partial_t\phi(t,x')dx'$$

for $\phi(t,x) \in C_0^\infty(\mathbb{R}^2)$, and let

$$\begin{aligned}\mathcal{M}[u](t) &:= \|u(\cdot,t)\|_{1,0} + \|\partial_x Ju(\cdot,t)\| + \|D^\alpha Ju(\cdot,t)\|\,, \\ \widetilde{\mathcal{M}}[u](t) &:= \left(\|u(\cdot,t)\|_{1,0} + \|Ju(\cdot,t)\|_{1,0}\right)\langle t\rangle^{-\frac{1}{6}} + \|\mathcal{F}U_0(-t)u(t)\|_\infty\,,\end{aligned}$$

where $U_0(t) = e^{-t\partial_x^3}$. We remark that

$$[\mathfrak{L}, J] = 0, \quad [\mathfrak{L}, I] = 3\int_{-\infty}^x \mathfrak{L}\phi dx', \quad [I, \partial_x]\phi = [J, \partial_x]\phi = -\phi \tag{4.35}$$

and that

$$(I-J)\phi = 3t\int_{-\infty}^x \mathfrak{L}\phi(t,x')dx'. \tag{4.36}$$

LEMMA 4.7 (see [19, 20]). *Let $T > 0$, $\alpha \in [0, 1/2)$, and $q \in (4, \infty]$ be constants.*

(i) *Suppose that $u(t,x)$ is a function satisfying $\sup_{0\le t\le T}\mathcal{M}[u](t) < \infty$. Then*

$$\begin{aligned}\|u\|_q &\le C(1+t)^{-\frac{1}{3}+\frac{1}{3q}}\mathcal{M}[u](t), \\ \|uu_x\|_\infty &\le Ct^{-\frac{2}{3}}(1+t)^{-\frac{1}{3}}\mathcal{M}[u]^2(t),\end{aligned}$$

where C is a constant which does not depend on u, t, and T.

(ii) *Suppose that $u(t,x)$ is a function satisfying $\sup_{0\le t\le T}\widetilde{\mathcal{M}}[u](t) < \infty$. Then*

$$\begin{aligned}\|u\|_q &\le C(1+t)^{-\frac{1}{3}+\frac{1}{3q}}\widetilde{\mathcal{M}}[u](t), \\ \|uu_x\|_\infty &\le Ct^{-\frac{2}{3}}(1+t)^{-\frac{1}{3}}\widetilde{\mathcal{M}}[u]^2(t),\end{aligned}$$

where C is a constant which does not depend on u, t, and T.

Let $z_i = x - x_i(t) + x_1(0)$ for $i = 1, 2$ and

$$\tilde{v}(x,t) = u(x + x_1(0), t) - \varphi_{c_1}(z_1) - \varphi_{c_2}(z_2). \tag{4.37}$$

Then

$$\mathfrak{L}\tilde{v} + \partial_x \bar{N} = G, \tag{4.38}$$

where

$$\bar{N} = f(\varphi_{c_1}(z_1) + \varphi_{c_2}(z_2) + \tilde{v}) - f(\varphi_{c_1}(z_1)) - f(\varphi_{c_2}(z_2)),$$
$$G = \sum_{i=1,2} \left\{ (\dot{x}_i - c_i)\partial_x \varphi_{c_i}(z_i) - \dot{c}_i \partial_c \varphi_{c_i}(z_i) \right\}.$$

To prove Theorem 1.1, we need the following proposition.

PROPOSITION 4.8. *Assume* (2.6). *Let* $3 < p < 5$ *and let* v_0, $c_{i,0}$, *and* $x_{i,0}$ $(i = 1, 2)$ *satisfy the assumptions in Theorem* 1.1. *Let* $0 < \alpha < 1/2 - \gamma$ *with* $0 < \gamma < (p-3)/3$ *if* $3 < p \le 7/2$ *and let* $\alpha = 0$ *if* $p > 7/2$. *Then there exist positive numbers* $\bar{\varepsilon}_1$, η, B, *and* C *such that if* $\varepsilon_1 \le \bar{\varepsilon}_1$ *and* $\varepsilon_0 \le C\varepsilon_1^2$,

$$\sup_{t \ge 0} \mathcal{M}[\tilde{v}] \le B(\varepsilon_0 + \varepsilon_1^{\eta}).$$

PROPOSITION 4.9. *Assume* (2.6). *Let* $p = 3$ *and let* v_0, $c_{i,0}$, *and* $x_{i,0}$ $(i = 1, 2)$ *satisfy the assumptions in Theorem* 1.1. *Then there exist positive numbers* $\bar{\varepsilon}_1$, η, B, *and* C *such that if* $\varepsilon_1 \le \bar{\varepsilon}_1$ *and* $\varepsilon_0 \le C\varepsilon_1^2$,

$$\sup_{t \ge 0} \widetilde{\mathcal{M}}[\tilde{v}] \le B(\varepsilon_0 + \varepsilon_1^{\eta}).$$

Proof of Propositions 4.8 and 4.9. Let $T_5 = \sup\left\{ \bar{t} \in [0, T_4] \;\middle|\; \mathcal{M}[\tilde{v}](t) \le \varepsilon \right\}$ and let ε be a small number such that Lemma 4.7 implies

$$\|v(t)\|_\infty = \|\tilde{v}\|_\infty \le C\varepsilon \le \varepsilon_2 \quad \text{for } t \in [0, T_5].$$

By the continuity of v, we have $T_5 \ge 1$ if ε_0 is sufficiently small.

Now, we estimate each term of $\mathcal{M}[\tilde{v}]$. By Lemmas 3.2 and 4.1–4.5, (3.20), (4.18), and (4.20),

$$\|v\|_{H^1_{\bar{a}}} \le C(e^{-\delta b t}\|v_0\|_{H^1_{\bar{a}}} + e^{(\bar{a}-d)h})$$

for $\bar{a} \in [a_1, a_2]$, and there exists a positive number C_1 such that

$$\begin{aligned} |\dot{x}_i - c_i| + |\dot{c}_i| &\le C\left(e^{-dh} + \mathcal{R}_1 + \mathcal{R}_2 + \mathcal{R}_3 + \mathcal{R}_4\right) \\ &\le C\left(e^{-\delta b t}\|v_0\|_{H^1_{\bar{a}}} + \beta^{2\tilde{\nu}_3} e^{-C_1 \beta t}\right) \end{aligned} \tag{4.39}$$

for $i = 1, 2$ and

$$|x_2 - \dot{c}_2| + |\dot{c}_2| \le C\left(e^{-dh} + \mathcal{R}_3 + \mathcal{R}_4\right) \le Ce^{-\bar{a}h}\left(e^{-\delta b t}\|v_0\|_{H^1_{\bar{a}}} + \beta^{2\tilde{\nu}_2} e^{-C_1 \beta t}\right), \tag{4.40}$$

where $\tilde{\nu}_i = \nu_i - C_i\varepsilon_0$ for nonnegative C_i $(i = 1, 2)$. We remark that Lemmas 4.3–4.5 imply $d(t) \ge \nu_3\sqrt{c_1(0)}$ for $0 \le t \le T_4$.

Let

$$\begin{aligned}
&\bar{N}_1 = \bar{N}_{1,1} + \bar{N}_{1,2},\\
&\bar{N}_2 = f(\varphi_{c_1}(z_1) + \varphi_{c_2}(z_2)) - f(\varphi_{c_1}(z_1)) - f(\varphi_{c_2}(z_2)),\\
&\bar{N}_{1,1} = f(\varphi_{c_1}(z_1) + \varphi_{c_2}(z_2) + \tilde{v}) - f(\varphi_{c_1}(z_1) + \varphi_{c_2}(z_2)) - f(\tilde{v}),\\
&\bar{N}_{1,2} = f(\tilde{v}).
\end{aligned}$$

Note that

$$\bar{N}_{1,1} = \int_0^1 \int_0^1 f''\left(\theta_1(\varphi_{c_1} + \varphi_{c_2} + \theta_2\tilde{v})\right) d\theta_1 d\theta_2 (\varphi_{c_1} + \varphi_{c_2})\tilde{v}.$$

By Lemma 4.5 and (1.3), we get

$$\begin{aligned}
\|\bar{N}_{1,1}\|_{1,0} &\le C \sum_{0\le i+j\le 1} \|\partial_x^i(\varphi_{c_1}(z_1) + \varphi_{c_2}(z_2))\partial_x^j\tilde{v}\|_{1,0}\\
&\le C\left(e^{-\delta bt}\|v_0\|_{H_a^1} + \beta^{2\tilde{\nu}_2}e^{-C_1\beta t}\right)
\end{aligned} \tag{4.41}$$

and

$$\|\bar{N}_2\|_{2,0} \le C\|\varphi_{c_1}(z_1)\varphi_{c_2}(z_2)\|_{2,\infty} \le Che^{-dh} \le C\beta^{2\tilde{\nu}_3}e^{-C_1\beta t}, \tag{4.42}$$

where C_1 is a positive constant. By Lemma 4.5 and Corollary 4.6,

$$\begin{aligned}
\left|\left\langle \partial_x^2\bar{N}_{1,1}, \tilde{v}_x\right\rangle\right| \le\ & \left|\langle\{f'(\varphi_{c_1} + \varphi_{c_2} + \tilde{v}) - f'(\tilde{v})\}\tilde{v}_{xx}, \tilde{v}_x\rangle\right|\\
&+ C(1 + \|w\|_{1,\infty})\|w\|_{1,0}\|\tilde{v}_x\|\\
&\le C(1 + t^{-\frac{1}{4}})\left(e^{-\delta bt}\|v_0\|_{H_a^1} + \beta^{2\tilde{\nu}_2}e^{-C_1\beta t}\right)\|\tilde{v}_x\|.
\end{aligned} \tag{4.43}$$

By Lemma 4.7,

$$\begin{aligned}
&\|\partial_x\bar{N}_{1,2}\| \le C\|\tilde{v}\tilde{v}_x\|_\infty\|\tilde{v}\|_\infty^{p-3}\|\tilde{v}\| \le Ct^{-\frac{2}{3}}\langle t\rangle^{-\frac{p-2}{3}}\mathcal{M}[\tilde{v}]^{p-1}\|\tilde{v}\|,\\
&\left|\left\langle \partial_x^2\bar{N}_{1,2}, \tilde{v}_x\right\rangle\right| \le C\|\tilde{v}\tilde{v}_x\|_\infty\|\tilde{v}\|_\infty^{p-3}\|\tilde{v}_x\|^2 \le C\mathcal{M}[\tilde{v}]^{p-1}t^{-\frac{2}{3}}\langle t\rangle^{-\frac{p-2}{3}}\|\tilde{v}_x\|^2.
\end{aligned} \tag{4.44}$$

Multiplying (4.38) by $2\tilde{v}$ and integrating the resulting equation, we have

$$\begin{aligned}
\frac{d}{dt}\|\tilde{v}\|^2 &= 2\sum_{i=1,2}\{(\dot{x}_i - c_i)\langle\partial_x\varphi_{c_i}, \tilde{v}\rangle - \dot{c}_i\langle\partial_c\varphi_{c_i}, \tilde{v}\rangle\} - 2(\partial_x\bar{N}, \tilde{v})\\
&\le C\left(e^{-\delta bt}\|v_0\|_{H_a^1} + \beta^{2\tilde{\nu}_2}e^{-C_1\beta t}\right)\|\tilde{v}\| + Ct^{-\frac{2}{3}}\langle t\rangle^{-\frac{p-2}{3}}\mathcal{M}[\tilde{v}]^{p-1}\|\tilde{v}\|^2.
\end{aligned}$$

Differentiating (4.38) with respect to x, multiplying by $\partial_x\tilde{v}$, and integrating the resulting equation by parts, we have

$$\begin{aligned}
\frac{d}{dt}\|\tilde{v}_x\|^2 &= 2\sum_{i=1,2}\left\{(\dot{x}_i - c_i)\langle\partial_x^2\varphi_{c_i}, \tilde{v}_x\rangle - \dot{c}_i\langle\partial_x\partial_c\varphi_{c_i}, \tilde{v}_x\rangle\right\} - 2(\partial_x^2\bar{N}, \tilde{v}_x)\\
&\le C(1 + t^{-\frac{1}{4}})\left(e^{-\delta bt}\|v_0\|_{H_a^1} + \beta^{2\tilde{\nu}_2}e^{-C_1\beta t}\right)\|\tilde{v}_x\|\\
&\quad + C\mathcal{M}[\tilde{v}]^{p-1}t^{-\frac{2}{3}}\langle t\rangle^{-\frac{p-2}{3}}\|\tilde{v}_x\|^2.
\end{aligned}$$

Hence, it follows that

$$\|\tilde{v}(t)\|_{1,0} \le \|\tilde{v}(0)\|_{1,0} + C\left(\|v_0\|_{H^1_a} + \beta^{2\tilde{\nu}_2}\right) + C\sup_{0\le s\le t}\mathcal{M}[\tilde{v}]^{p-1}\int_0^t s^{-\frac{2}{3}}\langle s\rangle^{-\frac{p-2}{3}}\|\tilde{v}(s)\|_{1,0}ds.$$

Thus, by Gronwall's inequality, we have

$$\|\tilde{v}(t)\|_{1,0} \le C(\|v_0\|_{1,0} + \|v_0\|_{H^1_a} + \beta^{2\tilde{\nu}_2-1}) \tag{4.45}$$

if $p > 3$ and

$$\|\tilde{v}(t)\|_{1,0} \le C(\|v_0\|_{1,0} + \|v_0\|_{H^1_a} + \beta^{2\tilde{\nu}_2-1})\langle t\rangle^{C'\varepsilon^2} \tag{4.46}$$

for a $C' > 0$ if $p = 3$.

Second, we will estimate $\|\partial_x J\tilde{v}\|$ and $\|J\tilde{v}\|$. Let us start by considering the case where $7/2 < p < 5$, that is, the case where $\|f'(\tilde{v})\tilde{v}\| = \|\tilde{v}\|_{2p}^p = O(\langle t\rangle^{-\frac{p}{3}+\frac{1}{6}}) \in L^1(0,\infty)$. By (4.35), (4.36), and (4.38), we have

$$\mathfrak{L}J\tilde{v} = J(G - \partial_x\bar{N}) = JG - I\partial_x\bar{N} + 3t\mathfrak{L}\bar{N},$$
$$I\partial_x\tilde{v} = J\partial_x\tilde{v} + 3t\mathfrak{L}\tilde{v} = \partial_x(J\tilde{v} - 3t\bar{N}) + 3tG - \tilde{v}.$$

Let $\Phi = J\tilde{v} - 3t\bar{N}$. Then Φ satisfies

$$\mathfrak{L}\Phi = K_1 + K_2 + K_3 + K_4 + K_5, \tag{4.47}$$

where

$$\begin{aligned}
K_1 &= -f'(\varphi_{c_1} + \varphi_{c_2} + \tilde{v})\partial_x\Phi,\\
K_2 &= f'(\varphi_{c_1} + \varphi_{c_2} + \tilde{v})\tilde{v} - 3\bar{N}_1,\\
K_3 &= \{(f'(\varphi_{c_1} + \varphi_{c_2}) - f'(\varphi_{c_1} + \varphi_{c_2} + \tilde{v})\}I\partial_x(\varphi_{c_1} + \varphi_{c_2}),\\
K_4 &= -(I\partial_x + 3)\bar{N}_2,\\
K_5 &= (J - 3tf'(\varphi_{c_1} + \varphi_{c_2} + \tilde{v}))G.
\end{aligned}$$

Multiplying (4.47) by Φ and integrating the resulting equation by parts, we obtain

$$\begin{aligned}
\frac{d}{dt}\|\Phi\|^2 &= 2(\mathfrak{L}\Phi, \Phi)\\
&\le 2\sum_{i=1,2}|\langle K_i, \Phi\rangle| + 2\sum_{3\le i\le 5}\|e^{-az_1}K_i\|\|e^{az_1}\Phi\|.
\end{aligned} \tag{4.48}$$

Using Lemmas 4.7 and 4.5 and Corollary 4.6, we have

$$\begin{aligned}
&|\langle K_1, \Phi\rangle|\\
&\le \int_{\mathbb{R}}|\partial_x f'(\varphi_{c_1} + \varphi_{c_2} + \tilde{v})|\,\Phi^2 dx\\
&\le C(\|(\varphi_{c_1} + \varphi_{c_2})_x\Phi^2\|_1 + \|(\varphi_{c_1} + \varphi_{c_2})\tilde{v}_x\Phi^2\|_1 + \|\tilde{v}\tilde{v}_x\|_\infty\|\tilde{v}\|^{p-3}\|\Phi\|^2)\\
&\le C(1 + t^{-\frac{1}{4}})\|e^{az_1}\Phi\|^2 + C\mathcal{M}[\tilde{v}]^{p-1}t^{-\frac{2}{3}}\langle t\rangle^{-\frac{p-2}{3}}\|\Phi\|^2
\end{aligned} \tag{4.49}$$

and

$$\begin{aligned}\|K_2\| &\le C(\|w\| + \|\tilde{v}\|_{2p}^p)\\ &\le C(e^{-\delta bt}\|v_0\|_{H_a^1} + \beta^{2\tilde{\nu}_2}e^{-C_1\beta t} + \mathcal{M}[\tilde{v}]^p\langle t\rangle^{-\frac{2p-1}{6}}).\end{aligned} \tag{4.50}$$

Noting that

$$I\partial_x\varphi_{c_i}(z_i) = (x - 3t\dot{x}_i)\partial_x\varphi_{c_i}(z_i) + 3t\dot{c}_i\partial_c\varphi_{c_i}(z_i),$$

we have

$$\begin{aligned}\|e^{-az_1}K_3\|_{1,0} &\le C\sum_{0\le i+j\le 1}\|e^{-az_1}(\partial_x^i\tilde{v})\partial_x^j I\partial_x(\varphi_{c_1}+\varphi_{c_2})\|\\ &\le C(\langle t\rangle + (t+h)e^{-ah})\|w\|_{1,0}\\ &\le C\langle t\rangle\left(e^{-\delta bt}\|v_0\|_{H_a^1} + \beta^{2\tilde{\nu}_2}e^{-C_1\beta t}\right),\end{aligned} \tag{4.51}$$

(4.52)

$$\begin{aligned}\|e^{-az_1}K_4\|_{1,0} &\le C\sum_{i=1,2}\left\|e^{-az_1}\{f'(\varphi_{c_1}+\varphi_{c_2}) - f'(\varphi_{c_i})\}I\partial_x\varphi_{c_i}\right\|_{1,0} + C\|e^{-az_1}\varphi_{c_1}\varphi_{c_2}\|_{1,0}\\ &\le C\langle t\rangle\beta^{2\tilde{\nu}_3}e^{-C_1\beta t}.\end{aligned}$$

By (4.39) and (4.40),

$$\begin{aligned}\|e^{-az_1}K_5\|_{1,0} &\le C\langle t\rangle\sum_{i=1,2}(|\dot{x}_i - c_i| + |\dot{c}_i|)\\ &\le C\langle t\rangle(e^{-\delta bt}\|v_0\|_{H_a^1} + \beta^{2\tilde{\nu}_2}e^{-C_1\beta t})\\ &\le C\langle t\rangle\left(e^{-\delta bt}(\|v_0\|_{H_{a_1}^1} + \|v_0\|_{H_{a_2}^1}) + \beta^{2\tilde{\nu}_2}e^{-C_1\beta t}\right).\end{aligned} \tag{4.53}$$

Multiplying (4.47) by Φ_{xx} and integrating the resulting equation by parts, we have

$$\begin{aligned}\frac{d}{dt}\|\Phi_x\|^2 &= 2(\mathfrak{L}\Phi_x, \Phi_x)\\ &\le 2\sum_{i=1,2}|\langle\partial_x K_i, \Phi_x\rangle| + 2\sum_{3\le i\le 5}\|e^{-az_1}\partial_x K_i\|\|e^{az_1}\Phi_x\|.\end{aligned} \tag{4.54}$$

Using Lemmas 4.7 and 4.5 and Corollary 4.6, we have

(4.55)

$$\begin{aligned}|\langle\partial_x K_1, \Phi_x\rangle| &\le \int_{\mathbb{R}}|\partial_x f'(\varphi_{c_1}+\varphi_{c_2}+\tilde{v})|\,\Phi_x^2 dx\\ &\le C(\|(\varphi_{c_1}+\varphi_{c_2})_x\Phi_x^2\|_1 + \|(\varphi_{c_1}+\varphi_{c_2})\tilde{v}_x\Phi_x^2\|_1 + \|\tilde{v}\tilde{v}_x\|_\infty\|\tilde{v}\|^{p-3}\|\Phi_x\|^2)\\ &\le C(1+t^{-\frac{1}{4}})\|e^{az_1}\Phi_x\|^2 + C\mathcal{M}[\tilde{v}]^{p-1}t^{-\frac{2}{3}}\langle t\rangle^{-\frac{p-2}{3}}\|\Phi_x\|^2\end{aligned}$$

and

$$\begin{aligned}\|\partial_x K_2\| &\le C(\|w\|_{1,0} + \|\tilde{v}\|_\infty^{p-3}\|\tilde{v}\tilde{v}_x\|_\infty\|\tilde{v}\|)\\ &\le C(e^{-\delta bt}\|v_0\|_{H_a^1} + \beta^{2\tilde{\nu}_2}e^{-C_1\beta t} + \mathcal{M}[\tilde{v}]^{p-1}t^{-\frac{2}{3}}\langle t\rangle^{-\frac{p-2}{3}}\|\tilde{v}\|).\end{aligned} \tag{4.56}$$

Next we will obtain the decay rate of $\|e^{az_1}\Phi\|_{1,0}$. Let $\widetilde{\Psi}(z_1,t)=e^{az_1}\Phi_x$. Then by (4.47),

$$\begin{aligned}\partial_t\widetilde{\Psi}&=\mathcal{A}_{\mathbf{c}(t),h(t)}\widetilde{\Psi}+(\partial_y-a)(k\widetilde{\Psi})+(\partial_{z_1}-a)(f'(\varphi_{c_1}+\varphi_{c_2}+\tilde{v})w\\&\quad+e^{az_1}\partial_x(K_3+K_4+K_5),\end{aligned}\tag{4.57}$$

where

$$k=\dot{x}-c_1-g\chi_2-f'(\varphi_{c_1}+\varphi_{c_2}+\tilde{v})+f'(\varphi_{c_1})+f'(\varphi_{c_2}).$$

We decompose $\widetilde{\Psi}$ into two parts. Let $\widetilde{\Psi}=\widetilde{\Psi}_{\|}+\widetilde{\Psi}_{\perp}$ and

$$\int_{\mathbb{R}}\widetilde{\Psi}_{\perp}(t,z_1)\bar{\eta}_i(z_1,\mathbf{c}(t),h(t))dz_1=0\quad\text{for } t\geq 0 \text{ and } 1\leq i\leq 4.$$

To start with, we will estimate $\|\widetilde{\Psi}_{\|}\|_{1,0}$.

$$\begin{aligned}\|\widetilde{\Psi}_{\|}\|_{1,0}&\leq\sum_{1\leq i\leq 4}\left|\left\langle(x\tilde{v})_x-3t\partial_x^3\tilde{v},e^{az_1}\bar{\eta}_i\right\rangle\right|\left\|\bar{\xi}_i\right\|_{1,0}\\&\leq C\sum_{1\leq i\leq 4}\left(|\langle e^{az_1}\tilde{v},x(\partial_x+a)\bar{\eta}_i(z_1)\rangle|+t\left|\left\langle e^{az_1}\tilde{v},(\partial_x+a)^3\bar{\eta}_i\right\rangle\right|\right)\\&\leq C\langle t\rangle\left(e^{-\delta bt}(\|v_0\|_{H^1_{a_1}}+\|v_0\|_{H^1_{a_2}})+\beta^{2\tilde{\nu}_2}e^{-C_1\beta t}\right).\end{aligned}\tag{4.58}$$

Next, we will obtain a decay estimate of $\|\widetilde{\Psi}_{\perp}\|$. By a simple computation, we have

$$\|\widetilde{\Psi}(0)\|=\|e^{az_1}(x\tilde{v}(0))_x\|\leq C(\|v_0\|_{H^1_{a_1}}+\|v_0\|_{H^1_{a_2}}),$$

$$\begin{aligned}\|e^{az_1}K_3\|&\leq C\langle t\rangle\|e^{a_2z_1}\tilde{v}\|\|e^{(a-a_2)z_1}I\partial_x(\varphi_{c_1}+\varphi_{c_2})\|_{\infty}\\&\leq C\langle t\rangle\left(e^{-\delta bt}\|v_0\|_{H^1_{a_2}}+\beta^{2\tilde{\nu}_2}e^{-C_1\beta t}\right),\end{aligned}\tag{4.59}$$

and

$$\|e^{az_1}K_4\|\leq C\langle t\rangle e^{\theta(a-d)h}\leq C\beta^{2\theta\tilde{\nu}_2-1}e^{-C_1\beta t}\tag{4.60}$$

for any $\theta<1$. By (4.39) and (4.40),

$$\begin{aligned}\|e^{az_1}K_5\|&\leq C\langle t\rangle(|\dot{x}_1-c_1|+|\dot{c}_1|)+C(t+h)e^{ah}(|\dot{x}_2-c_2|+|\dot{c}_2|)\\&\leq C\langle t\rangle(e^{-\delta bt}\|v_0\|_{H^1_a}+\beta^{2\tilde{\nu}_3}e^{-C_1\beta t})\\&\quad+C(t+h)e^{(a-a_2)h}(e^{-\delta bt}\|v_0\|_{H^1_{a_2}}+\beta^{2\tilde{\nu}_2}e^{-C_1\beta t})\\&\leq C\left(\langle t\rangle e^{-\delta bt}(\|v_0\|_{H^1_{a_1}}+\|v_0\|_{H^1_{a_2}})+\beta^{2\tilde{\nu}_2-1}e^{-C_1\beta t}\right).\end{aligned}\tag{4.61}$$

Since

$$\begin{aligned}\|k(t)\|_{\infty}&\leq C(|\dot{x}_1-c_1|+|g|+|\tilde{v}|_{\infty}+\|\varphi_{c_1}(z_1)\varphi_{c_2}(z_2)\|_{\infty})\\&=O(\|v_0\|_{H^1_a}+\varepsilon+\beta^{2\tilde{\nu}_3}),\end{aligned}$$

we can use the variation of constants formula and apply Lemma 2.6 as in the proof of Lemma 4.5 to obtain

$$\left\|\widetilde{\Psi}_{\perp}(t)\right\| \le C\left(\langle t\rangle e^{-\delta bt}(\|v_0\|_{H^1_{a_1}}+\|v_0\|_{H^1_{a_2}})+\beta^{2\theta\tilde{\nu}_2-1}e^{-C_1\beta t}\right) \tag{4.62}$$

for any $\theta\in(0,1)$ if $\|v_0\|_{H^1_a}$, ε, and β are sufficiently small.

Let $\Psi(z_1,t)=e^{az_1}\Phi$. Then (4.47) is transformed into

$$\partial_t\Psi-\dot{x}_1(\partial_{z_1}-a)\Psi+(\partial_{z_1}-a)^3\Psi=\sum_{1\le i\le 5}e^{az_1}K_i.$$

Making use of the variation of constants formula, we have

(4.63)

$$\begin{aligned}\|\Psi(t)\| &\le \left\|e^{tA_0(c_1(0))}\Psi(0)\right\|+\sup_{0\le s\le t}|\dot{x}_1(s)-c_1(0)|\left\|\int_0^t e^{(t-s)A_0(c_1(0))}(\partial_{z_1}-a)\Psi ds\right\|\\ &\quad+\left\|\int_0^t e^{(t-s)A_0(c_1(0))}e^{az_1}\left(K_1+K_2+K_3+K_4+K_5\right)\right\|\\ &\le Ce^{-b''t}\|\Psi(0)\|+C\sup_{0\le s\le t}|\dot{x}_1(s)-c_1(0)|\int_0^t(t-s)^{-\frac{1}{2}}e^{-b''(t-s)}\|\Psi\|ds\\ &\quad+C\sum_{1\le i\le 5}\int_0^t e^{-b''(t-s)}\|e^{az_1}K_i\|ds,\end{aligned}$$

where $b''=a^3-ac_1(0)$.

By (4.58) and (4.62),

$$\|e^{az_1}K_1\|\le C\|\widetilde{\Psi}\|\le C\left(\langle t\rangle e^{-\delta bt}(\|v_0\|_{H^1_{a_1}}+\|v_0\|_{H^1_{a_2}})+\beta^{2\theta\tilde{\nu}_2-1}e^{-C_1\beta t}\right). \tag{4.64}$$

By Lemma 4.5,

$$\|e^{az_1}K_2\|\le C\|w\|\le C\left(e^{-\delta bt}\|v_0\|_{H^1_a}+\beta^{2\tilde{\nu}_2}e^{-C_1\beta t}\right). \tag{4.65}$$

Combining (4.58)–(4.65) with $\|\Psi(0)\|\le C(\|v_0\|_{H^1_{a_1}}+\|v_0\|_{H^1_{a_2}})$ and taking $b>0$ smaller if necessary, we obtain

$$\|\Psi(t)\|\le C\left(\langle t\rangle e^{-\delta bt}(\|v_0\|_{H^1_{a_1}}+\|v_0\|_{H^1_{a_2}})+\beta^{2\theta\tilde{\nu}_2-1}e^{-C_1\beta t}\right) \tag{4.66}$$

for sufficiently small ε_i $(i=2,3,4)$.

By (4.48)–(4.56), (4.58), (4.62), and (4.66), we obtain

$$\begin{aligned}\frac{d}{dt}\|\Phi\|^2 &\le C(\varepsilon^{p-1}t^{-\frac{2}{3}}\langle t\rangle^{-\frac{p-2}{3}}\|\Phi\|+\varepsilon^p\langle t\rangle^{-\frac{2p-1}{6}})\|\Phi\|\\ &\quad+C(e^{-\delta bt}\|v_0\|_{H^1_a}+\beta^{2\tilde{\nu}_2}e^{-C_1\beta t})\|\Phi\|\\ &\quad+C(1+t^{-\frac{1}{4}})\left(\langle t\rangle e^{-\delta bt}(\|v_0\|_{H^1_{a_1}}+\|v_0\|_{H^1_{a_2}})+\beta^{2\theta\tilde{\nu}_2-1}e^{-C_1\beta t}\right)^2\end{aligned}$$

and

$$\begin{aligned}\frac{d}{dt}\|\Phi_x\|^2 &\le Ct^{-\frac{2}{3}}\langle t\rangle^{-\frac{p-2}{3}}\varepsilon^{p-1}(\|\Phi_x\|+\|\tilde{v}\|)\|\Phi_x\|\\ &\quad+C(e^{-\delta bt}\|v_0\|_{H^1_a}+\beta^{2\tilde{\nu}_2}e^{-C_1\beta t})\|\Phi_x\|\\ &\quad+C(1+t^{-\frac{1}{4}})\left(\langle t\rangle e^{-\delta bt}(\|v_0\|_{H^1_{a_1}}+\|v_0\|_{H^1_{a_2}})+\beta^{2\theta\tilde{\nu}_2-1}e^{-C_1\beta t}\right)^2.\end{aligned}$$

Using Gronwall's inequality, we obtain

$$\|\Phi\| \le C(\|v_0\|_{0,1} + \|v_0\|_{H^1_{a_1}} + \|v_0\|_{H^1_{a_2}} + \beta^{2\theta\tilde{\nu}_2-\frac{3}{2}} + \varepsilon^p) \tag{4.67}$$

if $p > 7/2$ and

$$\|\Phi_x\| \le C\langle t\rangle^{C'\varepsilon^2}(\|v_0\|_{1,1} + \|v_0\|_{H^1_{a_1}} + \|v_0\|_{H^1_{a_2}} + \beta^{2\theta\tilde{\nu}_2-\frac{3}{2}}), \tag{4.68}$$

where $C' = 0$ if $p > 3$ and C' is a positive number if $p = 3$. In the case where $p = 3$, we put $\mu(t) = t^\lambda\|\Phi\|^2$ for some $\lambda \in (0, 1/3)$. Applying Gronwall's inequality to μ, we obtain

$$\|\Phi\| \le C\langle t\rangle^{\frac{1}{6}}(\|v_0\|_{0,1} + \|v_0\|_{H^1_{a_1}} + \|v_0\|_{H^1_{a_2}} + \beta^{2\theta\tilde{\nu}_2-\frac{3}{2}} + \varepsilon^3). \tag{4.69}$$

Thus, by Lemma 4.7, (4.41), (4.42), (4.44)–(4.46), and (4.67)–(4.69), we obtain

$$\begin{aligned}\|J\tilde{v}\| &\le \|\Phi\| + 3t\|\bar{N}\| \\ &\le C(\|v_0\|_{0,1} + \|v_0\|_{H^1_{a_1}} + \|v_0\|_{H^1_{a_2}} + \beta^{2\theta\tilde{\nu}_2-\frac{3}{2}} + \varepsilon^p)\end{aligned} \tag{4.70}$$

for $7/2 < p < 5$,

$$\begin{aligned}\|J\tilde{v}\| &\le \|\Phi\| + 3t\|\bar{N}\| \\ &\le C\langle t\rangle^{\frac{1}{6}}(\|v_0\|_{0,1} + \|v_0\|_{H^1_{a_1}} + \|v_0\|_{H^1_{a_2}} + \beta^{2\theta\tilde{\nu}_2-\frac{3}{2}} + \varepsilon^3)\end{aligned} \tag{4.71}$$

for $p = 3$, and

$$\begin{aligned}\|\partial_x J\tilde{v}\| &\le \|\partial_x\Phi\| + 3t\|\partial_x\bar{N}\|C\langle t\rangle^{C'\varepsilon^2} \\ &\le C\langle t\rangle^{C'\varepsilon^2}(\|v_0\|_{1,1} + \|v_0\|_{H^1_{a_1}} + \|v_0\|_{H^1_{a_2}} + \beta^{2\theta\tilde{\nu}_2-\frac{3}{2}}),\end{aligned} \tag{4.72}$$

where $C' = 0$ if $p > 3$ and C' is a positive constant if $p = 3$.

By Remark 3.2, we have $\beta \le \varepsilon_1^{1+C\varepsilon_0}$ for some $C \in \mathbb{R}$. Since $2\theta\nu_2 > 3/2$, it follows from (4.45), (4.70), and (4.72) that there exist positive numbers B and η such that

$$\mathcal{M}[\tilde{v}] \le B(\varepsilon_0 + \varepsilon_1^\eta).$$

Thus, by the standard continuation argument, we have $T_4 = T_5 = \infty$.

To deal with the case where $3 < p \le 7/2$, we will estimate $\|D^\alpha J\tilde{v}\|$ with $\alpha = 1/2 - \gamma$ and $0 < \gamma < \min\{1/2, (p-3)/3\}$. Applying D^α to (4.47), multiplying by $D^\alpha\Phi$, and integrating the resulting equation by parts, we have

$$\begin{aligned}\frac{d}{dt}\|D^\alpha\Phi\|^2 &= 2\langle\mathfrak{L}D^\alpha\Phi, D^\alpha\Phi\rangle \\ &\le 2|\langle D^\alpha f'(\tilde{v})\partial_x\Phi, D^\alpha\Phi\rangle| + 2|\langle D^\alpha(f'(\varphi_{c_1}+\varphi_{c_2}+\tilde{v}) - f'(\tilde{v}))\partial_x\Phi, D^\alpha\Phi\rangle| \\ &\quad + 2|\langle D^\alpha K_2, D^\alpha\Phi\rangle| + 2\sum_{3\le i\le 5}\|e^{-az_1}D^\alpha K_i\|\|e^{az_1}D^\alpha\Phi\|.\end{aligned} \tag{4.73}$$

To estimate first and third terms of the right-hand side, we use the following lemma.

LEMMA 4.10 (see [19]). *Let $\alpha = 1/2-\gamma$ with $0<\gamma<\min\{1/2,(p-3)/3\}$. Then*

$$\|D^\alpha|u|^{p-1}u\| \le C\|u\|_{2(p-1)}^{p-1}(\|uu_x\|_\infty^{\frac{1}{2}} + \|u\|^{3\gamma}\|uu_x\|_\infty^{\frac{1-3\gamma}{2}}),$$

$$\begin{aligned}|\langle D^\alpha|u|^{p-1}\Phi_x, D^\alpha\Phi\rangle| \le{}& C\|D^\alpha\Phi\|(\|D^\alpha\Phi\|+\|\Phi_x\|)(\|u\|_\infty^{p-3}\|uu_x\|_\infty \\ &+ \|u\|_\infty^{p-3-2\gamma}\|u\|^{2\gamma}\|uu_x\|_\infty + \|u\|_\infty^{p-3+2\gamma}\|uu_x\|_\infty^{1-\gamma}).\end{aligned}$$

By Lemmas 4.7 and 4.10,

$$\begin{aligned}&|\langle D^\alpha f'(\tilde v)\partial_x\Phi, D^\alpha\Phi\rangle| \\ &\le C(\|\tilde v\|_\infty^{p-3}\|\tilde v\tilde v_x\|_\infty + \|\tilde v\|_\infty^{p-3-2\gamma}\|\tilde v\|^{2\gamma}\|\tilde v\tilde v_x\|_\infty + \|\tilde v\|_\infty^{p-3+2\gamma}\|\tilde v\tilde v_x\|_\infty^{1-\gamma}) \\ &\quad\times\|D^\alpha\Phi\|(\|D^\alpha\Phi\|+\|\Phi_x\|) \\ &\le C\varepsilon^{p-1-2\gamma}t^{-\frac{2}{3}}\langle t\rangle^{-\frac{p-2-2\gamma}{3}}\|D^\alpha\Phi\|(\|D^\alpha\Phi\|+\|\Phi_x\|).\end{aligned}$$

By Lemmas 4.10 and 4.5,

$$\begin{aligned}\|D^\alpha K_2\| &\le \|D^\alpha f'(\tilde v)\tilde v\| + \|D^\alpha\{f'(\varphi_{c_1}+\varphi_{c_2}+\tilde v)-f'(\tilde v)\}\tilde v\| \\ &\le C\|\tilde v\|_{2(p-1)}^{p-1}(\|\tilde v\tilde v_x\|_\infty^{\frac{1}{2}} + \|\tilde v\|_\infty^{3\gamma}\|\tilde v\tilde v_x\|_\infty^{\frac{1-3\gamma}{2}}) + C\|w\|_{1,0} \\ &\le C\varepsilon^p t^{-\frac{1}{3}}\langle t\rangle^{-\frac{p-1}{3}+\frac{\gamma}{2}} + C(e^{-\delta bt}\|v_0\|_{H^1_a} + \beta^{2\tilde\nu_2}e^{-C_1\beta t}).\end{aligned}$$

By (4.58) and (4.62),

$$\begin{aligned}&|\langle D^\alpha(f'(\varphi_{c_1}+\varphi_{c_2}+\tilde v)-f'(\tilde v))\partial_x\Phi, D^\alpha\Phi\rangle| \\ &\le C\|(\varphi_{c_1}+\varphi_{c_2})\partial_x\Phi\|\|D^{2\alpha}\Phi\| \\ &\le C\left(\langle t\rangle e^{-\delta bt}(\|v_0\|_{H^1_{a_1}}+\|v_0\|_{H^1_{a_2}}) + \beta^{2\theta\tilde\nu_2-1}e^{-C_1\beta t}\right)(\|D^\alpha\Phi\|+\|\Phi_x\|).\end{aligned}$$

Combining the above with

$$\begin{aligned}\|e^{az_1}D^\alpha\Phi\| &\le C\|e^{az_1}\Phi\|_{1,0} \\ &\le C\left(\langle t\rangle e^{-\delta bt}(\|v_0\|_{1,1}+\|v_0\|_{H^1_{a_1}}+\|v_0\|_{H^1_{a_2}}) + \beta^{2\theta\tilde\nu_2-1}e^{-C_1\beta t}\right),\end{aligned}$$

$$\|e^{-az_1}D^\alpha K_i\| \le C\|e^{-az_1}K_i\|_{1,0},$$

and (4.51)–(4.53), we obtain

$$\|D^\alpha\Phi\| \le C(\|v_0\|_{0,1}+\|v_0\|_{H^1_{a_1}}+\|v_0\|_{H^1_{a_2}}+\beta^{2\theta\tilde\nu_2-\frac{3}{2}}). \tag{4.74}$$

Thus we see that there exist positive constants B and η such that

$$\mathcal{M}[\tilde v] \le B(\varepsilon_0+\varepsilon_1^\eta)$$

holds for every $t\ge 0$ and $T_4=T_5=\infty$.

Finally, we will estimate $\|\mathcal{F}U_0(-t)\tilde v\|_\infty$ to deal with the case where $p=3$. Let $\phi(t)=\mathcal{F}U_0(-t)\tilde v$. Then (4.38) can be transformed into

$$\partial_t\phi = \mathcal{F}U_0(-t)(G-\partial_x\bar N). \tag{4.75}$$

Since $\|\mathcal{F}U_0(-t)f\|_\infty \le (2\pi)^{-\frac{1}{2}}\|f\|_1$ for $f \in L^1(\mathbb{R})$, it follows from (4.39), (4.41), and (4.42) that

$$\begin{aligned}
&\|\mathcal{F}U_0(-t)\partial_x(\bar{N}_{1,1}+\bar{N}_2)\|_\infty + \|\mathcal{F}U_0(-t)G\|_\infty \\
&\le C\int_0^t \left(\|\partial_x \bar{N}_{1,1}\|_1 + \|\partial_x \bar{N}_2\|_1 + \sum_{i=1,2}(|\dot{x}_i - c_i| + |\dot{c}_i|) \right) ds \\
&\le C\left(\|v_0\|_{H^1_{a_1}} + \|v_0\|_{H^1_{a_2}} + \beta^{2\tilde{\nu}_2 - 1} \right).
\end{aligned}$$

Hence we can prove that

$$\|\mathcal{F}U_0(-t)\tilde{v}\|_\infty \le C\left(\|v_0\|_{H^1_{a_1}} + \|v_0\|_{H^1_{a_2}} + \beta^{2\tilde{\nu}_2-1} + \widetilde{\mathcal{M}}[\tilde{v}]^3 \right), \tag{4.76}$$

following the arguments of Theorem 1.1 in [20]. By (4.46), (4.71), (4.72), and (4.76), there exist positive constants B and η such that

$$\widetilde{\mathcal{M}}[\tilde{v}] \le B(\|v_0\|_{1,1} + \|v_0\|_{H^1_{a_1}} + \|v_0\|_{H^1_{a_2}} + \beta^{2\theta\tilde{\nu}_2 - \frac{3}{2}})$$

for every $t \ge 0$. Thus we complete the proof. □

Now we are in position to prove Theorem 1.1.

Proof of Theorem 1.1. By Lemma 3.2 and (4.39), there exist

$$c_{i,\infty} := \lim_{t\to\infty} c_i(t) \quad \text{and} \quad x_{i,\infty} := \lim_{t\to\infty}(x_i(t) - c_{i,\infty}t)$$

satisfying

$$\begin{aligned}
|c_i(t) - c_{i,\infty}| &\le \int_t^\infty |\dot{c}_i(s)|ds \\
&= O\left(e^{-\delta b t}\|v_0\|_{H^1_a} + \beta^{2\tilde{\nu}_3 - 1} e^{-C_1\beta t} \right),
\end{aligned}$$

$$\begin{aligned}
|x_i(t) - x_{i,\infty} - c_{i,\infty}t| &\le \int_t^\infty |\dot{x}_i(s) - c_i(s)|ds + \int_t^\infty\int_s^\infty |\dot{c}_i(\tau)|d\tau ds \\
&= O\left(e^{-\delta b t}\|v_0\|_{H^1_a} + \beta^{2\tilde{\nu}_3 - 2} e^{-C_1\beta t} \right)
\end{aligned}$$

for $i = 1, 2$ and $C_1 > 0$. Combining the above with Lemmas 4.5 and 4.7, Propositions 4.8 and 4.9, and (4.45) yields (1.4)–(1.6).

If $3 < p < 5$, it follows from Proposition 4.8, (4.39), (4.41), (4.42), and (4.44) that

$$\begin{aligned}
\|U_0(-t)\tilde{v}(t) - U_0(-s)\tilde{v}(s)\| &\le \int_s^t (\|G\| + \|\partial_x \bar{N}\|)d\tau \\
&\le C\left(e^{-\delta b s}\|v_0\|_{H^1_a} + \beta^{2\tilde{\nu}_2 - 1} e^{-C_1\beta s} + \mathcal{M}[\tilde{v}]^{p-1}\langle s\rangle^{-\frac{p-3}{3}} \right)
\end{aligned}$$

for $t \ge s \ge 0$. Thus there exists a unique function $v_\infty \in L^2(\mathbb{R})$ such that (1.7) holds. □

Acknowledgments. The author would like to express his gratitude to Professor Shin-Ichiro Ei for stimulating discussions. The author would also like to express his gratitude to the referees for their careful reading of this article.

REFERENCES

[1] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform*, SIAM, Philadelphia, 1981.
[2] T. B. Benjamin, *The stability of solitary waves*, Proc. Roy. Soc. London A, 328 (1972), pp. 153–183.
[3] J. L. Bona, *On the stability of solitary waves*, Proc. Roy. Soc. London A, 344 (1975), pp. 363–374.
[4] J. L. Bona, *On solitary waves and their role in the evolution of long waves*, in Applications of Nonlinear Analysis in the Physical Sciences, H. Amann, N. Bazley, and K. Kirchgässner, eds., Pitman, London, 1981, pp. 183–205.
[5] J. L. Bona, P. E. Souganidis, and W. A. Strauss, *Stability and instability of solitary waves of Korteweg-de Vries type*, Proc. Roy. Soc. London., 411 (1987), pp. 395–412.
[6] J. L. Bona and A. Soyeur, *On the stability of solitary-wave solutions of model equations for long waves*, J. Nonlinear Sci., 4 (1994), pp. 449–470.
[7] J. Bourgain, *Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations* I, Geom. Funct. Anal., 3 (1993), pp. 107–156.
[8] J. Bourgain, *Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations* II, Geom. Funct. Anal., 3 (1993), pp. 209–262.
[9] V. S. Buslaev and G. S. Perelman, *Scattering for the nonlinear Schrödinger equation: States close to a soliton*, St. Petersburg Math. J., 4 (1993), pp. 1111–1142.
[10] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Sharp global well-posedness for KdV and modified KdV on* $\mathbb{R}$ *and* $\mathbb{T}$, J. Amer. Math. Soc., 16 (2003), pp. 705–749.
[11] F. M. Christ and M. I. Weinstein, *Dispersion of small amplitude solutions of the generalized Korteweg-de Vries equation,* J. Funct. Anal., 100 (1991), pp. 87–109.
[12] P. Deift and X. Zhou, *A steepest descent method for oscillatory Riemann-Hilbert problems. Asymptotics for the MKdV equation*, Ann. of Math. (2), 137 (1993), pp. 295–368.
[13] S.-I. Ei, *The motion of weakly interacting pulses in reaction-diffusion systems*, J. Dynam. Differential Equations, 14 (2002), pp. 85–137.
[14] S.-I. Ei, K. Fujii, and T. Kunihiro, *Renormalization-group method for reduction of evolution equations; invariant manifolds and envelopes*, Ann. Physics, 280 (2000), pp. 236–298.
[15] S.-I. Ei and T. Ohta, *The motion of interacting pulses*, Phys. Rev., 50 (1994), pp. 4672–4678.
[16] J. Ginibre, Y. Tsutsumi, and G. Velo, *Existence and uniqueness of solutions for the generalized Korteweg-de Vries equation*, Math. Z., 203 (1990), pp. 9–36.
[17] M. Grillakis, J. Shatah, and W. Strauss, *Stability theory of solitary waves in the presence of symmetry*, J. Funct. Anal., 74 (1987), pp. 160–197.
[18] D. Henry, *Geometric Theory of Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, Heidelberg, New York, 1981.
[19] N. Hayashi and P. Naumkin, *Large time asymptotics of solutions to the generalized Korteweg-de Vries equation*, J. Funct. Anal., 159 (1998), pp. 110–136.
[20] N. Hayashi and P. Naumkin, *On the modified Korteweg-de Vries equation*, Math. Phys. Anal. Geom., 4 (2001), pp. 197–201.
[21] T. Kato, *On the Cauchy problem for the (generalized) Korteweg-de Vries equation*, in Studies Applied Math., Adv. Math. Stud. Ser. 8, V. Guillemin, ed., Academic Press, New York, 1983, pp. 93–128.
[22] C. E. Kenig, G. Ponce, and L. Vega, *Well-posedness and scattering results for the generalized Korteweg-de Vries equation via the contraction principle*, Comm. Pure Appl. Math., 46 (1993), pp. 527–620.
[23] C. E. Kenig, G. Ponce, and L. Vega, *A bilinear estimate with applications to the KdV equation*, J. Amer. Math. Soc., 9 (1996), pp. 573–603.
[24] S. Klainerman, *Long time behavior of solutions to nonlinear evolution equations*, Arch. Ration. Mech. Anal., 78 (1982), pp. 73–89.
[25] S. Klainerman and G. Ponce, *Global small amplitude solutions to nonlinear evolution equations*, Comm. Pure Appl. Math., 36 (1983), pp. 133–141.
[26] D. J. Korteweg and G. de Vries, *On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves*, Philos. Mag., 539 (1895),

pp. 422–443.
[27] J. H. Maddocks and R. L. Sachs, *On the stability of KdV multi-solitons*, Comm. Pure Appl. Math., 46 (1993), pp. 867–901.
[28] Y. Martel and F. Merle, *A Liouville theorem for the critical generalized Korteweg-de Vries equation*, J. Math. Pures Appl. (9), 79 (2000), pp. 339–425.
[29] Y. Martel and F. Merle, *Instability of solitons for the critical generalized Korteweg-de Vries equation*, Geom. Funct. Anal., 11 (2001), pp. 74–123.
[30] Y. Martel and F. Merle, *Asymptotic stability of solitons for subcritical generalized KdV equations*, Arch. Ration. Mech. Anal., 157 (2001), pp. 219–254.
[31] Y. Martel and F. Merle, *Blow up in finite time and dynamics of blow up solutions for the L^2-critical generalized KdV equations*, J. Amer. Math. Soc., 15 (2002), pp. 617–664.
[32] Y. Martel, F. Merle, and T.-P. Tsai, *Stability and asymptotic stability in the energy space of the sum of N solitons for subcritical gKdV equations*, Comm. Math. Phys., 231 (2002), pp. 347–373.
[33] T. Mizumachi, *Large time asymptotics of solutions around solitary waves to the generaized Korteweg–de Vries equations*, SIAM J. Math. Anal., 32 (2001), pp. 1050–1080.
[34] G. S. Perelman, *Some results on the scattering of weakly interacting solitons for nonlinear Schrödinger equations*, in Spectral Theory, Microlocal Analysis, Singular Manifolds, Math. Top. 14, M. Demuth, E. Schrohe, B.-W. Schulze, and J. Sjöstrand, eds., Akademie Verlag, Berlin, 1997, pp. 78–137.
[35] R. L. Pego and M. I. Weinstein, *Eigenvalues and instabilities of solitary waves*, Phil. Trans. Roy. Soc. London A, 340 (1992), pp. 47–94.
[36] R. L. Pego and M. I. Weinstein, *Asymptotic stability of solitary waves*, Comm. Math. Phys., 164 (1994), pp. 305–349.
[37] G. Ponce and L. Vega, *Nonlinear small data scattering for the generalized Korteweg-de Vries equation*, J. Funct. Anal., 90 (1990), pp. 445–457.
[38] P. C. Schuur, *Asymptotic Analysis of Soliton Problems*, Lecture Notes in Math. 1232, Springer-Verlag, Berlin, Heidelberg, New York, 1986.
[39] J. Shatah, *Global existence of small solutions to nonlinear evolution equations*, J. Differential Equations, 46 (1982), pp. 409–425.
[40] W. A. Strauss, *Dispersion of low-energy waves for two conservative equations*, Arch. Ration. Mech. Anal., 55 (1974), pp. 86–92.
[41] A. Soffer and M. I. Weinstein, *Multichannel nonlinear scattering theory for nonintegrable equations*, Comm. Math. Phys., 133 (1990), pp. 119–146.
[42] A. Soffer and M. I. Weinstein, *Multichannel nonlinear scattering theory for nonintegrable equations* II*: The case of anisotropic potentials and data*, J. Differential Equations, 98 (1992), pp. 376–390.
[43] M. I. Weinstein, *Modulational stability of ground states of nonlinear Schrödinger equations*, SIAM J. Math. Anal., 16 (1985), pp. 472–491.
[44] M. I. Weinstein, *Lyapunov stability of ground states of nonlinear dispersive evolution equations*, Comm. Pure. Appl. Math., 39 (1986), pp. 51–68.

SIAM J. MATH. ANAL.
Vol. 35, No. 4, pp. 1081–1098

ON THE EFFECTS OF THERMAL DEGENERACY IN THE THERMISTOR PROBLEM*

XIANGSHENG XU†

Abstract. We investigate the boundedness of weak solutions to the initial boundary-value problem for the system

$$\begin{cases} \frac{\partial u}{\partial t} - \operatorname{div}(K(u)\nabla u) = \sigma(u)|\nabla\varphi|^2, \\ \operatorname{div}(\sigma(u)\nabla\varphi) = 0 \end{cases}$$

in the case where $K(u), \sigma(u)$ may both tend to 0 as $u \to \infty$. In particular, if solutions do blow up, our results impose a condition on the blow-up rates.

Key words. degeneracy, Hölder continuity, blow-up

AMS subject classifications. 35B65, 35K65

DOI. 10.1137/S0036141002417401

1. Introduction.

In this paper we consider the following problem:

$$\frac{\partial u}{\partial t} - \operatorname{div}(K(u)\nabla u) = \sigma(u)|\nabla\varphi|^2 \quad \text{in} \quad \Omega_T \equiv \Omega \times (0,T), \tag{1.1}$$

$$\operatorname{div}(\sigma(u)\nabla\varphi) = 0 \quad \text{in} \quad \Omega_T, \tag{1.2}$$

$$u(x,t) = u_0(x,t) \quad \text{on} \quad \partial_p\Omega_T, \tag{1.3}$$

$$\frac{\partial\varphi}{\partial\nu} = 0 \quad \text{on} \quad \Sigma_N \equiv \Gamma_N \times (0,T), \tag{1.4}$$

$$\varphi(x,t) = \varphi_0(x,t) \quad \text{on} \quad \Sigma_D \equiv \Gamma_D \times (0,T). \tag{1.5}$$

Here $T > 0$, Ω is a bounded domain in R^N with smooth boundary $\partial\Omega$, Γ_D is a nonempty open subset of $\partial\Omega$, $\Gamma_N = \partial\Omega \setminus \overline{\Gamma_D}$, $\partial_p\Omega_T$ is the parabolic boundary of Ω_T, and $u_0(x,t), \varphi_0(x,t), K(u), \sigma(u)$ are known functions of their arguments.

This problem is often called the thermistor problem, and it arises in the study of the electrical heating of a conductor (see [SSX, AC]). In this situation u is the temperature of the conductor, and φ is the electrical potential. The first equation describes the diffusion of heat in the presence of the Joule heating, which is the rate of energy generation associated with electrical current flow, while the second equation represents the conservation of electrical charges. The boundary conditions describe how the conductor is connected electrically and thermally to its surroundings. The function $K(u), \sigma(u)$ are the thermal conductivity and the electrical conductivity, respectively. Their precise forms are determined by the particular physical application one has in mind. See, e.g., [KO, XA] for various forms suggested for K, σ in industrial applications. In view of these, we assume the following.

(H1) $K(u), \sigma(u)$ are continuous and positive.

(H2) $u_0 \in W^{1,\infty}(\Omega_T), \varphi_0 \in L^\infty(0,T; W^{1,\infty}(\Omega))$.

Under (H1) and (H2), is there a bounded solution to (1.1)–(1.5) for any $T > 0$? If u in the solution does blow up, when and how does it blow up? These questions

*Received by the editors November 6, 2002; accepted for publication (in revised form) June 13, 2003; published electronically December 17, 2003.
http://www.siam.org/journals/sima/35-4/41740.html

†Department of Mathematics and Statistics, Mississippi State University, Mississippi State, MS 39762 (xxu@math.msstate.edu).

are important in both theory and application. What complicates the matter is the fact that assumption (H1) leaves open the possibility that $K(u), \sigma(u) \to 0$ as $u \to \infty$. Later we shall see that the interplay between thermal degeneracy and electrical degeneracy constitutes the main mathematical difficulty.

In all the previous papers [X1, X4, X5, X6], only the effects of electrical degeneracy are considered. Under the further assumption that $K(u) \equiv 1$ and $\sigma(u)$ is bounded above, several results concerning the existence of a weak solution, the so-called capacity solution, partial regularity, and global boundedness of weak solutions, are established there.

In this paper, we turn our attention to the thermal degeneracy and the interaction between thermal degeneracy and electrical degeneracy. This is, in part, motivated by the works of Allegretto and his collaborators [XA], where they suggest the following form for $K(u), \sigma(u)$:

$$K(u) \approx \frac{A}{u^2}, \tag{1.6}$$

$$\sigma(u) \approx \frac{B}{u^\gamma} \tag{1.7}$$

as $u \to \infty$. Here $A > 0, B > 0, 0 < \gamma \ll 1$. Note that in this example, γ is very close to 0. Thus we first consider the case where

$$c_1 \le \sigma \le c_2, \tag{1.8}$$

$$0 < K \le c_3, \tag{1.9}$$

where $c_i, i = 1, 2, 3$, are positive constants. That is to say, if σ behaves well, how degenerate can we allow K to be before u blows up? In this direction we obtain the following result.

THEOREM A. *Let* (H1), (H2), (1.8), *and* (1.9) *be satisfied. Assume*

$$K(s) \ge \frac{c}{s^\beta} \text{ on } [l, \infty) \text{ for some } l, c > 0, 0 < \beta < 2/N. \tag{1.10}$$

Then (1.1)–(1.5) *has a bounded solution on* Ω_T *for each* $T > 0$.

For simplicity, we assume that $u \ge 0$ throughout this paper. This can easily be achieved by assuming that $u_0 \ge 0$.

In view of (1.6), our restriction on β is obviously strong. This seems to suggest that the blow-up occurs when β is big. If this happens, is there any "life" after the blow-up? That is, do there exist any solutions in some weak sense after the temperature becomes unbounded? The following theorem attempts to answer this question.

THEOREM B. *Let* (H1), (H2) *be satisfied. Assume*

$$\sigma \le M \text{ for some } M > 0, \tag{1.11}$$

$$\int_0^\infty K(s)ds = \infty. \tag{1.12}$$

Then for each $T > 0$ *there is a capacity solution to* (1.1)–(1.5).

If $\int_0^\infty K(s)ds = a < \infty$, then the conclusion of the theorem is still valid if we impose the additional condition

$$\int_0^a \frac{1}{\sqrt{\sigma(F^{-1}(s))}} ds = \infty,$$

where $F(s) = \int_0^s K(\tau)d\tau$. This can be obtained easily by a result in [X2].

To recall the definition of a capacity solution, for $v \in L^2(0,T;W^{1,2}(\Omega))$, set

$$X_v = \{\varphi \in L^\infty(\Omega_T) : \rho(v)\varphi \in L^2(0,T;W^{1,2}(\Omega)) \text{ for each } \rho \in C_0^1(R)\}.$$

Then by a result in [X2], for each $\varphi \in X_v$, there is a unique function ξ defined on Ω_T such that

$$\xi = \bigtriangledown(\rho(v)\varphi) \quad \text{on} \quad \{(x,t) : |v(x,t)| \le M\}$$

for each $M > 0$ and each $\rho \in C_0^1(R)$ with $\rho = 1$ on $[-M, M]$. If we denote ξ by $\bigtriangledown\varphi$, we also have the product rule, i.e.,

$$\bigtriangledown(\varphi\rho(v)) = \bigtriangledown\varphi\rho(v) + \varphi\rho'(v)\nabla v \tag{1.13}$$

for each $\rho \in C_0^1(R)$. We can also define traces for $\varphi \in X_v$. We say $\varphi = \varphi_0$ on $\partial\Omega \times (0,T)$ for $\varphi_0 \in L^2(0,T;W^{1,2}(\Omega))$ if $\rho(v)\varphi = \rho(v)\varphi_0$ on $\partial\Omega \times (0,T)$ for all $\rho \in C_0^1(R)$.

DEFINITION. *A capacity solution to* (1.1)–(1.5) *is a couple* $\{u, \varphi\}$ *such that*

(i) $u \in L^\infty((0,T);L^1(\Omega)), v \equiv \int_0^u K(s)ds \in L^2(0,T;W^{1,2}(\Omega)), \varphi \in X_v$, $\sqrt{\sigma(u)}\nabla\varphi \in \left(L^2(\Omega_T)\right)^N$;

(ii) $v = v_0 \equiv \int_0^{u_0} K(s)ds$ *on* $\partial\Omega \times (0,T)$, $\varphi = \varphi_0$ *on* Σ_D;

(iii) $\int_{\Omega_T} \sigma(u) \bigtriangledown \varphi \bigtriangledown \xi \, dxdt = 0$ *for all* $\xi \in L^2(0,T;W^{1,2}(\Omega))$ *such that* $\xi = 0$ *on* Σ_D *and*

$$\begin{aligned} &-\int_{\Omega_T} u\xi_t \, dxdt + \int_{\Omega_T} K(u) \bigtriangledown u \bigtriangledown \xi dxdt \\ &= \int_{\Omega_T} \sigma(u)| \bigtriangledown \varphi|^2 \xi \, dxdt + \int_\Omega u_0\xi(x,0) \, dx \end{aligned}$$

for all $\xi \in W^{1,2}(\Omega_T) \cap L^\infty(\Omega_T)$ *with* $\xi(x,T) = 0, \xi = 0$ *on* $\partial\Omega \times (0,T)$.

The following theorem imposes a condition on the blow-up rate.

THEOREM C. *Let the assumptions of Theorem* B *be satisfied. Assume* (1.8) *holds. Then we have*

$$\int_0^t \int_0^{u(x,\tau)} K(s)dsd\tau \in L^\infty(\Omega_T). \tag{1.14}$$

Note that $u, \int_0^u K(s)ds$ blow up simultaneously because of (1.12). Thus (1.14) seems to indicate that if $\int_0^u K(s)ds$ does blow up at a finite time T_0, one should have

$$\int_0^{u(x,t)} K(s)ds \le \frac{c}{(T_0 - t)^\alpha} \text{ for some } \alpha \in (0,1). \tag{1.15}$$

Next we explore how thermal degeneracy and electrical degeneracy interact. In this direction, we establish the following theorem.

THEOREM D. *Let* (1.12), (H1), *and* (H2) *be satisfied. Assume*

(H3) $\sigma \in C^1(R), \sigma' < 0$, *and* $\lim_{s\to\infty} \sigma(s) = 0$;

(H4)

$$4\left(\|\varphi_0\|_{\infty,\Omega_T}\right)^2 \le \frac{K(s)}{-\sigma'(s)} \le c\int_0^s \tfrac{1}{\sigma(\tau)}d\tau + c \text{ on } [0,\infty) \text{ for some } c > 0. \tag{1.16}$$

Then we have

$$\int_0^t \int_0^{u(x,\tau)} K(s)dsd\tau \in L^\infty(0,T;L^\infty_{\rm loc}(\Omega)). \tag{1.17}$$

The second inequality in (1.16) is not essential and can be eliminated if we assume $\sigma \in C^2(R)$ and $\Delta u_0 \in L^\infty(\Omega_T)$. The first inequality in (1.16) roughly says that to expect u to behave well, $K(u)$ cannot be too degenerate in relation to $\sigma(u)$. For example, if $K(u) \approx c/u^\beta, \sigma(u) \approx c/u^\alpha$, then we should have $\beta - \alpha \le 1$. The condition $\lim_{s\to\infty} \sigma(s) = 0$ is not surprising. Recall that $\sigma(u)$ is the electrical conductivity. This condition says that the rise of temperature leads to a drop in electrical conductivity. This, in turn, results in the weakening of electrical current, thus slowing down the rise in temperature. This is one cancellation effect we try to explore here.

Finally, let us make some remarks about the notation. The letter c is used to denote the generic constant. Furthermore, if $r > 0, z = (x,t) \in \mathbf{R}^N \times (0,\infty)$, and u, φ are locally integrable, then

$$\begin{aligned} Q(z,r) &= \{(y,\tau) : |y-x| < r, t - r^2 < \tau < t\}, \\ u_{z,r} &= \fint_{Q(z,r)} u dy d\tau, \\ \varphi_{x,r}(\tau) &= \fint_{B(x,r)} \varphi(y,\tau) dy, \\ B(x,r) &= \{y : |y-x| < r\}. \end{aligned}$$

When the notation we use is standard, no explanation is given.

2. Proofs of Theorems A, B, and C. We first construct a capacity solution to (1.1)–(1.5) under the assumptions (H1), (H2), (1.11), and (1.12). For this purpose, define, for each k,

$$K_k(s) = \begin{cases} K(k) & \text{if } s > k, \\ K(s) & \text{if } s \le k, \end{cases}$$

$$\sigma_k = \sigma + \frac{1}{k}.$$

Consider the following approximate problem:

$$\frac{\partial}{\partial t} u_k - \operatorname{div}(K_k(u_k) \bigtriangledown u_k) = \sigma_k(u_k) |\bigtriangledown \varphi_k|^2 \quad \text{in} \quad \Omega_T, \tag{2.1}$$

$$\operatorname{div}(\sigma_k(u_k) \bigtriangledown \varphi_k) = 0 \quad \text{on} \quad \Omega_T, \tag{2.2}$$

$$\varphi_k = \varphi_0 \quad \text{on} \quad \Sigma_D, \tag{2.3}$$

$$\frac{\partial \varphi_k}{\partial \nu} = 0 \quad \text{on} \quad \Sigma_N, \tag{2.4}$$

$$u_k = u_0 \quad \text{on} \quad \partial_p \Omega_T. \tag{2.5}$$

The existence of a classical weak solution to (2.1)–(2.5) in the space

$$\begin{aligned} S_T \equiv\ & C^{\alpha,\alpha/2}(\overline{\Omega_T}) \cap L^p\left(0,T;W^{1,p}_{\rm loc}(\Omega)\right) \cap L^2(0,T;W^{1,2}(\Omega)) \\ & \times L^\infty\left(0,T;W^{1,p}_{\rm loc}(\Omega)\right) \cap L^\infty(0,T;W^{1,2}(\Omega)) \cap L^\infty(0,T;C^\delta(\overline{\Omega})) \\ & \text{for some } \alpha, \delta \text{ in } (0,1) \text{ and for each } p > 1 \end{aligned}$$

is established in [X3, X4]. Even though we have $K_k(u_k) = 1$ in [X3, X4], a careful examination of the proof of Theorem 7 in [X4] reveals that the argument employed there also works in the situation considered here.

Now we proceed to derive a priori estimates for approximate solutions. Let $v_k = \int_0^{u_k} K_k(s)ds$. We can easily conclude from [X1] that

$$\max_{\Omega_T} |\varphi_k| \le c, \tag{2.6}$$

$$\int_{\Omega_T} \sigma_k(u_k)| \bigtriangledown \varphi_k|^2 \, dxdt \le c, \tag{2.7}$$

$$\operatorname{ess\,sup}_{0\le t\le T} \int_\Omega (u_k(x,t) - l)^+ \, dx \le \int_{u_k>l} \sigma_k(u_k)| \bigtriangledown \varphi_k|^2 \, dxdt, \tag{2.8}$$

$$\int_{\Omega_T} | \bigtriangledown v_k|^2 \, dxdt \le c, \tag{2.9}$$

where $l \ge \|u_0\|_{\infty,\Omega_T}$. We may assume, passing to subsequences if necessary, that

$$\varphi_k \rightharpoonup \varphi \quad \text{weak}^* \text{ in } \ L^\infty(\Omega_T), \tag{2.10}$$

$$v_k \rightharpoonup v \quad \text{weakly in } \ L^2(0,T;W^{1,2}(\Omega)), \tag{2.11}$$

$$\sqrt{\sigma_k(u_k)} \bigtriangledown \varphi_k \rightharpoonup g \quad \text{weakly in } \ (L^2(\Omega_T))^N. \tag{2.12}$$

The remaining proof of the existence theorem is divided into the following several lemmas.

LEMMA 2.1. *$v_k \to v$ a.e. on Ω_T.*

Proof. First, it is easy to verify from (2.9) that $\{\rho(u_k)\}$ is bounded in $L^2(0,T;W^{1,2}(\Omega))$ for each $\rho \in C_0^1(R)$. Now pick a $\theta \in C_0^2(R)$. We can deduce from (2.1) and (2.9) that the sequence $\{\frac{\partial}{\partial t}\theta(u_k)\}$ is bounded $L^1(\Omega_T) + L^2(0,T;W^{-1,2}(\Omega))$. This puts us in a position to invoke a result in [BMR] to conclude that there is a measurable function u such that

$$u_k \to u \text{ a.e. on } \Omega_T. \tag{2.13}$$

This implies the desired result. □

LEMMA 2.2. *For each $\rho \in C_0^1(R)$, we have*

$$\rho(v_k)\varphi_k \rightharpoonup \rho(v)\varphi \quad \textit{weakly in} \quad L^2(0,T;W^{1,2}(\Omega)). \tag{2.14}$$

Note from (1.12) that the boundedness of $\{v_k\}$ on a set implies the boundedness of $\{u_k\}$ on the same set. Then (2.14) follows from (2.7) and an argument in [X1]. This lemma asserts that $\varphi \in X_v$.

LEMMA 2.3. *$g = \sqrt{\sigma(u)} \bigtriangledown \varphi$ a.e. on Ω_T.*

Once again, the proof can be found in [X1].

LEMMA 2.4.

$$\sqrt{\sigma_k(u_k)} \bigtriangledown \varphi_k \to g \quad \textit{strongly in} \ \ (L^2(\Omega_T))^N. \tag{2.15}$$

Proof. We modify slightly the proof in [X2]. Observe from (2.13) and (1.11) that $\sqrt{\sigma_k(u_k)} \to \sqrt{\sigma(u)}$ strongly in $L^q(\Omega_T)$ for each $q > 1$. Using $\varphi_k - \varphi_0$ as a test

function in (2.2) yields

$$\begin{aligned}\int_{\Omega_T} \sigma_k(u_k)|\bigtriangledown \varphi_k|^2\ dxdt &= \int_{\Omega_T} \sigma_k(u_k)\bigtriangledown \varphi_k \bigtriangledown\varphi_0\ dxdt\\ &= \int_{\Omega_T} \sqrt{\sigma_k(u_k)}\sqrt{\sigma_k(u_k)}\bigtriangledown \varphi_k\bigtriangledown\varphi_0\ dxdt\\ &\to \int_{\Omega_T} \sqrt{\sigma(u)}g\bigtriangledown\varphi_0\ dxdt. \qquad (2.16)\end{aligned}$$

On the other hand, we easily conclude from (2.2) that

$$\int_{\Omega_T} \sqrt{\sigma(u)}g\bigtriangledown\xi\ dxdt = 0 \qquad (2.17)$$

for all $\xi \in L^2(0,T;W^{1,2}(\Omega))$ with $\xi = 0$ on Σ_D. For each $l > 0$ choose $\theta_l \in C_0^1(R)$ so that

$$\theta_l(s) = \begin{cases} 0 & \text{if } |s| \geq 2l,\\ 1 & \text{if } |s| < l\end{cases}$$

and

$$|\theta_l{}'| \leq \frac{c}{l}.$$

Let $\xi = \theta_l(v)(\varphi - \varphi_0)$ in (2.17) and keep in mind (1.13) to get

$$\begin{aligned}&\int_{\Omega_T} \sqrt{\sigma(u)}g(\bigtriangledown\varphi - \bigtriangledown\varphi_0)\theta_l(v)\ dxdt\\ &= -\int_{\Omega_T} \sqrt{\sigma(u)}g\theta_l{}'(v)\nabla v(\varphi - \varphi_0)\ dxdt\\ &\to 0 \text{ as } l \to \infty. \qquad (2.18)\end{aligned}$$

This gives

$$\int_{\Omega_T} (\sigma(u)|\bigtriangledown \varphi|^2 - \sigma(u)\bigtriangledown \varphi\bigtriangledown\varphi_0)\ dxdt = 0. \qquad (2.19)$$

Combining (2.19), Lemma 2.3, and (2.16), we obtain

(2.20)
$$\lim_{k\to\infty}\int_{\Omega_T} \sigma_k(u_k)|\bigtriangledown \varphi_k|^2\ dxdt = \int_{\Omega_T} \sigma(u)\bigtriangledown \varphi\bigtriangledown\varphi_0\ dxdt = \int_{\Omega_T} \sigma(u)|\bigtriangledown \varphi|^2\ dxdt.$$

Consequently,

$$\begin{aligned}&\int_{\Omega_T} |\sqrt{\sigma_k(u_k)}\bigtriangledown \varphi_k - g|^2\ dxdt\\ &= \int_{\Omega_T} \sigma_k(u_k)|\bigtriangledown \varphi_k|^2\ dxdt - 2\int_{\Omega_T} \sqrt{\sigma_k(u_k)}\bigtriangledown \varphi_k g\ dxdt + \int_{\Omega_T} |g|^2\ dxdt\\ &\to \int_{\Omega_T} \sigma(u)|\bigtriangledown \varphi|^2\ dxdt - 2\int_{\Omega_T} |g|^2\ dxdt + \int_{\Omega_T} |g|^2\ dxdt\\ &= 0. \qquad \square\end{aligned}$$

LEMMA 2.5. *$u_k \to u$ strongly in $L^1(\Omega_T)$.*

This is a consequence of (2.15) and (2.8).

Now we can pass to the limit in (2.1)–(2.5) to conclude our proof of the existence theorem.

We now turn our attention to Theorem A. First we need to set up some technical machineries.

LEMMA 2.6. *Suppose that $g(x)$ is a measurable function on Ω with the property*

$$|\{|g| > s\}| \le \frac{A}{s^q} \text{ on } (0,\infty) \quad \text{for some } A > 0, q > 1. \tag{2.21}$$

Then for each $1 \le p < q$, we have

$$\left(\fint_\Omega |g|^p dx \right)^{1/p} \le \left(\frac{q}{q-p} \right)^{1/p} A^{1/q} \frac{1}{|\Omega|^{1/q}}. \tag{2.22}$$

This lemma is well known, and we will not offer a proof here.

In all our subsequent calculations, we assume that $(u, \varphi) \in S_T$. This is due to the fact that a capacity solution to (1.1)–(1.5) can be constructed as a limit of a sequence of approximate solutions in S_T. However, all the positive constants in our estimates depend only on the known data.

To find out how degenerate we can allow K to be before u blows up, we investigate Green's function associated with (1.1)–(1.5). For each $z = (x,t) \in \Omega_T$ define $\Gamma(y, \tau, x, t)$ to be the solution of the problem

$$\frac{\partial \Gamma}{\partial \tau} + \operatorname{div}(K(u)\nabla\Gamma) = 0 \text{ in } \Omega_t \equiv \Omega \times (0,t), \tag{2.23}$$

$$\Gamma = 0 \text{ on } \partial\Omega \times (0,t), \tag{2.24}$$

$$\Gamma|_{\tau=t} = \delta(x), \tag{2.25}$$

where $\delta(x)$ is the Dirac delta function concentrated at x. To construct a solution to this problem, select a sequence $\{\delta_n(y)\} \subset C_0^\infty(\Omega)$ with the properties

$$\delta_n \ge 0 \text{ on } \Omega, \tag{2.26}$$

$$\int_\Omega \delta_n(y) dy = 1, \tag{2.27}$$

$$\text{the support of } \delta_n \subset B(x, 1/n), \tag{2.28}$$

$$\int_\Omega \delta_n(y)\xi(y) dy \to \xi(x) \text{ as } n \to \infty \tag{2.29}$$

for each $\xi \in C_0^\infty(\Omega)$. Then it is easy to see that for each n the problem

$$\frac{\partial \Gamma_n}{\partial \tau} + \operatorname{div}(K(u)\nabla\Gamma_n) = 0 \text{ in } \Omega_t \equiv \Omega \times (0,t), \tag{2.30}$$

$$\Gamma_n = 0 \text{ on } \partial\Omega \times (0,t), \tag{2.31}$$

$$\Gamma_n|_{\tau=t} = \delta_n(y) \tag{2.32}$$

has a solution Γ_n in the space $C^{\beta,\beta/2}(\overline{\Omega}) \cap L^p(0,t; W^{1,p}(\Omega))$ for some $\beta \in (0,1)$ and all $p > 1$. Next we collect a few properties of Γ that are critical to the proof of Theorem A.

LEMMA 2.7. *There holds*

$$|\{\Gamma > s\}| \le \frac{c}{m}\frac{1}{s^{1+2/N}} \ \text{ on } (0,\infty), \tag{2.33}$$

where c depends only on N and c_3 in (1.9) and $m = \min_{\Omega_T} K(u)$.

This lemma can be obtained by modifying an argument in [GW, pp. 307–308].

LEMMA 2.8. *For each $r > 0$ there is a positive number c such that for every $z_0 = (y_0, \tau_0) \in \Omega_t, R > 0, 0 < \rho \le t - \tau_0$ there holds*

$$\begin{aligned}&\max_{Q_\infty \equiv B(u_0,R/2)\times(\tau_0,\tau_0+\rho/2)\cap\Omega_t} \Gamma \\ &\le \frac{c}{m^{\frac{N}{2(r+1)}}}\left((1/\rho + 1/R^2)^{\frac{N+2}{2}} \int_{Q_0 \equiv B(y_0,R)\times(\tau_0,\tau_0+\rho)\cap\Omega_t} \Gamma^{r+1} dy d\tau\right)^{\frac{1}{r+1}}.\end{aligned} \tag{2.34}$$

This lemma can be inferred from the proof of Lemma 1 in [M].

LEMMA 2.9. *Let (1.8) hold. Then there exist two numbers $c > 0, 0 < \alpha < 1$ such that*

$$\int_{B(x,R)\cap\Omega} \sigma(u)|\nabla\varphi|^2 dy \le cR^{N-2+2\alpha}. \tag{2.35}$$

Proof. Assumption (1.8) enables us to appeal to the classical regularity theory for uniformly elliptic equations to conclude

$$|\varphi(x,t) - \varphi(y,t)| \le c|x-y|^\alpha, \quad (y,t),(x,t) \in \Omega_T \tag{2.36}$$

for some $c > 0, \alpha \in (0,1)$. Fix $(x,t) \in \Omega_T, R > 0$. Then choose $\xi \in C_0^\infty(B(x,R))$ so that

$$0 \le \xi \le 1 \ \text{ on } \ B(x,R), \tag{2.37}$$

$$|\nabla\xi| \le \frac{c}{R} \ \text{ on } \ B(x,R), \tag{2.38}$$

$$\xi = 1 \ \text{ on } \ B(x,R/2). \tag{2.39}$$

If $\partial B(x,R) \cap \Gamma_D = \emptyset$, we use $\xi^2(\varphi(y,t) - \varphi(x,t))$ as a test function in (1.2) to get

$$\begin{aligned}&\int_{B(x,R)\cap\Omega} \sigma(u)|\nabla\varphi|^2\xi^2 dy = -\int_{B(x,R)\cap\Omega} \sigma(u)\nabla\varphi(\varphi(y,t)-\varphi(x,t))2\xi\nabla\xi dy \\ &\le \frac{1}{2}\int_{B(x,R)\cap\Omega} \sigma(u)|\nabla\varphi|^2\xi^2 dy + \frac{1}{2}\int_{B(x,R)\cap\Omega} 4\sigma(u)(\varphi(y,t)-\varphi(x,t)^2|\nabla\xi|^2 dy;\end{aligned} \tag{2.40}$$

from this it follows that

$$\int_{B(x,R/2)\cap\Omega} \sigma(u)|\nabla\varphi|^2 dy \le cR^{N-2+2\alpha}. \tag{2.41}$$

If $\partial B(x,R)\cap\Gamma_D\neq\emptyset$, we use $(\varphi-\varphi_0)\xi^2$ as a test function in (1.2) to get

$$\begin{aligned}\int_{B(x,R)\cap\Omega}\sigma(u)|\nabla\varphi|^2\xi^2dy &= \int_{B(x,R)\cap\Omega}\sigma(u)\nabla\varphi\nabla\varphi_0\xi^2dy\\ &\quad+\int_{B(x,R)\cap\Omega}\sigma(u)\nabla\varphi(\varphi-\varphi_0)2\xi\nabla\xi dy\\ &\le\frac{1}{2}\int_{B(x,R)\cap\Omega}\sigma(u)|\nabla\varphi|^2\xi^2dy\\ &\quad+\int_{B(x,R)\cap\Omega}\sigma(u)|\nabla\varphi_0|^2\xi^2dy\\ &\quad+\int_{B(x,R)\cap\Omega}4|\nabla\xi|^2\sigma(u)|\varphi-\varphi_0|^2dy.\end{aligned}\tag{2.42}$$

Let $y_0\in\partial B(x,R)\cap\Gamma_D$. In view of the fact that $\varphi(y_0,t)=\varphi_0(y_0,t)$, we have, for $y\in B(x,R)\cap\Omega$,

$$\begin{aligned}|\varphi(y,t)-\varphi_0(y,t)| &\le|\varphi(y,t)-\varphi_0(y_0,t)|+|\varphi_0(y_0,t)-\varphi_0(y,t)|\\ &\le c|y-y_0|^\alpha+c|y-y_0|\\ &\le cR^\alpha+cR.\end{aligned}\tag{2.43}$$

This, along with (2.42) and (H2), implies (2.41). □

LEMMA 2.10. *There holds*

$$u(x,t)\le\|u_0\|_{\infty,\Omega_T}+\int_{\Omega_t}\sigma(u)|\nabla\varphi|^2\Gamma(y,\tau,x,t)dyd\tau.\tag{2.44}$$

Proof. We may write

$$u=v+w,$$

where v is the solution of the problem

$$\frac{\partial v}{\partial\tau}-\operatorname{div}(K(u)\nabla v)=\sigma(u)|\nabla\varphi|^2\ \text{ in }\ \Omega_T,\tag{2.45}$$

$$v=0\ \text{ on }\ \partial_p\Omega_T\tag{2.46}$$

and w solves the problem

$$\frac{\partial w}{\partial\tau}-\operatorname{div}(K(u)\nabla w)=0\ \text{ in }\ \Omega_T,\tag{2.47}$$

$$w=u_0\ \text{ on }\ \partial_p\Omega_T.\tag{2.48}$$

Using Γ_n as a test function in (2.45) yields

$$\left(\frac{\partial v}{\partial\tau},\Gamma_n\right)+\int_\Omega K(u)\nabla v\nabla\Gamma_n dy=\int_\Omega\sigma(u)|\nabla\varphi|^2\Gamma_n dy,\tag{2.49}$$

where $(\cdot,\cdot)$ denotes the duality pairing between $W^{-1,2}(\Omega)$ and $W_0^{1,2}(\Omega)$. By using v as a test function in (2.30), we obtain

$$\left(\frac{\partial\Gamma_n}{\partial\tau},v\right)-\int_\Omega K(u)\nabla\Gamma_n\nabla v dy=0.\tag{2.50}$$

We deduce, with the aid of the chain rule, that

$$\int_0^t \left(\left(\frac{\partial v}{\partial \tau}, \Gamma_n \right) + \left(\frac{\partial \Gamma_n}{\partial \tau}, v \right) \right) d\tau = \int_0^t \frac{d}{d\tau} \int_\Omega v \Gamma_n dy d\tau \tag{2.51}$$
$$= \int_\Omega v(y,t) \delta_n(y) dy \to v(x,t) \text{ as } n \to \infty.$$

Add (2.50) to (2.49), integrate the resulting equation over $(0,t)$, and thereby obtain

$$v(x,t) = \int_{\Omega_t} \sigma(u) |\nabla \varphi|^2 \Gamma(y,\tau,x,t) dy d\tau. \tag{2.52}$$

By the maximum principle, we have

$$|w| \le \|u_0\|_{\infty, \Omega_T}. \tag{2.53}$$

Then the lemma follows. □

To complete the proof of Theorem A, for each $(y,\tau) \in \Omega \times [0,t)$, set $\rho = t - \tau > 0, R = \sqrt{\rho}$. Choose $r > 0$ so that $1 + r < 1 + 2/N$. We estimate, with the aid of Lemmas 2.6, 2.7, and 2.8, that

$$\Gamma(y,\tau,x,t) \le \frac{c}{m^{\frac{N}{2(r+1)}}} \left(\fint_{B(y,R) \times (\tau, \tau + R^2)} \Gamma^{r+1} dy d\tau \right)^{\frac{1}{r+1}} \tag{2.54}$$
$$\le \frac{c}{m^{\frac{N}{2(r+1)} + \frac{N}{N+2}} (t-\tau)^{N/2}}.$$

If $|y - x| = R > 0$, we infer from the proof of Lemma 2.8 that we can take $\rho = R^2$ in (2.34). Therefore, we obtain

$$\Gamma(y,\tau,x,t) \le \frac{c}{m^{\frac{N}{2(r+1)} + \frac{N}{N+2}} |y-x|^N}. \tag{2.55}$$

Let $R > 0$ be so big that $\Omega_t \subset Q(z,R)$. For each $(y,\tau) \in Q(z, \frac{R}{2^i}) \setminus Q(z, \frac{R}{2^{i+1}}) \cap \Omega_t$, where $i = 0, 1, 2, \ldots$, we have either $|y - x| \ge \frac{R}{2^{i+1}}$ or $t - \tau \ge \left(\frac{R}{2^{i+1}} \right)^2$. Thus it follows from (2.54) and (2.55) that the following always holds:

$$\max_{Q(z, \frac{R}{2^i}) \setminus Q(z, \frac{R}{2^{i+1}}) \cap \Omega_t} \Gamma(y,\tau,x,t) \le \frac{c}{m^{\frac{N}{2(r+1)} + \frac{N}{N+2}}} \left(\frac{2^{i+1}}{R} \right)^N. \tag{2.56}$$

We easily conclude from Lemma 2.9 that

$$\int_{Q(z, \frac{R}{2^i})} \sigma(u) |\nabla \varphi|^2 dy d\tau \le c \left(\frac{R}{2^i} \right)^{N+2\alpha}. \tag{2.57}$$

Now we are ready to estimate

$$\int_{\Omega_t} \sigma(u) |\nabla \varphi|^2 \Gamma(y,\tau,x,t) dy d\tau \tag{2.58}$$
$$= \sum_{i=0}^\infty \int_{Q(z, \frac{R}{2^i}) \setminus Q(z, \frac{R}{2^{i+1}}) \cap \Omega_t} \sigma(u) |\nabla \varphi|^2 \Gamma(y,\tau,x,t) dy d\tau$$
$$\le \sum_{i=0}^\infty \frac{c}{m^{\frac{N}{2(r+1)} + \frac{N}{N+2}}} \left(\frac{R}{2^{i+1}} \right)^{2\alpha}$$
$$\le \frac{c}{m^{\frac{N}{2(r+1)} + \frac{N}{N+2}}}.$$

In view of (1.10), we have

$$m \geq \frac{1}{(\max_{\Omega_T} u)^{\beta}}. \tag{2.59}$$

Thus by (2.44) and (2.58), we derive

$$\max_{\Omega_T} u \leq c(\max_{\Omega_T} u)^{\beta(\frac{N}{2(r+1)}+\frac{N}{N+2})} + \|u_0\|_{\infty,\Omega_T}. \tag{2.60}$$

Since $\beta < 2/N$, we can find $r > 0$ so that the following two inequalities hold:

$$1 + r < 1 + 2/N, \tag{2.61}$$

$$\beta\left(\frac{N}{2(r+1)} + \frac{N}{N+2}\right) < 1. \tag{2.62}$$

The proof is complete.

Before we prove Theorem C, let us state the following lemma.

LEMMA 2.11. *Let $u \in W_0^{1,1}(\Omega)$. Then*

$$u(x) = \frac{1}{N\omega_N} \int_{\Omega} \tfrac{x-y}{|x-y|^N} \cdot \nabla u dy \ \ a.e.\ on\ \Omega. \tag{2.63}$$

We refer the reader to [GT, p. 161] for proof.

Proof of Theorem C. For $\epsilon > 0$ define

$$\theta_\epsilon(s) = \begin{cases} 1 & \text{if } s > \epsilon, \\ s/\epsilon & \text{if } 0 \leq s \leq \epsilon, \\ 0 & \text{if } s < 0. \end{cases}$$

Fix $l \geq \|u_0\|_{\infty,\Omega_T}$. Remember that $x \in \Omega$ and $u \in L^p(0,T;W^{1,p}_{\text{loc}}(\Omega))$ for each $p > 1$. Thus $\theta_\epsilon(u-l)\frac{1}{|x-y|^{N-2}}$ is a legitimate test function for (1.1). Thus use it in (1.1) to obtain

$$\begin{aligned} &\int_\Omega \int_0^{u(y,\tau)} \theta_\epsilon(s-l)ds \frac{1}{|x-y|^{N-2}} dy \\ &\quad + \int_{\Omega_t} K(u)\theta_\epsilon'(u-l)|\nabla u|^2 \frac{1}{|x-y|^{N-2}} dyd\tau \\ &\quad + (N-2)\int_{\Omega_t} K(u)\nabla u \theta_\epsilon(u-l)\frac{x-y}{|x-y|^N} dyd\tau \\ &= \int_{\Omega_t} \sigma(u)|\nabla\varphi|^2 \theta_\epsilon(u-l)\frac{1}{|x-y|^{N-2}} dyd\tau \\ &\leq c. \end{aligned} \tag{2.64}$$

The last step is due to (2.35). The third integral on the left-hand side of (2.64) can be evaluated by using Lemma 2.11 as follows:

$$\begin{aligned} &\int_{\Omega_t} K(u)\nabla u\theta_\epsilon(u-l)\frac{x-y}{|x-y|^N} dyd\tau \\ &= \int_0^t \int_\Omega \nabla \int_0^{u(y,\tau)} K(s)\theta_\epsilon(s-l)ds \frac{x-y}{|x-y|^N} dyd\tau \\ &= \int_0^t \int_0^{u(x,\tau)} K(s)\theta_\epsilon(s-l)dsd\tau. \end{aligned} \tag{2.65}$$

By taking $\epsilon \to 0$, we get

$$\int_\Omega (u-l)^+ \frac{1}{|x-y|^{N-2}} dy + \int_0^t \left(\int_l^{u(x,\tau)} K(s) ds \right)^+ \le c. \tag{2.66}$$

This implies the desired result.

3. Proof of Theorem D. We first present some preparatory lemmas which eventually leads to the proof of Theorem D. Once again, we assume that $(u, \varphi) \in S_T$ in our calculations.

LEMMA 3.1. *Let $h \in L^\infty_{\text{loc}}(\Omega) \cap L^1(\Omega), g \in L^\infty_{\text{loc}}(\Omega)$ be such that $h, g \ge 0$. Assume that there holds*

$$h(x) \le A \int_\Omega h(y) \frac{1}{|x-y|^{N-1}} dy + g(x) \ \text{ a.e. on } \Omega. \tag{3.1}$$

Then there exist two positive numbers $c = c(A, N), c_1 = c_1(A, N)$ such that

$$\max_{B(x_0, R/2)} h \le 2 \max_{B(x_0,R)} g + \frac{c_1}{R^{N-1}} \int_\Omega h dy \tag{3.2}$$

for all $x_0 \in \Omega, 0 < R \le \min\{c, d(x_0, \partial\Omega)\}$.

Proof. Fix $x_0 \in \Omega$. Let $0 < r < R \le d(x, \partial\Omega)$. Then for $x \in B(x_0, r)$, we have

$$\begin{aligned} &\int_\Omega h(y) \frac{1}{|x-y|^{N-1}} dy \\ &= \int_{B(x_0,R)} h(y) \frac{1}{|x-y|^{N-1}} dy + \int_{\Omega \setminus B(x_0,R)} h(y) \frac{1}{|x-y|^{N-1}} dy \\ &\le \max_{B(x_0,R)} h \int_{B(x_0,R)} \frac{1}{|x-y|^{N-1}} dy + \frac{1}{|R-r|^{N-1}} \int_\Omega h dy \\ &= c(N) R \max_{B(x_0,R)} h + \frac{1}{|R-r|^{N-1}} \int_\Omega h dy. \end{aligned} \tag{3.3}$$

On account of (3.1), we obtain

$$\max_{B(x_0,r)} h \le cR \max_{B(x_0,R)} h + \frac{A}{|R-r|^{N-1}} \int_\Omega h dy + \max_{B(x_0,R)} g \tag{3.4}$$

for all $0 < r < R \le d(x, \partial\Omega)$. Now fix such an R. For each $i = 0, 1, 2, \ldots$, set

$$R_i = R - \frac{R}{2^{i+1}}.$$

Let $r = R_n, R = R_{i+1}$ in (3.4) to get

$$Y_i \le \varepsilon Y_{i+1} + B(2^{N-1})^i, \tag{3.5}$$

where

$$\begin{aligned} \varepsilon &= cR, \\ B &= \max_{B(x_0,R)} g + \frac{c_1}{R^{N-1}} \int_\Omega h dy. \end{aligned}$$

Then the lemma follows from a result in [D, p. 13]. ∎

LEMMA 3.2. *There holds*

$$\begin{aligned}(3.6)\qquad &\int_\Omega |\nabla\varphi|^2 \frac{1}{|x-y|^{N-2}}dy \\ &\le 2\left(\|\varphi_0\|_{\infty,\Omega_T}\right)^2 \int_\Omega \left(\frac{\sigma'(u)}{\sigma(u)}\right)^2 |\nabla u|^2 \frac{1}{|x-y|^{N-2}}dy + \frac{c}{(d(x,\partial\Omega))^{N-1}}.\end{aligned}$$

Proof. First we infer from (1.2) that

$$(3.7)\qquad -\Delta\varphi = \frac{\sigma'(u)}{\sigma(u)}\nabla u\nabla\varphi.$$

By using $(\varphi-\varphi_0)\frac{1}{|x-y|^{N-2}}$ as a test function here, we derive

$$\begin{aligned}(3.8)\qquad &\int_\Omega \nabla\varphi(\nabla\varphi-\nabla\varphi_0)\frac{1}{|x-y|^{N-2}}dy + (N-2)\int_\Omega \nabla\varphi(\varphi-\varphi_0)\frac{x-y}{|x-y|^N}dy \\ &= \int_\Omega \frac{\sigma'(u)}{\sigma(u)}\nabla u\nabla\varphi(\varphi-\varphi_0)\frac{1}{|x-y|^{N-2}}dy \\ &\le 1/4\int_\Omega |\nabla\varphi|^2\frac{1}{|x-y|^{N-2}}dy + \int_\Omega \left(\frac{\sigma'(u)}{\sigma(u)}\right)^2(\varphi-\varphi_0)^2|\nabla u|^2\frac{1}{|x-y|^{N-2}}dy.\end{aligned}$$

To bound the second integral on the left-hand side of (3.8), we see that for each $\rho\in(0,d(x,\partial\Omega))$ there holds

$$\begin{aligned}(3.9)\qquad \int_{\Omega\setminus B(x,\rho)} \nabla\varphi(\varphi-\varphi_0)\frac{x-y}{|x-y|^N}dy &= \int_{\Omega\setminus B(x,\rho)} \nabla\frac{1}{2}(\varphi-\varphi_0)^2\frac{x-y}{|x-y|^N}dy \\ &\quad + \int_{\Omega\setminus B(x,\rho)} \nabla\varphi_0(\varphi-\varphi_0)\frac{x-y}{|x-y|^N}dy \\ &= \int_{\partial\Omega}\frac{1}{2}(\varphi-\varphi_0)^2\frac{x-y}{|x-y|^N}\cdot\nu d\mathcal{H}^{N-1} \\ &\quad + \frac{1}{\rho^{N-1}}\int_{\partial B(x,\rho)}\frac{1}{2}(\varphi-\varphi_0)^2 d\mathcal{H}^{N-1} \\ &\quad + \int_{\Omega\setminus B(x,\rho)} \nabla\varphi_0(\varphi-\varphi_0)\frac{x-y}{|x-y|^N}dy,\end{aligned}$$

where ν is the unit outward normal to $\partial\Omega$. Consequently, we have

$$\begin{aligned}(3.10)\qquad \left|\int_\Omega \nabla\varphi(\varphi-\varphi_0)\frac{x-y}{|x-y|^N}dy\right| &= \lim_{\rho\to 0}\left|\int_{\Omega\setminus B(x,\rho)} \nabla\varphi(\varphi-\varphi_0)\frac{x-y}{|x-y|^N}dy\right| \\ &\le \frac{c}{(d(x.\partial\Omega))^{N-1}}.\end{aligned}$$

Then the lemma follows from (3.10) and (3.8). □

Now we are ready to prove Theorem D. Let $f\in C^1(\mathbf{R})$ be such that

$$(3.11)\qquad f>0, f'>0 \text{ on } \mathbf{R}.$$

Then for each $x \in \Omega$, the function $(f(u)-f(u_0))\frac{1}{|x-y|^{N-2}}$ is a legitimate test function. Upon using it, we obtain

$$\begin{aligned}
&\left(\frac{\partial u}{\partial \tau}, (f(u)-f(u_0))\frac{1}{|x-y|^{N-2}}\right) + \int_\Omega K(u)f'(u)|\nabla u|^2 \frac{1}{|x-y|^{N-2}}dy \\
&\quad +(N-2)\int_\Omega K(u)\nabla u(f(u)-f(u_0))\frac{x-y}{|x-y|^N}dy \\
&= \int_\Omega K(u)f'(u_0)\nabla u \nabla u_0 \frac{1}{|x-y|^{N-2}}dy \\
&\quad + \int_\Omega (f(u)-f(u_0))\sigma(u)|\nabla \varphi|^2 \frac{1}{|x-y|^{N-2}}dy. \qquad (3.12)
\end{aligned}$$

Using a standard approximation argument, we can show

$$\begin{aligned}
(3.13)\quad &\left(\frac{\partial u}{\partial \tau}, (f(u)-f(u_0))\frac{1}{|x-y|^{N-2}}\right) = \frac{d}{d\tau}\int_\Omega \int_0^{u(y,\tau)} f(s)ds \frac{1}{|x-y|^{N-2}}dy \\
&-\frac{d}{d\tau}\int_\Omega uf(u_0)\frac{1}{|x-y|^{N-2}}dy + \int_\Omega uf'(u_0)\frac{\partial u_0}{\partial \tau}\frac{1}{|x-y|^{N-2}}dy.
\end{aligned}$$

To evaluate the third integral on the right-hand of (3.12), we appeal to Lemma 2.11, thereby obtaining

$$\begin{aligned}
(3.14)\quad &\int_\Omega K(u)\nabla u(f(u)-f(u_0))\frac{x-y}{|x-y|^N}dy \\
&= \int_\Omega \nabla \int_{u_0(y,\tau)}^{u(y,\tau)} K(s)f(s)ds \frac{x-y}{|x-y|^N}dy - \int_\Omega f(u_0)\nabla \int_{u_0(y,\tau)}^{u(y,\tau)} K(s)ds \frac{x-y}{|x-y|^N}dy \\
&= N\omega_N \int_{u_0(x,\tau)}^{u(x,\tau)} K(s)f(s)ds - \int_\Omega \nabla\left(f(u_0)\int_{u_0(y,\tau)}^{u(y,\tau)} K(s)ds\right)\frac{x-y}{|x-y|^N}dy \\
&\quad + \int_\Omega f'(u_0)\nabla u_0 \int_{u_0(y,\tau)}^{u(y,\tau)} K(s)ds \frac{x-y}{|x-y|^N}dy \\
&= N\omega_n \int_{u_0(x,\tau)}^{u(x,\tau)} K(s)(f(s)-f(u_0))ds + \int_\Omega f'(u_0)\nabla u_0 \int_{u_0(y,\tau)}^{u(y,\tau)} K(s)ds \frac{x-y}{|x-y|^N}dy.
\end{aligned}$$

It is easy to see that

$$\begin{aligned}
(3.15)\quad &\int_\Omega K(u)f'(u_0)\nabla u \nabla u_0 \frac{1}{|x-y|^{N-2}}dy \\
&\le \frac{1}{2}\int_\Omega K(u)f'(u)|\nabla u|^2 \frac{1}{|x-y|^{N-2}}dy \\
&\quad +\frac{1}{2}\int_\Omega K(u)\frac{(f'(u_0))^2}{f'(u)}|\nabla u_0|^2 \frac{1}{|x-y|^{N-2}}dy.
\end{aligned}$$

Now use (3.14), (3.13), and (3.15) in (3.12) and integrate the resulting equation with

respect to τ over $(0,t)$, thereby obtaining

$$
\begin{aligned}
(3.16)\quad & \int_\Omega \int_0^{u(y,t)} f(s)ds \frac{1}{|x-y|^{N-2}}dy + \frac{1}{2}\int_{\Omega_t} K(u)f'(u)|\nabla u|^2 \frac{1}{|x-y|^{N-2}}dyd\tau \\
& + c(N)\int_0^t \int_{u_0}^u K(s)(f(s)-f(u_0))dsd\tau \\
\le\ & \frac{1}{2}\int_{\Omega_t} \frac{K(u)}{f'(u)}(f'(u_0))^2|\nabla u_0|^2 \frac{1}{|x-y|^{N-2}}dyd\tau \\
& + \int_{\Omega_t} f(u)\sigma(u)|\nabla\varphi|^2 \frac{1}{|x-y|^{N-2}}dyd\tau \\
& + \int_\Omega uf(u_0)\frac{1}{|x-y|^{N-2}}dy - \int_{\Omega_t} uf'(u_0)\frac{\partial u_0}{\partial \tau}\frac{1}{|x-y|^{N-2}}dyd\tau \\
& + \int_{\Omega_t}\int_{u_0}^u K(s)dsf'(u_0)\nabla u_0 \frac{x-y}{|x-y|^N}dyd\tau \\
& + \int_\Omega \left(\int_0^{u_0} f(s)ds - u_0 f(u_0)\right)\frac{1}{|x-y|^{N-2}}dy.
\end{aligned}
$$

Recall that

$$
(3.17)\qquad \sigma'(s) < 0,
$$

and thus we can take $f(s) = 1/\sigma(s)$ in (3.16) and then apply Lemma 3.2 to get

$$
\begin{aligned}
(3.18)\quad & \int_\Omega \int_0^u \tfrac{1}{\sigma(s)}ds \frac{1}{|x-y|^{N-2}}dy + \frac{1}{2}\int_{\Omega_t} \frac{K(u)(-\sigma'(u))}{\sigma^2(u)}|\nabla u|^2 \frac{1}{|x-y|^{N-2}}dyd\tau \\
& + c(N)\int_0^t\int_{u_0}^u K(s)\left(\frac{1}{\sigma(s)} - \frac{1}{\sigma(u_0)}\right)dsd\tau \\
\le\ & c\int_{\Omega_t}\frac{K(u)}{f'(u)}\frac{1}{|x-y|^{N-2}}dyd\tau \\
& + 2\left(\|\varphi_0\|_{\infty,\Omega_T}\right)^2\int_\Omega \left(\frac{\sigma'(u)}{\sigma(u)}\right)^2|\nabla u|^2\frac{1}{|x-y|^{N-2}}dy + \frac{c}{(d(x,\partial\Omega))^{N-1}} \\
& + c\int_\Omega u\frac{1}{|x-y|^{N-2}}dy + c\int_{\Omega_t} u\frac{1}{|x-y|^{N-2}}dyd\tau \\
& + c\int_{\Omega_t}\int_{u_0}^u K(s)ds\frac{1}{|x-y|^{N-1}}dyd\tau + c.
\end{aligned}
$$

From (H4) we see that

$$
(3.19)\qquad \frac{K(u)(-\sigma'(u))}{\sigma^2(u)} \ge 4\left(\|\varphi_0\|_{\infty,\Omega_T}\right)^2\left(\frac{\sigma'(u)}{\sigma(u)}\right)^2.
$$

This enables us to drop both the second integral on the right and the second integral

on the left in (3.18). Also, since $\lim_{s\to 0}\sigma(s)=0$, we can find l so that

$$\tfrac{1}{\sigma(s)} \geq \max_{\Omega_T} \frac{2}{\sigma(u_0)} \text{ on } [l,\infty).$$

Keeping this in mind, we estimate

$$\begin{aligned}
&\int_0^t \int_{u_0(y,\tau)}^{u(y,\tau)} K(s)\left(\frac{1}{\sigma(s)}-\frac{1}{\sigma(u_0)}\right) ds d\tau \\
&\geq \int_{u(y,\tau)>l} \int_{u_0(y,\tau)}^{u(y,\tau)} K(s)\left(\frac{1}{\sigma(s)}-\frac{1}{\sigma(u_0)}\right) ds d\tau - c \\
&= \int_{u(y,\tau)>l} \int_{u_0(y,\tau)}^{l} K(s)\left(\frac{1}{\sigma(s)}-\frac{1}{\sigma(u_0)}\right) ds d\tau \\
&\quad + \int_{u(y,\tau)>l} \left(\int_l^{u(y,\tau)} K(s)\left(\frac{1}{\sigma(s)}-\frac{1}{\sigma(u_0)}\right) ds\right)^+ d\tau - c \\
&\geq -c.
\end{aligned} \tag{3.20}$$

The fact that $\lim_{s\to\infty}\sigma(s)=0$ also enables us to show that for each $\varepsilon>0$ there is a positive number c such that

$$u \leq \varepsilon \int_0^u \frac{1}{\sigma(s)} ds + c \text{ on } [0,\infty). \tag{3.21}$$

Use this and (3.20) in (3.18) to obtain

$$\int_\Omega \int_0^{u(x,\tau)} \frac{1}{\sigma(s)} ds \frac{1}{|x-y|^{N-2}} dy \leq c \int_{\Omega_t} \int_0^{u(y,\tau)} \frac{1}{\sigma(s)} \frac{1}{|x-y|^{N-2}} dy d\tau + g(x,t), \tag{3.22}$$

where

$$g(x,t) = c\int_{\Omega_t} \int_0^{u(y,\tau)} K(s) ds \frac{1}{|x-y|^{N-1}} dy d\tau + \frac{c}{(d(x,\partial\Omega))^{N-1}}. \tag{3.23}$$

An application of Gronwall's inequality gives

$$\begin{aligned}
&\int_{\Omega_t} \int_0^{u(y,\tau)} \frac{1}{\sigma(s)} ds \frac{1}{|x-y|^{N-2}} dy d\tau \\
&\leq c\int_0^t g(x,\tau) d\tau \\
&\leq c\int_{\Omega_t} \int_0^{u(y,\tau)} K(s) ds \frac{1}{|x-y|^{N-1}} dy d\tau + \frac{c}{(d(x,\partial\Omega))^{N-1}}.
\end{aligned} \tag{3.24}$$

Plugging this into (3.18) again, we obtain

$$\begin{aligned}
&\int_0^t \int_{u_0(x,\tau)}^{u(x,\tau)} K(s)\left(\frac{1}{\sigma(s)}-\frac{1}{\sigma(u_0)}\right) ds d\tau \\
&\leq c\int_{\Omega_t} \int_0^{u(y,\tau)} K(s) ds \frac{1}{|x-y|^{N-1}} dy d\tau + \frac{c}{(d(x,\partial\Omega))^{N-1}},
\end{aligned} \tag{3.25}$$

from which it follows that

$$\int_0^t \int_0^{u(x,\tau)} K(s)dsd\tau \leq c\int_\Omega \int_0^t \int_0^{u(y,\tau)} K(s)dsd\tau \frac{1}{|x-y|^{N-1}}dy + \frac{c}{(d(x,\partial\Omega))^{N-1}}. \tag{3.26}$$

Recall from (2.9) that $\int_0^{u(x,\tau)} K(s)ds \in L^2(\Omega_T)$. The theorem follows from Lemma 3.1.

Note that if we assume that $f(s) = \frac{1}{\sigma(s)} \in C^2(R)$ and $u_0 \in W^{2,\infty}(\Omega)$, we can also evaluate the first integral on the right-hand side of (3.12) as follows:

$$\begin{aligned}
&\left|\int_\Omega K(u)f'(u_0)\nabla u\nabla u_0 \frac{1}{|x-y|^{N-2}}dy\right| \\
&= \left|\int_\Omega \nabla \int_{u_0(y,\tau)}^{u(y,\tau)} K(s)ds\nabla f(u_0)\frac{1}{|x-y|^{N-2}}dy + \int_\Omega |\nabla f(u_0)|^2 \frac{1}{|x-y|^{N-2}}dy\right| \\
&\leq \left|-\int_\Omega \int_{u_0(y,\tau)}^{u(y,\tau)} K(s)ds\Delta f(u_0)\frac{1}{|x-y|^{N-2}}dy\right| \\
&\quad + \left|(N-2)\int_\Omega \int_{u_0(y,\tau)}^{u(y,\tau)} K(s)ds\nabla f(u_0)\frac{x-y}{|x-y|^N}dy\right| + c \\
&\leq c\int_\Omega \int_0^{u(y,\tau)} K(s)ds\frac{1}{|x-y|^{N-1}}dy + c.
\end{aligned}$$

In this case the second inequality in (1.16) can be removed.

The difficulty we are facing here is the lack of a comparison principle of any kind. In fact, this represents the fundamental difference between systems of equations and single equations.

REFERENCES

[AC] S.N. Antontsev and M. Chipot, *The thermistor problem: Existence, smoothness, uniqueness, blow up*, SIAM J. Math. Anal., 25 (1994), pp. 1128–1156.

[BMR] D. Blanchard, F. Murat, and H. Redwane, *Existence and uniqueness of a renormalized solution for a fairly general class of nonlinear parabolic problems*, J. Differential Equations, 177 (2001), pp. 331–374.

[D] E. DiBenedetto, *Degenerate Parabolic Equations*, Springer-Verlag, New York, 1993.

[GT] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, New York, 1983.

[GW] M. Gruter and K.-O. Widman, *The Green function for uniformly elliptic equations*, Manuscripta Math., 37 (1982), pp. 303–342.

[KO] B.S. Kerner and V.V. Osipov, *Autosolitons. A New Approach to Problems of Self-Organization and Turbulence*, Fund. Theories Phys. 61, Kluwer Academic, Boston, 1994.

[LSU] O.A. Ladyzenskaja, V.A. Solonnikov, and N.N. Uralceva, *Linear and Quasilinear Equations of Parabolic Type*, AMS, Providence, RI, 1968.

[M] J. Moser, *On a pointwise estimate for parabolic differential equations*, Comm. Pure Appl. Math., 24 (1971), pp. 727–740.

[SSX] P. Shi, M. Shillor, and X. Xu, *Existence of a solution to the Stefan problem with Joule's heating*, J. Differential Equations, 105 (1993), pp. 239–263.

[XA] H. Xie and W. Allegretto, *$C^\alpha(\overline{\Omega})$ solutions of a class of nonlinear degenerate elliptic systems arising in the thermister problem* , SIAM J. Math. Anal., 22 (1991), pp. 1491–1499.

[X1] X. Xu, *A strongly degenerate system involving an equation of parabolic type and an equation of elliptic type*, Comm. Partial Differential Equations, 18 (1993), pp. 199–213.

[X2] X. Xu, *Existence for a model arising from the IN Situ Vitrification process*, J. Math. Anal. Appl., 271 (2002), pp. 333–342.

[X3] X. Xu, *Local and global existence of continuous temperature in the electrical heating of conductors*, Houston J. Math., 22 (1996), pp. 435–455.

[X4] X. Xu, *Partial regularity of solutions to a class of degenerate systems*, Trans. Amer. Math. Soc., 349 (1997), pp. 1973–1992.

[X5] X. Xu, *A local partial regularity theorem for weak solutions of degenerate elliptic equations and its application to the thermistor problem*, Differential Integral Equations, 12 (1999), pp. 83–100.

[X6] X. Xu, *Local partial regularity theorems for suitable weak solutions of a class of degenerate systems*, Appl. Math. Optim., 34 (1996), pp. 299–324.

SIAM J. MATH. ANAL.
Vol. 35, No. 5, pp. 1099–1109

THE IMAGE OF A WEAKLY DIFFERENTIABLE MAPPING*

DAVID SWANSON† AND WILLIAM P. ZIEMER‡

Abstract. Let $\Omega \subset \mathbf{R}^n$ be an open ball, $n \geq 2$. Suppose that $f, g : \Omega \to \mathbf{R}^n$, $f = g$ on $\partial\Omega$, and that f is injective. In case f and g are continuous, then $f(\overline{\Omega}) \subset g(\overline{\Omega})$. We extend this result to generally discontinuous mappings belonging to suitable Sobolev spaces under appropriate notions of injectivity and boundary equality.

Key words. Sobolev mappings, topological degree, topological multiplicity, injectivity almost everywhere

AMS subject classifications. 46E35, 55M25, 74B99

DOI. 10.1137/S0036141002412069

1. Introduction.

Let $\Omega \subset \mathbf{R}^n$ be an open ball. It is widely known that if $f, g : \overline{\Omega} \to \mathbf{R}^n$ are continuous injective mappings which agree on $\partial\Omega$, then $f(\overline{\Omega}) = g(\overline{\Omega})$. In connection with their work in the theory of nonlinear elasticity, Müller, Spector, and Tang [11] considered the following related question:

> If $f \in W^{1,p}(\Omega; \mathbf{R}^n)$, d is a continuously differentiable diffeomorphism on a neighborhood of $\overline{\Omega}$, and f and d agree on $\partial\Omega$, what conditions will guarantee that $f(x) \in d(\overline{\Omega})$ for almost all $x \in \Omega$?

Here $W^{1,p}(\Omega; \mathbf{R}^n)$ is the class of all mappings $f : \Omega \to \mathbf{R}^n$ whose component functions f_i belong to the usual Sobolev space $W^{1,p}(\Omega)$, and the values of f on $\partial\Omega$ are in the sense of the trace operator; cf. [2, section 4.3].

Mappings in the Sobolev space $W^{1,p}(\Omega; \mathbf{R}^n)$, $n = 3$, are used in nonlinear elasticity to model physical deformations of matter. Regarding $f(\Omega)$ as a deformation of an elastic body Ω (in $\mathbf{R}^3$), it is natural that f should satisfy reasonable injectivity conditions. See, for example, the papers [10] and [11] and the references therein. Generally, such deformations f need not be homeomorphisms since cavities may appear in the body during the deformation and Ω, $f(\Omega)$ may be homologically distinct.

One such condition is that f is injective almost everywhere, which is the case if there exists a set $N \subset \Omega$ with $|N| = 0$ such that $f(A \setminus N) \cap f(B \setminus N) = \emptyset$ whenever A and B are disjoint sets in Ω. This corresponds to the physical notion that matter cannot interpenetrate itself. The question posed above concerns topological consequences of the injectivity conditions imposed on the mapping f. It seeks to find conditions which disallow the possibility that a nonnegligible volume of Ω will pass through the boundary during the deformation, again related to the physical hypothesis of noninterpenetrability.

Müller, Spector, and Tang [11] proved that if $f \in W^{1,p}(\Omega; \mathbf{R}^n)$ with $p > n - 1$, then in fact $f(x) \in d(\Omega)$ for almost all $x \in \Omega$ provided that

1. $f(x) = d(x)$ for $\mathcal{H}^1$ almost all $x \in \partial\Omega$,
2. there is a set $N \subset \Omega$ with $\mathcal{H}^1(N) = 0$ so that $f|_{\Omega \setminus N}$ is injective, and
3. $Jf(x) \neq 0$ for almost every $x \in \Omega$ with $Jf(x) > 0$ on a set of positive measure.

*Received by the editors July 24, 2002; accepted for publication (in revised form) March 28, 2003; published electronically January 6, 2004.
http://www.siam.org/journals/sima/35-5/41206.html
†Department of Mathematics, University of Louisville, Louisville, KY 40292 (david.swanson@louisville.edu).
‡Department of Mathematics, Indiana University, Bloomington, IN 47405 (ziemer@indiana.edu).

$\mathcal{H}^k$ denotes k-dimensional Hausdorff measure and Jf denotes the Jacobian of f.

The authors [14] showed that the conclusion above remains valid if (2) is replaced with the weaker condition $\mathcal{H}^{n-1}(N) = 0$, provided that N may be decomposed as

$$N = N_1 \cup N_2, \quad \mathcal{H}^1(N_1) = 0, \quad N_2 \text{ closed},$$

and that for (almost) every pair S_1, S_2 of spheres in Ω, the preimage of $f(S_2 \cap N)$ in S_1 has $\mathcal{H}^{n-2}$ measure zero. Roughly speaking, in the case of disjoint spheres S_1 and S_2, this states that $f(S_1)$ and $f(S_2)$ must have very little overlap.

The key to the argument in [14] is a variant of the Jordan separation theorem, which states that if $S \subset \Omega$ is a sphere and $f \in W^{1,p}(S; \mathbf{R}^n)$ with $p > n-1$ has an almost everywhere nonvanishing Jacobian, then $\mathbf{R}^n \setminus f(S)$ has precisely two components, and the bounded component has finite perimeter provided that there is a closed set $N \subset S$ for which $f|_{S \setminus N}$ is injective and

$$\mathcal{H}^{n-2}(S \cap f^{-1}(f(S \cap N))) = 0.$$

In general one cannot expect to obtain such a result if the exceptional set N has Hausdorff dimension exceeding $n-1$. Indeed, an example is given in section 5 of [11], with B denoting the unit ball in $\mathbf{R}^2$, of a mapping f satisfying

$$f \in \bigcap_{1 \le p < 2} W^{1,p}(B; \mathbf{R}^2), \quad f|_{\partial B} \equiv Id,$$

which is injective off a set $N \subset B$ with $0 < \mathcal{H}^1(N) < \infty$, which satisfies $Jf(x) > 0$ for almost all $x \in \Omega$ yet nevertheless carries a set of positive measure in B to the exterior of B.

In contrast to this example, in this paper we derive a result in which the exceptional set may have dimension n—that it is possible to assume only that $|N| = 0$, provided that additional integrability conditions are added to the derivatives of f. Our analysis requires that if $\{f_k\}$ is a sequence of regularizers of f, then the Jacobians of the f_k converge locally in L^1 to the Jacobian of f. That is,

$$\int_K |Jf_k - Jf| \, dx \to 0$$

for all compact sets $K \subset \Omega$. This is the case if $f \in W^{1,n}(\Omega; \mathbf{R}^n)$ and, more generally, if the component functions lie in appropriate (but possibly different) Sobolev spaces. For $\bar{p} = (p_1, p_2, \dots, p_n) \in \mathbf{R}^n$ with each $p_i \ge 1$ we define

$$W^{1,\bar{p}}(\Omega; \mathbf{R}^n) = \left\{ f = (f_1, f_2, \dots, f_n) : \Omega \to \mathbf{R}^n : f_i \in W^{1,p_i}(\Omega) \right\}.$$

Our interest lies in mappings $f \in W^{1,\bar{p}}(\Omega; \mathbf{R}^n)$, where $\bar{p}$ belongs to the class $\mathcal{B}$ defined as follows:

$$\mathcal{B} = \left\{ \bar{p} = (p_1, p_2, \dots, p_n) : p_i > n-1 \text{ for all } i \text{ and } \sum_{i=1}^{n} 1/p_i \le 1 \right\}.$$

Goffman and Ziemer [6] developed a theory of area for such mappings in which the area of f is given by the integral of its Jacobian. See Proposition 2.4 below.

We do not require d to be a diffeomorphism and we make no a priori assumption regarding the positivity of Jf. In the case of continuous mappings $f, g : \overline{\Omega} \to \mathbf{R}^n$

agreeing on $\partial\Omega$, the containment $f(\overline{\Omega}) \subset g(\overline{\Omega})$ is valid, provided only that f is injective. Our theorem may be regarded as an extension of this fact.

THEOREM 1.1. *Let $\Omega \subset \mathbf{R}^n$ be an open ball and let $f \in W^{1,\bar{p}}(\Omega;\mathbf{R}^n)$, where $\bar{p} \in \mathcal{B}$. If*

1. *there is a set $N \subset \Omega$ with $|N| = 0$ so that $f|_{\Omega\setminus N}$ is injective;*
2. *there is a continuous mapping $d : \overline{\Omega} \to \mathbf{R}^n$ belonging to $W^{1,\bar{p}}(\Omega;\mathbf{R}^n)$ so that $\mathrm{Tr} f = d$ on $\partial\Omega$; and*
3. *$Jf(x) \neq 0$ for almost all $x \in \Omega$,*

then $f(x) \in d(\overline{\Omega})$ for almost all $x \in \Omega$.

Remarks.

1. The trace operator Tr maps $W^{1,\bar{p}}(\Omega;\mathbf{R}^n)$ continuously to $L^{\bar{p}}(\partial\Omega;\mathbf{R}^n)$ endowed with the Hausdorff measure $\mathcal{H}^{n-1}$. Thus $\mathrm{Tr} f = d$ on $\partial\Omega$ is understood to mean that $\mathrm{Tr} f$ and d belong to the same $L^{\bar{p}}$ equivalence class and in particular coincide $\mathcal{H}^{n-1}$ almost everywhere on $\partial\Omega$.
2. Both the hypotheses and conclusion of Theorem 1.1 are independent of any particular choice of representative of f so it suffices to consider only a particular representative. In our development we will assume without loss of generality that f is p-quasicontinuous. For a thorough discussion of p-quasicontinuous representatives, see [1, Chapter 6].
3. If, in addition, $|d(\partial\Omega)| = 0$, then in fact $f(x) \in d(\Omega)$ for almost all $x \in \Omega$. For conditions under which d may satisfy $|d(\partial\Omega)| = 0$, see the papers of Malý [7] and Malý and Martio [8].
4. In section 4 we strengthen the conclusion of Theorem 1.1 and show that in fact $f(x)$ belongs to the topological image of $\overline{\Omega}$ under d for almost all $x \in \Omega$. This fact is used in section 5 to establish monotonicity of maps in $W^{1,\bar{p}}(\Omega;\mathbf{R}^n)$ for arbitrary open sets Ω.

For the remainder of the paper we consider cubes rather than balls. This is simply for technical convenience and provides no loss of generality.

2. Preliminaries. Let $Q \subset \mathbf{R}^n$ be a closed cube and let $f : Q \to \mathbf{R}^n$ be continuous. We denote by

$$\deg(f, Q, y)$$

the topological degree of f at a point $y \in \mathbf{R}^n \setminus f(\partial Q)$. We recall the following properties of the degree (cf. [12, II.2], [4, Chapters 1, 2]).

PROPOSITION 2.1. *Let f and Q be as above. Then $\deg(f, Q, \cdot)$ is an integer-valued function satisfying the following:*

1. *$\deg(f, Q, y)$ is defined for all $y \notin f(\partial Q)$ and is constant on the connected components of $\mathbf{R}^n \setminus f(\partial Q)$;*
2. *$y \notin f(Q)$ implies $\deg(f, Q, y) = 0$;*
3. *if $g : Q \to \mathbf{R}^n$ is continuous, $y \in \mathbf{R}^n$, and*

$$\sup_{z\in\partial Q} |f(z) - g(z)| < \mathrm{dist}(y, g(\partial Q)),$$

   then $\deg(f, Q, y) = \deg(g, Q, y)$;
4. *if $\{Q_j\}$ is a sequence of pairwise disjoint closed cubes contained in the interior of Q, $y \in \mathbf{R}^n$, and $f^{-1}(y)$ is contained in the union of the interiors of the Q_j, then*

$$\deg(f, Q, y) = \sum_j \deg(f, Q_j, y),$$

where at most finitely many terms in the sum are nonzero;

5. *if $f \in C^1(Q)$ and $y \notin f(\partial Q) \cap f(\{x : Jf(x) = 0\})$, then $f^{-1}(y)$ is a finite set and*

$$\deg(f, Q, y) = \sum_{x \in f^{-1}(y)} \operatorname{sgn} Jf(x).$$

Property 3 is a real variables counterpart to Rouché's theorem. In particular it implies that $\deg(f, Q, \cdot)$ is uniquely determined by $f|_{\partial Q}$, making it possible to define the degree for continuous maps $f : \partial Q \to \mathbf{R}^n$ as the degree of any continuous extension to the whole cube.

Let $\Omega \subset \mathbf{R}^n$ be an open cube. Recall that a mapping $f : \Omega \to \mathbf{R}^n$ is said to have an approximate differential L at a point $x \in \Omega$ if there is a set E with density 1 at x and a linear function $L : \mathbf{R}^n \to \mathbf{R}^n$ such that

$$\lim_{\substack{z \to x \\ z \in E}} \frac{|f(x) - f(z) - L(x - z)|}{|x - z|} = 0 \tag{2.1}$$

and that L is said to be regular if E may be written as a union of boundaries of concentric cubes centered at x. The following was obtained by Goffman and Ziemer [6].

PROPOSITION 2.2. *Let $f \in W^{1,p}(\Omega; \mathbf{R}^n)$, $p > n - 1$. Then f has a regular approximate differential coinciding with its distributional differential $Df(x)$ at almost all points $x \in \Omega$ and $f|_{\partial Q}$ is continuous for almost every closed cube $Q \subset \Omega$.*

Standard examples show that the condition $p > n - 1$ is necessary. Of course the continuity of a mapping is affected by the choice of representative in its L^p class, but the proposition is valid for any quasicontinuous representative.

The topological multiplicity of a mapping $f : \Omega \to \mathbf{R}^n$ is defined for $y \in \mathbf{R}^n$ by

$$M(f, \Omega, y) = \sup_{\mathcal{Q}} \sum_{Q \in \mathcal{Q}} |\deg(f, Q, y)|,$$

where the supremum is taken over all finite families $\mathcal{Q}$ of nonoverlapping closed cubes $Q \subset \Omega$ for which $f|_{\partial Q}$ is continuous and $y \notin f(\partial Q)$. The following proposition illustrates the relationship between Jf and $M(f, \Omega, \cdot)$.

PROPOSITION 2.3. *Let $f \in W^{1,p}(\Omega; \mathbf{R}^n)$, $p > n - 1$. Then $M(f, \Omega, f(x)) \geq 1$ for almost all $x \in \{Jf \neq 0\}$.*

Proof. Let x be a point where $Df(x)$ is a regular approximate differential of f and where $Jf(x) \neq 0$. Almost every $x \in \{Jf \neq 0\}$ has this property. We write $Q(x, r)$ for the cube of side-length $2r$ centered at x and define $g(z) = f(x) - Df(x) \cdot (x - z)$. For every $\varepsilon > 0$, Proposition 2.2 and (2.1) imply that $f|_{\partial Q(x,r)}$ is continuous and

$$\sup_{z \in \partial Q(x,r)} |f(z) - g(z)| = \sup_{z \in \partial Q(x,r)} |f(x) - f(z) - Df(x) \cdot (x - z)| < \varepsilon r$$

for infinitely many small $r > 0$. On the other hand, for arbitrary $z \in \partial Q(x, r)$ we may write

$$f(x) - g(z) = Df(x) \cdot (x - z);$$

hence

$$r \leq |x - z| = |Df(x)^{-1} \cdot (f(x) - g(z))| \leq |Df(x)^{-1}|\, |f(x) - g(z)|.$$

Taking the infimum over all $z \in \partial Q(x,r)$, this implies

$$|Df(x)^{-1}|\, r \le \operatorname{dist}(f(x), g(\partial Q(x,r))).$$

With $\varepsilon = |Df(x)^{-1}|$, we conclude that there exists $r > 0$ so that $Q(x,r) \subset \Omega$, $f|_{\partial Q(x,r)}$ is continuous, and

$$\sup_{z \in \partial Q(x,r)} |f(z) - g(z)| < \operatorname{dist}(f(x), g(\partial Q(x,r))).$$

Let $Q = Q(x,r)$. Proposition 2.1(3) implies that $\deg(f, Q, f(x)) = \deg(g, Q, f(x))$, and since g is an injective affine mapping, part 5 of the same proposition implies that

$$\deg(g, Q, f(x)) = \pm 1.$$

Therefore $|\deg(f, Q, f(x))| = 1$, implying that $M(f, \Omega, f(x)) \ge 1$ as desired. □

For our purposes, the importance of the topological multiplicity function lies in the following proposition, in which the area of a mapping f is given by the integral of its topological multiplicity. See Goffman and Ziemer [6] and Gariepy [5].

PROPOSITION 2.4. *Let $f \in W^{1,\bar{p}}(\Omega; \mathbf{R}^n)$ with $\bar{p} \in \mathcal{B}$. Let $\{f_k\}$ denote a sequence of regularizers of f. Then*

1. $\lim_{k\to\infty} \int_\Omega |Jf(x) - Jf_k(x)|\, dx = 0$;
2. $\lim_{k\to\infty} \int_{\mathbf{R}^n} |M(f, \Omega, y) - M(f_k, \Omega, y)|\, dy = 0$; *and*
3. $\int_\Omega |Jf(x)|\, dx = \int_{\mathbf{R}^n} |M(f, \Omega, y)|\, dy = L(f)$,

where $L(f)$ is the Lebesgue area of f.

3. Proof of Theorem 1.1. For $r > 0$ let Ω_r denote the open cube with side-length $2r$ centered at 0. We may assume without loss of generality that $\Omega = \Omega_1$. Let $f, d \in W^{1,\bar{p}}(\Omega; \mathbf{R}^n)$ and $N \subset \Omega$ satisfy the hypotheses of the theorem.

Step 1. There exists a sequence $\{E_k\}$ of disjoint measurable subsets of Ω with the property that $|\Omega \setminus \cup E_k| = 0$ and $f|_{E_k}$ is Lipschitz for all k. For functions $f \in W^{1,1}(\Omega)$ this follows from [3, Theorem 3.1.8] and the fact that f is approximately differentiable almost everywhere. The extension to mappings $f \in W^{1,1}(\Omega; \mathbf{R}^n)$ is immediate. Define $E = \cup E_j$. Replacing E by $E \setminus N$ we may assume that $f|_E$ is injective. By hypothesis we may assume that $Jf(x) \ne 0$ for all $x \in E$, and by Proposition 2.3 we may assume further that $M(f, \Omega, f(x)) \ge 1$ for every $x \in E$.

Step 2. Let $A \subset E$ be measurable. Applying the area formula for Lipschitz mappings [3, Corollary 3.2.20] we have

$$\int_{A \cap E_k} |Jf(x)|\, dx = \int_{\mathbf{R}^n} N(f, A \cap E_k, y)\, dy = |f(A \cap E_k)|$$

for every E_k, where $N(f, A \cap E_k, y)$ is the crude multiplicity function counting the number of solutions $x \in A \cap E_k$ to $f(x) = y$. Since Lipschitz mappings carry measurable sets to measurable sets, the monotone convergence theorem and the injectivity of f then imply

$$\int_A |Jf(x)|\, dx = |f(A)|. \tag{3.1}$$

In particular, $|f(A)| = 0$ if and only if $|A| = 0$ since $|Jf| > 0$ on E. Since this holds for all measurable $A \subset E$, the Radon–Nikodým theorem implies that for every $\varepsilon > 0$ there exists $\delta > 0$ so that

$$|A| < \varepsilon \text{ whenever } |f(A)| < \delta. \tag{3.2}$$

Step 3. Extend $d \in W^{1,\bar{p}}(\Omega;\mathbf{R}^n) \cap C(\overline{\Omega};\mathbf{R}^n)$ to a mapping (again denoted by d) in $W^{1,\bar{p}}(\Omega;\mathbf{R}^n) \cap C(\Omega_2;\mathbf{R}^n)$. This extension may be obtained using a reflection argument as in [9, Theorem 1.63]. Since Tr satisfies

$$\lim_{r\to 0} r^{-n} \int_{B(x,r)\cap\Omega} |f(y) - \mathrm{Tr} f(x)|\, dy = 0$$

for $\mathcal{H}^{n-1}$-a.e. $x \in \partial\Omega$ (see, for instance, Theorem 2 in [2, section 5.3]), the continuity of d and the fact $\mathrm{Tr} f = d$ on $\partial\Omega$ imply that

$$\lim_{r\to 0} r^{-n} \int_{B(x,r)\cap\Omega} |f(y) - d(y)|\, dy = 0 \tag{3.3}$$

for $\mathcal{H}^{n-1}$-a.e. $x \in \partial\Omega$. Now define

$$h(x) = \begin{cases} f(x) - d(x), & x \in \Omega, \\ 0, & x \in \mathbf{R}^n \setminus \Omega. \end{cases}$$

Clearly every component function of h belongs to $BV(\mathbf{R}^n)$ since Ω is a cube; cf. [16, Lemma 5.10.4]. At all points $x \in \partial\Omega$ where (3.3) holds we have

$$\lim_{r\to 0} \fint_{B(x,r)} |h(y)|\, dy = 0,$$

implying that each component function of h is approximately continuous at all such x. It follows from [3, Theorem 4.5.9 (30)] that the component functions are absolutely continuous on almost all lines in $\mathbf{R}^n$ parallel to the coordinate axes, and since these partial derivatives vanish outside Ω, we conclude that

$$h \in W^{1,\bar{p}}(\mathbf{R}^n;\mathbf{R}^n).$$

Now define

$$\tilde{f}(x) = \begin{cases} f(x), & x \in \Omega, \\ d(x), & x \in \Omega_2 \setminus \Omega. \end{cases}$$

Then $\tilde{f} \in W^{1,\bar{p}}(\Omega_2;\mathbf{R}^n)$ since

$$\tilde{f} = h + d \text{ on } \Omega_2.$$

For convenience denote $\tilde{f}$ again by f.

Step 4. Fix $1 < R < 2$. We will prove that

$$|\{x \in \Omega : f(x) \notin d(\Omega_R)\}| = 0.$$

Let $\varepsilon > 0$. Choose $\delta > 0$ satisfying (3.2) and choose $1 < r < R$ sufficiently close to 1 so that

$$\int_{\Omega_r\setminus\Omega} |Jf(x)|\, dx < \delta. \tag{3.4}$$

Let $\{f_k\}$ be a sequence of regularizers of f. Then

$$\lim_{k\to\infty} \int_{\mathbf{R}^n} |M(f,\Omega_r,y) - M(f_k,\Omega_r,y)|\, dy = 0$$

by Proposition 2.4. Thus there is a subsequence of the f_k, again denoted by the full sequence, with the property that $M(f_k, \Omega_r, y) \to M(f, \Omega_r, y)$ for almost all $y \in \mathbf{R}^n$ as $k \to \infty$.

Step 5. Define

$$X = \{y \in f(E) : M(f, \Omega_r, y) \geq 2\}.$$

Note that $M(f, \Omega_r, y) = 1$ for all $y \in f(E) \setminus X$ since $M(f, \Omega_r, f(x)) \geq 1$ whenever $x \in E$. By Proposition 2.4 we have

$$\int_{f(E)} M(f, \Omega_r, y)\, dy \leq \int_{\mathbf{R}^n} M(f, \Omega_r, y)\, dy = \int_{\Omega_r} |Jf|\, dx,$$

where

$$\int_{\Omega_r} |Jf|\, dx = \int_E |Jf|\, dx + \int_{\Omega_r \setminus \Omega} |Jf|\, dx < |f(E)| + \delta$$

by (3.1) and (3.4). On the other hand,

$$\begin{aligned} \int_{f(E)} M(f, \Omega_r, y)\, dy &= \int_{f(E)\setminus X} M(f, \Omega_r, y)\, dy + \int_X M(f, \Omega_r, y)\, dy \\ &\geq |f(E) \setminus X| + 2|X| \\ &= |f(E)| + |X|, \end{aligned}$$

from which it follows that $|X| < \delta$.

Step 6. For each $k \geq 1$ the classical area formula states

$$\int_{\mathbf{R}^n} N(f_k, \Omega_r, y)\, dy = \int_{\Omega_r} |Jf_k(x)|\, dx < \infty,$$

implying that $N(f_k, \Omega_r, y)$ is finite for almost all $y \in \mathbf{R}^n$. In light of Steps 4 and 5 there exists a set $G \subset f(E)$ with the property that

1. $|f(E) \setminus G| < \delta$;
2. $N(f_k, \Omega_r, y) < \infty$ for all $y \in G$ and $k \geq 1$; and
3. $M(f_k, \Omega_r, y) \to 1$ as $k \to \infty$ for all $y \in G$.

Since M is integer valued, note that property 3 states that for fixed $y \in G$, $M(f_k, \Omega_r, y) = 1$ for all sufficiently large k.

Now choose $y \in G$ and suppose that $y \notin d(\overline{\Omega}_r)$. Since $f|_{\partial\Omega_r}$ is continuous, the regularizers f_k converge uniformly to f on $\partial\Omega_r$, and so Proposition 2.1(3) implies

$$\deg(f_k, \Omega_r, y) = \deg(f, \Omega_r, y) = \deg(d, \Omega_r, y) = 0 \tag{3.5}$$

for all sufficiently large k. It is thus possible to fix k (depending on y) so that $M(f_k, \Omega_r, y) = 1$ and $\deg(f_k, \Omega_r, y) = 0$. By the definition of M there exists a finite family $\mathcal{Q}$ of closed cubes in Ω_r with the property that $y \notin f_k(\partial Q)$ for each $Q \in \mathcal{Q}$ and

$$\sum_{Q \in \mathcal{Q}} |\deg(f_k, Q, y)| = 1. \tag{3.6}$$

Since $y \in G$ implies that $f_k^{-1}(y)$ is a finite set, and since the sum in (3.6) cannot exceed 1 for any choice of the family $\mathcal{Q}$, we may assume (by adding finitely many additional

cubes Q to $\mathcal{Q}$ if necessary) that $f_k^{-1}(y)$ is contained in the union of the interiors of the cubes $Q \subset \mathcal{Q}$. Since deg is integer valued, this implies that $\deg(f_k, \Omega_r, y) = \pm 1$ for exactly one cube $Q \in \mathcal{Q}$ and is zero otherwise. Thus, by Proposition 2.1(4), we conclude

$$0 = \deg(f_k, \Omega_r, y) = \sum_{Q \in \mathcal{Q}} \deg(f_k, Q, y) = \pm 1,$$

a contradiction. Therefore $y \in d(\overline{\Omega}_r) \subset d(\Omega_R)$. We conclude that $G \subset d(\Omega_R)$, and hence that

$$|\{y \in f(E) : y \notin d(\Omega_R)\}| \leq |f(E) \setminus G| < \delta.$$

By (3.2) and the choice of δ, this implies

$$|\{x \in E : f(x) \notin d(\Omega_R)\}| < \varepsilon,$$

and since $\varepsilon > 0$ was arbitrary we have

$$|\{x \in \Omega : f(x) \notin d(\Omega_R)\}| = 0,$$

as desired.

Step 7. Now we may complete the argument. Let R_j be a sequence of real numbers decreasing to 1, and observe that

$$d(\overline{\Omega}) = \bigcap_{j=1}^{\infty} d(\Omega_{R_j}).$$

To see this, if z belongs to the intersection, then for every j there exists $x_j \in \Omega_{R_j}$ with $d(x_j) = z$. Since $\{x_j\}$ is a bounded sequence there exists a subsequence (denoted again by the full sequence) so that x_j converges to a point $x \in \overline{\Omega}$. The continuity of d implies that $d(x) = z$, implying that $z \in d(\overline{\Omega})$. The other containment is obvious. It follows that

$$\{x \in \Omega : f(x) \notin d(\overline{\Omega})\} = \bigcup_{j=1}^{\infty} \{x \in \Omega : f(x) \notin d(\Omega_{R_j})\},$$

so by the result of Step 6 we conclude that

$$|\{x \in \Omega : f(x) \notin d(\overline{\Omega})\}| = 0.$$

This completes the proof.

4. Topological image. Let $Q \subset \mathbf{R}^n$ be a closed cube and let $g : \partial Q \to \mathbf{R}^n$ be continuous. The topological image of Q under g is defined as

$$\mathrm{im}_T(g, Q) = g(\partial Q) \cup \{y \in \mathbf{R}^n \setminus g(\partial Q) : \deg(g, Q, y) \neq 0\}.$$

If $g : Q \to \mathbf{R}^n$ is continuous, then $\mathrm{im}_T(g, Q) \subset g(Q)$ and equality holds if g is a homeomorphism. We will show that the conclusion of Theorem 1.1 may be strengthened to

$$f(x) \in \mathrm{im}_T(d, \overline{\Omega}) \text{ for almost every } x \in \Omega.$$

Observe that (3.5) remains valid under the assumption that

$$y \notin \mathrm{im}_T(d, \overline{\Omega}_r)$$

rather than $y \notin d(\overline{\Omega}_r)$, so the contradiction obtained in Step 6 above allows us to conclude that $G \subset \mathrm{im}_T(d, \overline{\Omega}_r)$. Since this is valid for all sufficiently small r (dependent on ε and R) we obtain the following corollary to the proof of Theorem 1.1: for every $\varepsilon > 0$ there exists r_0 so that $1 < r < r_0$ implies

$$|\{x \in \Omega : f(x) \notin \mathrm{im}_T(d, \overline{\Omega}_r)\}| < \varepsilon. \tag{4.1}$$

We will not make use of Ω_R. The set of points $x \in \Omega$ with the property that $f(x) \notin \mathrm{im}_T(d, \overline{\Omega})$ may be written as

$$B = \{x \in \Omega : f(x) \notin d(\partial\Omega),\ \deg(d, \Omega, f(x)) = 0\}.$$

Let us assume for the sake of obtaining a contradiction that $|B| > 0$, and for every positive integer k define

$$B_k = B \cap \{x \in \Omega : \mathrm{dist}(f(x), d(\partial\Omega)) \geq k^{-1}\}.$$

Thus there is an index k for which $|B_k| > 0$.

For each $r > 1$ denote by ψ_r the homothetic transformation

$$\psi_r(x) = rx$$

and observe that

$$\deg(\psi_r, \Omega, p) = \begin{cases} 1, & p \in \Omega_r, \\ 0, & p \notin \overline{\Omega}_r. \end{cases}$$

It follows from the multiplication theorem for degree [4, Theorem 2.10] that

$$\deg(d \circ \psi_r, \Omega, p) = \deg(d, \Omega_r, p) \tag{4.2}$$

for all $p \notin d(\partial\Omega_r)$. Define $\varepsilon = |B_k|/2$ and choose $r > 1$ so that (4.1) holds and that

$$\sup_{z \in \partial\Omega} |d(z) - d \circ \psi_r(z)| < \frac{1}{2k}.$$

This is possible due to the uniform continuity of d. Let $x \in B_k$ and choose an arbitrary point $z \in \partial\Omega$. Then

$$\frac{1}{k} \leq |f(x) - d(z)| \leq |f(x) - d \circ \psi_r(z)| + |d(z) - d \circ \psi_r(z)| < |f(x) - d \circ \psi_r(z)| + \frac{1}{2k},$$

implying that $|f(x) - d \circ \psi_r(z)| > 1/(2k)$ for all $z \in \partial\Omega$. It follows that

$$\sup_{z \in \partial\Omega} |d(z) - d \circ \psi_r(z)| < \mathrm{dist}(f(x), d \circ \psi_r(\partial\Omega)).$$

Therefore $f(x) \notin d \circ \psi_r(\partial\Omega) = d(\partial\Omega_r)$, and Proposition 2.1 along with (4.2) above implies

$$\deg(d, \Omega_r, f(x)) = \deg(d \circ \psi_r, \Omega, f(x)) = \deg(d, \Omega, f(x)) = 0$$

for all $x \in B_k$. Thus

$$B_k \subset \{x \in \Omega : f(x) \notin \mathrm{im}_T(d, \Omega_r)\},$$

but according to (4.1) this implies

$$0 < |B_k| < \varepsilon = |B_k|/2,$$

which is the desired contradiction. It follows that $|B_k| = 0$, and hence that $|B| = 0$. We conclude that $f(x) \in \mathrm{im}_T(d, \overline{\Omega})$ for almost all $x \in \Omega$, as desired.

5. Monotonicity. In this section we assume that $\Omega \subset \mathbf{R}^n$ is an arbitrary open set.

THEOREM 5.1. *Let $f \in W^{1,\bar{p}}(\Omega;\mathbf{R}^n)$, $\bar{p} \in \mathcal{B}$, and suppose there is a set $N \subset \Omega$ so that $f|_{\Omega\setminus N}$ is injective and $Jf(x) \neq 0$ for all $x \in \mathbf{R}^n \setminus N$. Then for almost every cube $Q \subset \Omega$,*

$$f(x) \in \operatorname{im}_T(f, Q)$$

for almost all $x \in Q$.

Theorem 5.1 was proven for mappings in the class $W^{1,n}(\Omega;\mathbf{R}^n)$ by Vodop'yanov and Gol'dshtein [15], who refer to this property as monotonicity. A closely related concept is condition (INJ) introduced by Müller and Spector [10]. Their definition of the topological image is at a slight variance with ours, as it does not include the image of the boundary of the domain.

To prove the theorem we fix a point $a \in \Omega$ and let

$$r_a = \sup\{r > 0 : Q(a,r) \subset \Omega\}.$$

It suffices to show that for almost all r belonging to the interval $(0, r_a)$, the conclusion holds for the cube $Q(a,r)$. Since each $p_i > n-1$ we may appeal to [6, Theorem 3.2] to conclude that for almost all $r \in (0, r_a)$, the regularizers f_k of f converge uniformly to f on $\partial Q(a,r)$. Fix such an r. Then $f|_{\partial Q(a,r)}$ is continuous and may be extended to a function in $W^{1,\bar{p}}(Q(a,r);\mathbf{R}^n) \cap C(Q(a,r);\mathbf{R}^n)$ as follows. Let $\delta \in C^\infty(Q(a,r))$ satisfy

$$C_1 \operatorname{dist}(x, \partial Q(x,r)) \leq \delta(x) \leq C_2 \operatorname{dist}(x, \partial Q(a,r))$$

for all $x \in Q(a,r)$, where C_1 and C_2 are independent of x and f. Let $\phi \in C_0^\infty(B(0,1))$ be a regularizing kernel with the property that

$$P = \phi_\varepsilon * P$$

for every linear polynomial P. Writing

$$\psi_z(x) = \delta(x)^{-n} \phi\left(\frac{x-z}{\delta(x)}\right)$$

for $x \in \Omega$ and $z \in \mathbf{R}^n$, define

$$d(x) = \begin{cases} \displaystyle\int_{\mathbf{R}^n} \psi_z(x) f(z)\, dz, & x \in Q(a,r)^\circ, \\ f(x), & x \in \partial Q(a,r), \end{cases}$$

where $Q(a,r)^\circ$ denotes the topological interior of $Q(a,r)$. Then

$$d(x) \in W^{1,\bar{p}}(Q(a,r)^\circ;\mathbf{R}^n) \cap C(Q(a,r);\mathbf{R}^n)$$

(in fact, d is C^∞ on the interior) and satisfies

$$\lim_{r\to 0} r^{-n} \int_{B(x,r)\cap Q} |f(y) - d(y)|\, dy = 0$$

for all $x \in \partial\Omega$. In particular, $\operatorname{Tr} f = d$ on $\partial\Omega$. See [13], [16, Chapter 3] for details. Applying the result of the preceding section, we have

$$f(x) \in \operatorname{im}_T(d, Q(a,r)) = \operatorname{im}_T(f, Q(a,r))$$

for almost all $x \in Q(a,r)$, completing the proof of Theorem 5.1. □

REFERENCES

[1] D. R. Adams and L. I. Hedberg, *Function Spaces and Potential Theory*, Grundlehren Math. Wiss. 314, Springer-Verlag, Berlin, 1996.

[2] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992.

[3] H. Federer, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

[4] I. Fonseca and W. Gangbo, *Degree Theory in Analysis and Applications*, Oxford University Press, New York, 1995.

[5] R. Gariepy, *Multiplicity and the area of an* $(n-1)$ *continuous mapping*, Pacific J. Math., 47 (1973), pp. 427–433.

[6] C. Goffman and W. P. Ziemer, *Higher dimensional mappings for which the area formula holds*, Ann. Math., 92 (1970), pp. 482–488.

[7] J. Malý, *Sufficient conditions for change of variables in integral*, in Proceedings on Analysis and Geometry, Sobolev Institute Press, Novosibirsk, 2000, pp. 370–386.

[8] J. Malý and O. Martio, *Lusin's condition (N) and mappings of the class* $W^{1,n}$, J. Reine Angew. Math., 458 (1995), pp. 19–36.

[9] J. Malý and W. P. Ziemer, *Regularity of Solutions to Elliptic Partial Differential Equations*, Math. Surveys Monogr. 51, AMS, Providence, RI, 1997.

[10] S. Müller and S. Spector, *An existence theory for nonlinear elasticity which allows for cavitation*, Arch. Ration. Mech. Anal., 131 (1995), pp. 1–66.

[11] S. Müller, S. J. Spector, and Q. Tang, *Invertibility and a topological property of Sobolev maps*, SIAM J. Math. Anal., 27 (1996), pp. 959–976.

[12] T. Rado and P. V. Reichelderfer, *Continuous Transformations in Analysis*, Grundlehren Math. Wiss. 75, Springer-Verlag, New York, 1955.

[13] D. Swanson, *Pointwise inequalities and approximation in fractional Sobolev spaces*, Studia Math., 149 (2002), pp. 147–174.

[14] D. Swanson and W. P. Ziemer *A topological aspect of Sobolev mappings*, Calc. Var. Partial Differential Equations, 14 (2002), pp. 69–84.

[15] S. K. Vodop'yanov and V. M. Gol'dshtein, *Quasiconformal mappings and spaces of functions with generalized first derivatives*, Siberian Math. J., 17 (1976), pp. 399–411.

[16] W. P. Ziemer, *Weakly Differentiable Functions*, Grad. Texts in Math. 120, Springer-Verlag, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 35, No. 5, pp. 1110–1132

ON THE COMPRESSIBILITY OF OPERATORS IN WAVELET COORDINATES*

ROB STEVENSON†

Abstract. In [*Found. Comput. Math.*, 2 (2002), pp. 203–245], Cohen, Dahmen, and DeVore proposed an adaptive wavelet algorithm for solving operator equations. Assuming that the operator defines a boundedly invertible mapping between a Hilbert space and its dual, and that a Riesz basis of wavelet type for this Hilbert space is available, the operator equation can be transformed into an equivalent well-posed infinite matrix-vector system. This system is solved by an iterative method, where each application of the infinite stiffness matrix is replaced by an adaptive approximation. For a certain range of $s > 0$, determined by the compressibility of the stiffness matrix, i.e., by how well it can be approximated by sparse matrices, it was proven that if the errors of best linear combinations from the wavelet bases with N terms are $\mathcal{O}(N^{-s})$, then approximations yielded by the adaptive method with N terms also have errors of $\mathcal{O}(N^{-s})$, where their computation takes only $\mathcal{O}(N)$ operations. With the available estimates for both differential and singular integral operators, the compressibility of the stiffness matrix appears to limit the rate of convergence of the adaptive method, in the sense that for solutions that have a sufficiently high (Besov) regularity, these best N-term approximations converge with a better rate than can be shown for the approximations produced by the adaptive method. In this paper, considering piecewise smooth wavelets as spline or finite element wavelets, and using modified sparse matrix approximations, we derive improved results concerning compressibility. From these results it will follow that for the full range of s for which, under appropriate smoothness conditions, convergence of the best N-term approximations of $\mathcal{O}(N^{-s})$ can be shown, the adaptive method converges with that rate.

Key words. wavelets, matrix compression, differential operators, boundary integral operators, adaptivity

AMS subject classifications. 41A25, 47A20, 65F50, 65N30, 65N38

DOI. 10.1137/S0036141002411520

1. Introduction.

As was first observed in [BCR91], the stiffness matrix resulting from a Galerkin discretization of a singular integral operator, which, using standard single scale bases, is densely populated, turns out to be close to a sparse matrix when wavelet bases are exploited. Responsible for this phenomenon is the fact that the kernel of such an operator is increasingly smooth away from the diagonal, and that the wavelets have vanishing moments. Quantitative analyses in [Sch98, DHS02, vPS97] show that for wavelets that have a sufficient number of vanishing moments, the stiffness matrix can be compressed to a sparse one, whose application requires only $\mathcal{O}(n)$ operations, with n being the number of unknowns, whereas the order of convergence is maintained.

Compared to alternative approaches for compression such as panel clustering [HN89] and multipole expansions [GR87], the wavelet approach has the additional advantage that properly scaled wavelets generate Riesz bases for a range of Sobolev spaces. Therefore, in any case for strongly elliptic equations, if the operator defines a boundedly invertible mapping between a Sobolev space in this range and its dual, then the stiffness matrices with respect to the wavelet bases are well-conditioned uniformly

*Received by the editors July 19, 2002; accepted for publication (in revised form) May 23, 2003; published electronically January 6, 2004. This work was supported by the Netherlands Organization for Scientific Research and by the EC-IHP project "Breaking Complexity."
http://www.siam.org/journals/sima/35-5/41152.html
†Department of Mathematics, Utrecht University, P.O. Box 80.010, NL-3508 TA Utrecht, The Netherlands (stevenso@math.uu.nl).

in their sizes, allowing for a fast iterative solution. In summary, with suitable wavelets the discretization error accuracy can be realized in $\mathcal{O}(n)$ operations.

This Riesz basis property inspired Cohen, Dahmen, and DeVore in [CDD01, CDD02] to go one step further. Instead of first discretizing the problem, i.e., replacing the underlying infinite-dimensional space by some fixed finite-dimensional one, and then solving the resulting finite-dimensional system by some iterative method, they transformed the original problem into an equivalent well-posed infinite matrix-vector system. This system can be solved iteratively, where in each iteration the application of the infinite matrix has to be approximated. The main advantage of their approach is that in the course of the iteration the spaces in which the approximations are sought, which are always spanned by a finite linear combination of wavelets, will adapt to the solution in an optimal way. Because of this adaptivity, the method is attractive for both integral and differential equations in variational form.

In the following we assume that the problem to be solved has the form

$$Lu = g,$$

where for some closed subspace $\mathcal{H} \subset H^t$, being a Sobolev space of order $t \in \mathbb{R}$, the linear operator $L : \mathcal{H} \to \mathcal{H}'$ is boundedly invertible, the right-hand side $g \in \mathcal{H}'$, and thus the unknown solution $u \in \mathcal{H}$. With $\Psi = \{\psi_\lambda : \lambda \in \Lambda\}$ being a Riesz basis for $\mathcal{H}$ of wavelet type, the equivalent infinite matrix-vector problem reads as

$$\mathbf{Mu} = \mathbf{g}, \tag{1.1}$$

where $\mathbf{M} := \langle \Psi, L\Psi \rangle : \ell_2(\Lambda) \to \ell_2(\Lambda)$ is boundedly invertible, $\mathbf{g} := \langle \Psi, g \rangle \in \ell_2(\Lambda)$, with $\langle\, , \rangle$ denoting the duality product on $\mathcal{H} \times \mathcal{H}'$, and $u = \mathbf{u}^T \Psi$.

In [CDD01, CDD02], the quality of the proposed iterative method is assessed by comparing the $\ell_2(\Lambda)$-error of the obtained approximation with, say, N coefficients with that of a *best N-term approximation* for $\mathbf{u}$, i.e., a vector $\mathbf{u}_N$ with at most N nonzero coefficients that has distance to $\mathbf{u}$ less than or equal to that of any vector with a support of that size. Recall that since Ψ is a Riesz basis, the sizes of the error measured in $\ell_2(\Lambda)$- or H^t-metric differ by at most a constant factor.

In any case for wavelets that are sufficiently smooth, the theory of nonlinear approximation [DeV98, Coh00] shows that if *both*

$$0 < s < \frac{d-t}{n},$$

where n is the space dimension and d is the order of the wavelets, *and* u is in the Besov space $B^{sn+t}_\tau(L_\tau)$ with $\tau = (\frac{1}{2} + s)^{-1}$, then

$$\sup_{N \in \mathbb{N}} N^s \|\mathbf{u} - \mathbf{u}_N\| < \infty. \tag{1.2}$$

Here with the order of the wavelets we mean the order of the primal multiresolution analysis or, equivalently, the number of vanishing moments of the dual wavelets.

The attractive feature of these best N-term approximations is the fact that the condition involving Besov regularity is much milder that the corresponding condition $u \in H^{sn+t}$ involving Sobolev regularity that would be needed to guarantee the same rate of convergence with linear approximation in the spaces spanned by N wavelets on the coarsest scales. Indeed, assuming a sufficiently smooth right-hand side, for several boundary value problems it has been proven that the solution has a much higher Besov

regularity than Sobolev regularity [DD97, Dah99]. Note that with wavelets of order d, the maximum rate that can be expected by only imposing appropriate smoothness conditions on the solution is $\frac{d-t}{n}$.

Returning to the adaptive wavelet algorithm from [CDD02], besides a clean-up step that is applied after every K iterations to remove small coefficients in order to control the vector length, the other crucial ingredient is the adaptive way in which, in each iteration, the application of the infinite matrix $\mathbf{M}$ to a finitely supported vector is approximated. Given such a vector, before multiplication each column of $\mathbf{M}$ that corresponds to a nonzero entry in this vector is replaced by a finitely supported approximation within a tolerance that decreases as a function of the size of this coefficient. To prove results about complexity, information is needed about the number of entries that is necessary to approximate any column within some given tolerance. We recall the following definition from [CDD02].

DEFINITION 1.1. *$\mathbf{M}$ is called* s^*-compressible*, when for each $j \in \mathbb{N}$ there exists an infinite matrix $\tilde{\mathbf{M}}_j$ with at most $\alpha_j 2^j$ nonzero entries in each row and column with $\sum_{j\in\mathbb{N}} \alpha_j < \infty$ such that for any $s < s^*$, $\sum_{j\in\mathbb{N}} 2^{js}\|\mathbf{M} - \tilde{\mathbf{M}}_j\| < \infty$.*

An equivalent definition is obtained by requiring that for any $s < s^*$ and any $N \in \mathbb{N}$, there exists a matrix on distance of order N^{-s} having at most N nonzeros in each row and column.

The main theorem from [CDD02] now says that if (1.2) is valid for some s and $\mathbf{M}$ is s^*-compressible with $s^* > s$, then the number of arithmetic operations and storage locations used by the adaptive wavelet algorithm for computing an approximation for $\mathbf{u}$ within tolerance ϵ is of the order $\epsilon^{-1/s}$. Since in view of (1.2) the same order of operations is already needed to approximate $\mathbf{u}$ within this tolerance using best N-term approximations, assuming these would be available, this result shows that this solution method has *optimal computational complexity*.

It remains to determine the value of s^*. First of all, note that even for a differential operator, $\mathbf{M}$ itself is not sparse. Indeed, any two wavelets ψ_λ, $\psi_{\lambda'}$ from the infinite collection with $\mathrm{vol}(\operatorname{supp}\psi_\lambda \cap \operatorname{supp}\psi_\lambda) > 0$ give rise to a generally nonzero entry. Furthermore, in contrast to the nonadaptive setting, here we do not have the possibility for the matrix-vector multiplication to switch to a single-scale representation, which for differential operators would be sparse. On the other hand, it can be shown that for wavelets that have both vanishing moments and some global smoothness, the modulus of an entry decreases with increasing distance in scale of the involved wavelets. Assuming that for some $\sigma > 0$, L and its adjoint L' are bounded from $H^{t+\sigma} \to H^{-t+\sigma}$, by substituting the estimates [Dah97, eqn. (9.4.5), eqn. (9.4.8)] into [CDD01, Prop. 6.6.2] we infer that $\mathbf{M}$ is s^*-compressible with

$$s^* = \frac{\min\{t + \tilde{d}, \sigma, \gamma - t\}}{n} - \frac{1}{2}, \tag{1.3}$$

where $\tilde{d}$ is the order of the dual wavelets, and $\gamma = \sup_s\{\Psi \subset H^s\}$ (here we used that the condition $\sigma < t + \tilde{\gamma}$ imposed for [Dah97, eqn. (9.4.8)] can actually be relaxed to $\sigma \le t + \tilde{d}$). This result holds true for differential operators as well as for singular integral operators. Note that in contrast to the nonadaptive setting discussed at the beginning of this introduction, global smoothness of the wavelets is required.

The result (1.3), however, is not satisfactory. Indeed, since $\gamma < d$ and so $s^* \le \frac{\gamma-t}{n} - \frac{1}{2} < \frac{d-t}{n}$, on the basis of (1.3) optimal computational complexity of the adaptive wavelet method can be concluded only for solutions u that have limited Besov regularity. Indeed, when $u \in B^{sn+t}_\tau(L_\tau)$ with $\tau = (\frac{1}{2} + s)^{-1}$ and $s > s^*$, then the

best N-term approximations converge at a faster rate than can be shown for the approximations yielded by the adaptive wavelet method.

For the special case of L being the Laplace operator and spline wavelets, in [BBC+01, DDU02] it was proved that $s^* = \frac{\gamma - t}{n}$, which, however, is still less than $\frac{d-t}{n}$.

In this paper we consider piecewise smooth wavelets as spline or finite element wavelets. For both differential and singular integral operators, and for wavelets that have a sufficient global smoothness and have cancellation properties of a sufficiently high order, we will prove that $\mathbf{M} = \langle \Psi, L\Psi \rangle$ is actually s^*-compressible with $s^* > \frac{d-t}{n}$. This result shows that if for some arbitrary s from the full range $(0, \frac{d-t}{n}]$ the best N-term approximations convergence with errors of $\mathcal{O}(N^{-s})$, then so do the approximations yielded by the adaptive method. The key to obtaining this improved result is that we use slightly different sparse approximations and that on essential places we estimate directly norms of blocks of the matrix $\mathbf{M}$ instead of deriving such estimates in terms of the sizes of the individual entries via the Schur lemma.

This paper is organized as follows: In section 2 we prove s^*-compressibility with $s^* > \frac{d-t}{n}$ for differential operators on a domain.

In section 3 we prove this result for a class of singular integral operators on a sufficiently regular manifold, which includes operators resulting from applying the boundary integral method. Since the regularity of the manifold imposes a limit to both the smoothness of the wavelets and the continuity properties of the singular integral operators, depending on the other parameters it may restrict the compressibility. In section 3.3, we give a general proof of a decay estimate for entries corresponding to wavelets with disjoint supports, which so far was shown only in specific situations. In section 3.4, relying on techniques developed in [Sch98, DHS02], we prove a new decay estimate for entries corresponding to wavelets, or, more generally, corresponding to a linear combination of wavelets and another wavelet, that may have overlapping supports but for which the support of the linear combination has empty intersection with the singular support of the other wavelet. In contrast to the bounds from [Sch98, DHS02], this estimate benefits from global smoothness of the wavelets. Apart from its use in the analysis of adaptive schemes, this estimate also results in quantitatively better compression rates when applied in the analysis of nonadaptive schemes.

At the end of this introduction, we fix some notation. We always think of the space L_2 of all measurable square integrable functions on a domain Ω or manifold Γ as being equipped with the *standard* scalar product $\langle\, , \rangle$ and corresponding norm $\| \|$, defined by $\langle u, v \rangle = \int_\Omega u(x)v(x)dx$ or $\langle u, v \rangle = \int_\Gamma u(x)v(x)d\mu(x)$, with $d\mu$ being the induced Lebesgue measure.

For H a Hilbert space embedded in L_2 and any $u \in L_2$, the mapping $v \mapsto \langle v, u \rangle$ is continuous on H. This procedure defines an embedding of L_2 into H', or, equivalently, it fixes an interpretation of a function in L_2 as a functional in H'. Different scalar products on L_2, defining equivalent norms on L_2, give rise to different embeddings of L_2 into H', which may lead to nonequivalent H'-norms of L_2-functions (cf. [NS03, sect. 4]). This observation, together with the fact that on a few places in the wavelet literature nonstandard L_2-scalar are applied, is the reason we emphasize here our choice of the L_2-scalar product.

If H is dense in L_2, then the above embedding of L_2 into H' is dense, meaning that the L_2-scalar product restricted to $H \times L_2$ has an unique extension to the duality product on $H \times H'$. For such an H, without risk of confusion, we may use $\langle\, , \rangle$ to denote either product.

For any countable index set Λ, the notation $\langle\, , \rangle$ and $\| \|$ will also be used to denote

the standard scalar product and norm, as well as the resulting operator norm, on the space $\ell_2(\Lambda)$ of square summable scalar sequences.

Finally, in order to avoid the repeated use of generic but unspecified constants, by $C \lesssim D$ we mean that C can be bounded by a multiple of D, independently of parameters which C and D may depend on. Obviously, $C \gtrsim D$ is defined as $D \lesssim C$, and $C \eqsim D$ as $C \lesssim D$ and $C \gtrsim D$.

2. Compressibility of differential operators. For some domain $\Omega \subset \mathbb{R}^n$, $t \in \mathbb{N}_0$, and $\Gamma_D \subset \partial\Omega$, possibly with $\Gamma_D = \emptyset$, let

$$H^t_{0,\Gamma^D}(\Omega) = \mathrm{clos}_{H^t(\Omega)}\{u \in H^t(\Omega) \cap C^\infty(\Omega) : \mathrm{supp}\, u \cap \Gamma^D = \emptyset\},$$

and let $L : H^t_{0,\Gamma^D}(\Omega) \to (H^t_{0,\Gamma^D}(\Omega))'$ be defined by

$$\langle u, Lv\rangle = \sum_{|\alpha|,|\beta|\le t} \langle \partial^\alpha u, a_{\alpha\beta}\partial^\beta v\rangle,$$

where $a_{\alpha\beta} \in L_\infty(\Omega)$ so that L is bounded. Obviously L has an extension, which we will also denote by L, as a bounded operator from $H^t(\Omega) \to H^{-t}(\Omega)$. For completeness, $H^s(\Omega)$ for $s < 0$ denotes the dual of $H^{-s}(\Omega)$.

We assume that there exists a $\sigma > 0$ such that

$$L, L' : H^{t+\sigma}(\Omega) \to H^{t-\sigma}(\Omega) \quad \text{are bounded.} \tag{2.1}$$

It is sufficient that for arbitrary $\varepsilon > 0$ and all α, β with $\min\{|\alpha|, |\beta|\} > m - \sigma$, the following holds:

$$a_{\alpha\beta} \in \begin{cases} W_\infty^{\sigma-m+\min\{|\alpha|,|\beta|\}}(\Omega) & \text{when } \sigma \in \mathbb{N}, \\ C^{\sigma-m+\min\{|\alpha|,|\beta|\}+\varepsilon}(\Omega) & \text{when } \sigma \notin \mathbb{N}. \end{cases}$$

Let

$$\Psi = \{\psi_\lambda : \lambda \in \Lambda\}$$

be a *Riesz basis for* $H^t_{0,\Gamma^D}(\Omega)$ of wavelet type. The index λ encodes both the level, denoted by $|\lambda| \in \mathbb{N}_0$, and the location of the wavelet ψ_λ.

We assume that the wavelets are *local* in the sense that

$$\mathrm{diam}(\mathrm{supp}\,\psi_\lambda) \lesssim 2^{-|\lambda|} \quad \text{and} \quad \sup_{x\in\Omega,\ell\in\mathbb{N}_0} \#\{|\lambda| = \ell : B(x; 2^{-\ell}) \cap \mathrm{supp}\,\psi_\lambda\} < \infty,$$

and that they are *piecewise smooth*, with which we mean that $\mathrm{supp}\,\psi_\lambda \backslash \mathrm{sing\,supp}\,\psi_\lambda$ is the disjoint union of m open "uniformly Lipschitz" domains $\Xi_{\lambda,1}, \dots, \Xi_{\lambda,m}$, with $\cup_{i=1}^m \overline{\Xi_{\lambda,i}} = \mathrm{supp}\,\psi_\lambda$, and that $\psi_\lambda|_{\Xi_{\lambda,i}}$ is smooth with, for any $\beta \in \mathbb{N}_0^n$,

$$\sup_{x\in\Xi_{\lambda,i}} |\partial^\beta \psi_\lambda(x)| \lesssim 2^{(|\beta|+\frac{n}{2}-t)|\lambda|}. \tag{2.2}$$

Examples of piecewise smooth wavelets are tensor products of univariate spline wavelets, or finite element wavelets subordinate to a subdivision of the domain into n-simplices.

Remark 2.1. Precisely, we will call a collection of domains $\{A_\nu\} \subset \mathbb{R}^n$ uniformly Lipschitz domains when there exist affine mappings B_ν with $|DB_\nu| \lesssim \mathrm{vol}(A_\nu)^{-1}$ and $|(DB_\nu)^{-1}| \lesssim \mathrm{vol}(A_\nu)$ such that the sets $B_\nu(A_\nu)$ satisfy the condition of "minimal

smoothness" [Ste70, Chap. VI, sect. 3] with uniform parameters "ε," "N," and "M." Examples are given of the interiors of "nondegenerate" polygons.

For minimally smooth domains it is known [Ste70] that there exist universal extension operators, with Sobolev norms dependent only on the aforementioned parameters. So, in particular, by first transforming $\psi_\lambda|_{\Xi_{\lambda,i}}$ using such an affine mapping to a function on a minimally smooth domain with volume one, then making the extension, and finally applying the inverse transformation, we conclude by (2.2) that there exists a smooth function $\varphi_{\lambda,i}$ on $\mathbb{R}^n$, equal to ψ_λ on $\Xi_{\lambda,i}$, with, for any $s \geq 0$, $p \in [1,\infty]$,

$$\|\varphi_{\lambda,i}\|_{W_p^s(\mathbb{R}^n)} \lesssim 2^{(s-t+\frac{n}{2}-\frac{n}{p})|\lambda|}.$$

Furthermore, we assume that there exist $\gamma > t$, $\tilde{d} > -t$ such that for $r \in [-\tilde{d}, \gamma)$, $s < \gamma$,

$$\|\cdot\|_{H^r(\Omega)} \lesssim 2^{\ell(r-s)} \|\cdot\|_{H^s(\Omega)} \qquad \text{on } W_\ell := \operatorname{span}\{\psi_\lambda : |\lambda| = \ell\}. \tag{2.3}$$

Remark 2.2. It is known that the above wavelet assumptions are satisfied by biorthogonal wavelets when the primal and dual spaces have regularity indices $\gamma > \max\{0,t\}$, $\tilde{\gamma} > \max\{0,-t\}$ and orders $d > \gamma$, $\tilde{d} > \tilde{\gamma}$, respectively (cf. [Dah96, DS99c]), the primal spaces consist of "piecewise" smooth functions, and, finally, no boundary conditions are imposed on the dual spaces ("complementary boundary conditions"; see [DS98]). In particular, (2.3) for $r \in [-\tilde{d}, -\tilde{\gamma}]$ can be deduced from the lines following (A.2) in [DS99c]. In Remark 2.5 we will comment on the case when the dual spaces satisfy the same boundary conditions as the primal ones.

Let us consider two wavelets $\psi_{\lambda'}$, ψ_λ with $\operatorname{vol}(\operatorname{supp}\psi_{\lambda'} \cap \operatorname{supp}\psi_\lambda) > 0$ and $|\lambda'| \gg |\lambda|$. Then on the scale of the size of $\operatorname{supp}\psi_{\lambda'}$, ψ_λ and so by the continuity assumption (2.1), $L\psi_\lambda$ are smooth, and therefore the latter function can be well approximated by a polynomial that integrated against $\psi_{\lambda'}$ vanishes because of its vanishing moments, showing that the entry $\langle \psi_{\lambda'}, L\psi_\lambda\rangle$ is small. The quantification of the smoothness of ψ_λ is governed by the parameter γ, whereas the number of vanishing moments of $\psi_{\lambda'}$ is given by $\tilde{d}$. Since $t > 0$, usually $t + \tilde{d} \geq \gamma - t$, with which γ becomes the restricting factor in compression results. To relax the restriction imposed by γ, in the following theorem we will make use of the fact that most of these $\psi_{\lambda'}$ will have their support inside some patch $\Xi_{\lambda,i}$ on which ψ_λ is infinitely smooth, and that only for $\psi_{\lambda'}$ with supports that intersect the $(n-1)$-dimensional singular support of ψ_λ, the estimate of the corresponding entry has to rely on the global smoothness parameter γ. Furthermore, to obtain sharp results, instead of estimating individual entries, we will directly estimate norms of blocks of the infinite matrix containing all entries that will be dropped.

THEOREM 2.3. *Let* $\mathbf{M} = \langle \Psi, L\Psi\rangle$. *For* $j \in \mathbb{N}$, *and with*

$$k(j,n) := \frac{j}{n-1} \quad \textit{when } n > 1,$$

and $j \leq k(j,1) \leq 2^j$, $k(j,1) > j\frac{\min\{t+\tilde{d},\sigma\}}{\gamma - t}$, *we define the infinite matrix* $\mathbf{M}_j$ *by replacing all entries* $\mathbf{M}_{\lambda,\lambda'} = \langle \psi_\lambda, L\psi_{\lambda'}\rangle$ *by zeros when*

$$\big||\lambda| - |\lambda'|\big| > k(j,n), \quad \textit{or} \tag{2.4}$$

$$\big||\lambda| - |\lambda'|\big| > \tfrac{j}{n} \quad \textit{and for some } 1 \leq i \leq m, \begin{cases} \operatorname{supp}\psi_\lambda \subset \overline{\Xi_{\lambda',i}} & \textit{when } |\lambda| > |\lambda'|, \\ \operatorname{supp}\psi_{\lambda'} \subset \overline{\Xi_{\lambda,i}} & \textit{when } |\lambda| < |\lambda'|. \end{cases} \tag{2.5}$$

Then the number of nonzero entries in each row and column of $\mathbf{M}_j$ *is of order* 2^j*, and for any*

$$s \le \min\{\tfrac{t+\tilde{d}}{n}, \tfrac{\sigma}{n}\}, \text{ with } s < \tfrac{\gamma-t}{n-1} \text{ when } n > 1,$$

it holds that $\|\mathbf{M} - \mathbf{M}_j\| \lesssim 2^{-sj}$.

Remark 2.4. In view of Definition 1.1, by taking $\tilde{\mathbf{M}}_j = \mathbf{M}_{\lceil j+\log(\alpha_j)\rceil}$, with, for example, $\alpha_j = j^{-(1+\varepsilon)}$ for some $\varepsilon > 0$, we infer that $\mathbf{M}$ is s^*-compressible with

$$s^* = \min\{\tfrac{t+\tilde{d}}{n}, \tfrac{\sigma}{n}, \tfrac{\gamma-t}{n-1}\}$$

($s^* = \min\{t+\tilde{d}, \sigma\}$ when $n = 1$). So, with d being the order of the wavelets, if $\tilde{d} > d-2t$, $\sigma > d-t$ and when $n > 1$, $\frac{\gamma-t}{n-1} > \frac{d-t}{n}$, then indeed $s^* > \frac{d-t}{n}$. The condition involving γ when $n > 1$ is satisfied, for instance, when $\frac{d-t}{n} > \frac{1}{2}$ and $\gamma = d - \frac{1}{2}$ (spline wavelets).

Proof of Theorem 2.3. Let λ be some given index. By the locality of the wavelets, the number of indices λ' with fixed $|\lambda'|$ for which $\mathrm{vol}(\mathrm{supp}\,\psi_{\lambda'} \cap \mathrm{supp}\,\psi_\lambda) > 0$ is of order $\max\{1, 2^{(|\lambda'|-|\lambda|)n}\}$. By using in addition the piecewise smoothness of the wavelets, the number of indices λ' with fixed $|\lambda'| \ge |\lambda|$ for which $\mathrm{vol}(\mathrm{supp}\,\psi_{\lambda'} \cap \mathrm{supp}\,\psi_\lambda) > 0$ and $\mathrm{supp}\,\psi_{\lambda'}$ is not contained in some $\overline{\Xi_{\lambda,i}}$ is of order $2^{(|\lambda'|-|\lambda|)(n-1)}$. We conclude that the number of nonzero entries in the λth row and column of $\mathbf{M}_j$ is of order

$$\sum_{||\lambda'|-|\lambda||\le \frac{j}{n}} \max\{1, 2^{(|\lambda'|-|\lambda|)n}\} + \sum_{\frac{j}{n}<||\lambda'|-|\lambda||\le k(j,n)} \max\{1, 2^{(|\lambda'|-|\lambda|)(n-1)}\} \eqsim 2^j.$$

Let $\hat{\mathbf{M}}_j$ be defined by

$$(\mathbf{M} - \hat{\mathbf{M}}_j)_{\lambda,\lambda'} = \begin{cases} \mathbf{M}_{\lambda,\lambda'} & \text{when } \big||\lambda| - |\lambda'|\big| > k(j,n), \\ 0 & \text{otherwise.} \end{cases}$$

The continuity assumptions on L, L', together with (2.3), show that for

$$r \in (0, t+\tilde{d}] \cap (0, \sigma] \cap (0, \gamma - t)$$

and $w_\ell \in W_\ell$, $w_{\ell'} \in W_{\ell'}$,

$$\begin{aligned} |\langle w_\ell, Lw_{\ell'}\rangle| &\lesssim \|w_\ell\|_{H^{t-r}(\Omega)} \|Lw_{\ell'}\|_{H^{-t+r}(\Omega)} \\ &\lesssim \|w_\ell\|_{H^{t-r}(\Omega)} \|w_{\ell'}\|_{H^{t+r}(\Omega)} \lesssim 2^{r(\ell'-\ell)} \|w_\ell\|_{H^t(\Omega)} \|w_{\ell'}\|_{H^t(\Omega)}, \end{aligned} \tag{2.6}$$

and, analogously, $|\langle w_\ell, Lw_{\ell'}\rangle| = |\langle L'w_\ell, w_{\ell'}\rangle| \lesssim 2^{r(\ell-\ell')} \|w_\ell\|_{H^t(\Omega)} \|w_{\ell'}\|_{H^t(\Omega)}$. So for arbitrary $\mathbf{c}, \mathbf{d} \in \ell_2(\Lambda)$, we have

$$\begin{aligned} |\langle \mathbf{c}, (\mathbf{M} - \hat{\mathbf{M}}_j)\mathbf{d}\rangle| &= \Big| \sum_{|\ell-\ell'|>k(j,n)} \Big\langle \sum_{|\lambda|=\ell} \mathbf{c}_\lambda \psi_\lambda, L \sum_{|\lambda'|=\ell'} \mathbf{d}_{\lambda'} \psi_{\lambda'} \Big\rangle \Big| \\ &\lesssim \sum_{|\ell-\ell'|>k(j,n)} 2^{-r|\ell-\ell'|} \Big\| \sum_{|\lambda|=\ell} \mathbf{c}_\lambda \psi_\lambda \Big\|_{H^t(\Omega)} \Big\| \sum_{|\lambda'|=\ell'} \mathbf{d}_{\lambda'} \psi_{\lambda'} \Big\|_{H^t(\Omega)} \\ &\lesssim 2^{-k(j,n)r} \sqrt{\sum_\ell \Big\| \sum_{|\lambda|=\ell} \mathbf{c}_\lambda \psi_\lambda \Big\|^2_{H^t(\Omega)}} \sqrt{\sum_{\ell'} \Big\| \sum_{|\lambda'|=\ell'} \mathbf{d}_{\lambda'} \psi_{\lambda'} \Big\|^2_{H^t(\Omega)}} \\ &\eqsim 2^{-k(j,n)r} \|\mathbf{c}\| \|\mathbf{d}\|, \end{aligned}$$

or $\|\mathbf{M}-\hat{\mathbf{M}}_j\| \lesssim 2^{-k(j,n)r}$.

Finally, we analyze the error as a consequence of dropping entries with indices that satisfy criterion (2.5). For $\lambda \in \Lambda$, $1 \le i \le m$, and $\ell > |\lambda'|$, let

$$A_{\ell,\lambda',i} := \{|\lambda| = \ell : \operatorname{supp}\psi_\lambda \subset \overline{\Xi_{\lambda',i}}\}.$$

For $w_{\ell,\lambda',i} \in \operatorname{span}\{\psi_\lambda : \lambda \in A_{\ell,\lambda',i}\} \subset W_\ell$ and

$$q \in (0, t+\tilde{d}] \cap (0,\sigma],$$

we will prove that

$$|\langle w_{\ell,\lambda',i}, L\psi_{\lambda'}\rangle| \lesssim 2^{q(|\lambda'|-\ell)}\|w_{\ell,\lambda',i}\|_{H^t(\Omega)}. \tag{2.7}$$

Because of (2.6), it is sufficient to consider $q \ge \gamma - t \ge -t$. From Remark 2.1, recall that $\varphi_{\lambda',i}$ is the extension of $\psi_{\lambda'}|_{\Xi_{\lambda',i}}$ to a smooth function on $\mathbb{R}^n$ with, for $s \ge 0$, $\|\varphi_{\lambda',i}\|_{H^s(\mathbb{R}^n)} \lesssim 2^{(s-t)|\lambda'|}$. From the locality and continuity of L, we conclude that

$$\begin{aligned}|\langle w_{\ell,\lambda',i}, L\psi_{\lambda'}\rangle| = |\langle w_{\ell,\lambda',i}, L\varphi_{\lambda',i}\rangle| &\lesssim \|w_{\ell,\lambda',i}\|_{H^{t-q}(\Omega)}\|L\varphi_{\lambda',i}\|_{H^{-t+q}(\Omega)}\\ &\lesssim \|w_{\ell,\lambda',i}\|_{H^{t-q}(\Omega)}\|\varphi_{\lambda',i}\|_{H^{t+q}(\Omega)} \lesssim 2^{-q(\ell-|\lambda'|)}\|w_{\ell,\lambda',i}\|_{H^t(\Omega)}.\end{aligned}$$

For any $\mathbf{c},\mathbf{d} \in \ell_2(\Lambda)$ and $q \in (0, t+\tilde{d}] \cap (0,\sigma]$, from (2.7) we have

$$\begin{aligned}&\Big|\sum_{\frac{j}{n}<\ell-\ell'\le k(j,n)}\sum_{|\lambda'|=\ell'}\mathbf{d}_{\lambda'}\Big\langle\sum_{i=1}^m\sum_{\lambda\in A_{\ell,\lambda',i}}\mathbf{c}_\lambda\psi_\lambda, L\psi_{\lambda'}\Big\rangle\Big|\\
&\lesssim \sum_{\frac{j}{n}<\ell-\ell'\le k(j,n)}\sum_{|\lambda'|=\ell'}|\mathbf{d}_{\lambda'}|2^{-q(\ell-\ell')}\sum_{i=1}^m\Big\|\sum_{\lambda\in A_{\ell,\lambda',i}}\mathbf{c}_\lambda\psi_\lambda\Big\|_{H^t(\Omega)}\\
&\lesssim \sum_{\frac{j}{n}<\ell-\ell'\le k(j,n)}2^{-q(\ell-\ell')}\sqrt{\sum_{|\lambda'|=\ell'}|\mathbf{d}_{\lambda'}|^2}\sqrt{\sum_{|\lambda'|=\ell'}\Big(\sum_{i=1}^m\sqrt{\sum_{\lambda\in A_{\ell,\lambda',i}}|\mathbf{c}_\lambda|^2}\Big)^2}\\
&\lesssim \sum_{\frac{j}{n}<\ell-\ell'\le k(j,n)}2^{-q(\ell-\ell')}\sqrt{\sum_{|\lambda'|=\ell'}|\mathbf{d}_{\lambda'}|^2}\sqrt{\sum_{|\lambda|=\ell}|\mathbf{c}_\lambda|^2} \lesssim 2^{-\frac{j}{n}q}\|\mathbf{d}\|\|\mathbf{c}\|,\end{aligned}$$

where for the last line we have used that for fixed $|\lambda'|$, each λ is contained in at most a uniformly bounded number of sets $A_{|\lambda|,\lambda',i}$.

Since, analogous to (2.7), $|\langle\psi_\lambda, Lw_{\ell',\lambda,i}\rangle| = |\langle L'\psi_\lambda, w_{\ell',\lambda,i}\rangle| \lesssim 2^{-q(\ell'-|\lambda|)}\|w_{\ell',\lambda,i}\|_{H^t(\Omega)}$ when $w_{\ell',\lambda,i} \in \operatorname{span}\{\psi_{\lambda'} : \lambda' \in A_{\ell',\lambda,i}\}$, and so

$$\Big|\sum_{\frac{j}{n}<\ell'-\ell\le k(j,n)}\sum_{|\lambda|=\ell}\mathbf{c}_\lambda\Big\langle\psi_\lambda, L\sum_{i=1}^m\sum_{\lambda'\in A_{\ell',\lambda,i}}\mathbf{d}_{\lambda'}\psi_{\lambda'}\Big\rangle\Big| \lesssim 2^{-\frac{j}{n}q}\|\mathbf{c}\|\|\mathbf{d}\|,$$

we conclude that $\|\hat{\mathbf{M}}_j - \mathbf{M}_j\| \lesssim 2^{-\frac{j}{n}q}$.

A combination of the estimates for $\mathbf{M}-\hat{\mathbf{M}}_j$ and $\hat{\mathbf{M}}_j-\mathbf{M}_j$ shows that for $s \le \min\{\frac{t+\tilde{d}}{n}, \frac{\sigma}{n}\}$, with $s < \frac{\gamma-t}{n-1}$ when $n>1$, it holds that $\|\mathbf{M}-\mathbf{M}_j\| \lesssim 2^{-sj}$. ∎

Remark 2.5. Let us consider the situation that $t \in \mathbb{N}$, $\Gamma_D \ne \emptyset$, and that Ψ is a biorthogonal basis for $H^t_{0,\Gamma_D}(\Omega)$, where now also the dual spaces satisfy homogeneous

Dirichlet boundary conditions on Γ_D. Then since for $s \geq 0$, $H^s(\Omega) \cap H^t_{0,\Gamma_D}(\Omega)$ is only dense in $H^s(\Omega)$ when $s < \frac{1}{2}$, we can expect (2.3) only for $r \in [-\tilde{d}, \gamma)$, $s < \gamma$ with $r > -\frac{1}{2}$. As a consequence, in the proof of Theorem 2.3, the range of r for which (2.6) holds is restricted to $r \in (0, t + \tilde{d}] \cap (0, \sigma] \cap (0, \gamma - t) \cap (0, t + \frac{1}{2})$.

The same problems are encountered for proving (2.7). However, instead of restricting the range of q, here another solution is possible. The homogeneous Dirichlet boundary conditions on the dual spaces affect only wavelets with supports near Γ_D. More precisely, one can expect that there exists a constant $\theta > 0$ such that for any $r \in [-\tilde{d}, \gamma)$, $s < \gamma$,

(2.8)
$$\|\cdot\|_{H^r(\Omega)} \lesssim 2^{\ell(r-s)} \|\cdot\|_{H^s(\Omega)} \quad \text{on } \operatorname{span}\{\psi_\lambda : |\lambda| = \ell,\ \operatorname{dist}(\operatorname{supp}\psi_\lambda, \Gamma_D) \geq \theta 2^{-|\lambda|}\}.$$

Let us now add to the dropping criterion (2.5) the condition that $\operatorname{dist}(\operatorname{supp}\psi_\lambda, \Gamma_D) \geq \theta 2^{-|\lambda|}$ when $|\lambda| > |\lambda'|$, or $\operatorname{dist}(\operatorname{supp}\psi_{\lambda'}, \Gamma_D) \geq \theta 2^{-|\lambda'|}$ when $|\lambda| < |\lambda'|$. Then it is easily verified that the resulting $\mathbf{M}_j$, although a little bit less sparse, still has at most order 2^j nonzero entries in each row and column. On the other hand, changing the definition of $A_{\ell,\lambda',i}$ into

$$A_{\ell,\lambda',i} := \{|\lambda| = \ell : \operatorname{supp}\psi_\lambda \subset \overline{\Xi_{\lambda',i}},\ \operatorname{dist}(\operatorname{supp}\psi_\lambda, \Gamma_D) \geq \theta 2^{-|\lambda|}\}$$

for $w_{\ell,\lambda',i} \in \operatorname{span}\{\psi_\lambda : \lambda \in A_{\ell,\lambda',i}\}$ and $q \in (0, t + \tilde{d}] \cap (0, \sigma)$, using (2.8) we can prove that $|\langle w_{\ell,\lambda',i}, L\psi_{\lambda'}\rangle| \lesssim 2^{-q(\ell - |\lambda'|)} \|w_{\ell,\lambda',i}\|_{H^t(\Omega)}$. By copying the remainder of the proof, and for $k(j,1) > j \frac{\min\{t+\tilde{d},\sigma\}}{\min\{\gamma - t, t + \frac{1}{2}\}}$, we conclude that for

$$s \leq \min\{\tfrac{t+\tilde{d}}{n}, \tfrac{\sigma}{n}\}, \quad \text{with } s < \min\{\tfrac{\gamma - t}{n-1}, \tfrac{t+\frac{1}{2}}{n-1}\} \text{ when } n > 1,$$

it holds that $\|\mathbf{M} - \mathbf{M}_j\| \lesssim 2^{-sj}$.

Remark 2.6. As follows from Remark 2.4, to show s^*-compressibility with $s^* > \frac{d-t}{n}$, it will be necessary that both $\tilde{d}$ and γ increase linearly as a function of d. To benefit from an often much higher regularity of the solution in Besov than in Sobolev scale, we are mainly interested in applying the adaptive method with a relatively large value of $d - t$. Indeed, for small $d - t$, the adaptive method can give at most a small improvement in the order of convergence compared to nonadaptive methods, which in practice might not compensate for the overhead they require.

In general, nontensor product domains, smooth wavelets, i.e., with large values of γ, are not easy to construct. The approach from [DS99b], based on a nonoverlapping domain decomposition, yields wavelet bases that in principal for any d satisfy all requirements concerning smoothness and cancellation properties to obtain $s^* > \frac{d-t}{n}$. Other approaches based on nonoverlapping domain decomposition (see [DS99a, CTU99, CM00]) yield wavelets which over the interfaces between subdomains are only continuous. Note that although for nonadaptive wavelet methods the fact that wavelets along some lower-dimensional interface are less smooth or have reduced cancellation properties might not influence the overall complexity-accuracy balance, for adaptive methods it generally does. Indeed, it might happen that the solution is smooth everywhere except exactly along that interface, meaning that the adaptive method mainly produces coefficients corresponding to wavelets with degenerated properties. Also finite element wavelets as constructed in [DS99c, CES00, Ste00] are only continuous. For example, for $t = 1$ and $n = 2$, with continuous wavelets only for orders $d \leq 2$, $s^* \geq \frac{d-t}{n}$ can be shown.

Still having in mind a nonoverlapping decomposition of the domain into a number of subdomains or patches $\Omega_1, \ldots, \Omega_M$, as an alternative it seems not too difficult, in any case, if one refrains from having local dual wavelets, which are not needed here anyway, to construct wavelets of some given order d, which restricted to each patch are again wavelets characterized by parameters $\gamma \leq d - \frac{1}{2}$ and $\tilde{d}$ ("*patchwise*" smoothness and cancellation properties) that can be chosen at one's convenience. If, in addition, these wavelets have a sufficient global smoothness such that they generate a Riesz basis for $H^t_{0,\Gamma_D}(\Omega)$, then $\langle \Psi, L\Psi \rangle = \sum_q \langle \Psi|_{\Omega_q}, L\Psi|_{\Omega_q} \rangle$. Theorem 2.3 now directly applies to the matrices in the sum with conditions in terms of the "local" γ and $\tilde{d}$, and so when these are sufficiently large, $s^* > \frac{d-t}{n}$ follows.

In [Ste02] we generalized the adaptive wavelet method from [CDD02] to the case that Ψ is a *frame* for $H^t_{0,\Gamma_D}(\Omega)$ instead of a Riesz basis. Writing the domain as an *overlapping* union of subdomains Ω_q, a suitable frame Ψ is given by $\cup_{q=1}^M \omega_q \Psi_q$, where Ψ_q is a Riesz basis of wavelet type order d for a Sobolev space of order t on Ω_q, and ω_q is a smooth weight function that vanishes at the internal boundary of Ω_q. Since the presence of smooth weight functions and the fact that for $q \neq q'$, Ψ_q and $\Psi_{q'}$ are different are both harmless, Theorem 2.3 can be applied to verify s^*-compressibility of any of the matrices $\langle \omega_q \Psi_q, L\omega_{q'} \Psi_{q'} \rangle$ and with that of $\mathbf{M} := \langle \Psi, L\Psi \rangle = \sum_{q,q'} \langle \omega_q \Psi_q, L\omega_{q'} \Psi_{q'} \rangle$. Indeed, note that in the proof of this theorem it was not used that the wavelets are piecewise smooth with respect to partitions that are nested as a function of the level. The advantage of this frame approach is that smoothness requirements on the wavelets Ψ_q are easily satisfied, since this construction requires no linking of functions from different subdomains over interfaces, and so that $s^* > \frac{d-t}{n}$ can easily be realized.

3. Compressibility of boundary integral operators. In this section, we will generalize the compression result Theorem 2.3 to a certain class of nonlocal operators L. Since in particular we think of integral operators resulting from the boundary integral method, in addition we replace the underlying domain Ω by a manifold Γ.

3.1. Definitions and main result. For some $\mu \in \mathbb{N}$, let Γ be a patchwise smooth, compact n-dimensional, globally $C^{\mu-1,1}$ -manifold in $\mathbb{R}^{n+1}$. Following [DS99b], we assume that $\Gamma = \cup_{q=1}^M \overline{\Gamma_q}$, with $\Gamma_q \cap \Gamma_{q'} = \emptyset$ when $q \neq q'$, and that for each $1 \leq q \leq M$, there exists

- a domain $\Omega_q \subset \mathbb{R}^n$ and a C^∞-parametrization $\kappa_q : \mathbb{R}^n \to \mathbb{R}^{n+1}$ with $\mathrm{Im}(\kappa_q|_{\Omega_q}) = \Gamma_q$;
- a domain $\mathbb{R}^n \supset \hat{\Omega}_q \supset\supset \Omega_q$ and an extension of $\kappa_q|_{\Omega_q}$ to a $C^{\mu-1,1}$ parametrization $\hat{\kappa}_q : \hat{\Omega}_q \to \mathrm{Im}(\hat{\kappa}_q) \subset \Gamma$.

For $|s| \leq \mu$, the Sobolev spaces $H^s(\Gamma)$ are well-defined, where for $s < 0$, $H^s(\Gamma)$ is the dual of $H^{-s}(\Gamma)$. For some $|t| \leq \mu$, let L be a bounded operator from $H^t(\Gamma) \to H^{-t}(\Gamma)$, where we have in mind a singular integral operator of order $2t$. Let

$$\Psi = \{\psi_\lambda : \lambda \in \Lambda\}$$

be a *Riesz basis for* $H^t(\Gamma)$ of wavelet type.

We assume that the wavelets are *local*, in the sense that

$$\mathrm{diam}(\mathrm{supp}\, \psi_\lambda) \lesssim 2^{-|\lambda|} \quad \text{and} \quad \sup_{x \in \Gamma, \ell \in \mathbb{N}_0} \#\{|\lambda| = \ell : B(x; 2^{-\ell}) \cap \mathrm{supp}\, \psi_\lambda\} < \infty,$$

where for $A \subset \Gamma$ now $B(A; \varepsilon) := \{y \in \mathbb{R}^{n+1} : \mathrm{dist}(A, y) < \varepsilon\}$, and that they are *piecewise smooth*, by which we mean that $\mathrm{supp}\, \psi_\lambda \backslash \mathrm{sing\,supp}\, \psi_\lambda$ is the disjoint union

of sets $\Xi_{\lambda,1},\dots,\Xi_{\lambda,m}$, with $\cup_{i=1}^m \overline{\Xi_{\lambda,i}} = \operatorname{supp}\psi_\lambda$, such that each $\Xi_{\lambda,i}$ is contained in some Γ_q, $\kappa_q^{-1}(\Xi_{\lambda,i})$ are uniformly Lipschitz domains, and $(\psi_\lambda\circ\kappa_q)|_{\kappa_q^{-1}(\Xi_{\lambda,i})}$ is smooth with, for any $\beta\in N_0^n$,

$$\sup_{\xi\in\kappa_q^{-1}(\Xi_{\lambda,i})} |\partial^\beta(\psi_\lambda\circ\kappa_q)(\xi)| \lesssim 2^{(|\beta|+\frac{n}{2}-t)|\lambda|}. \tag{3.1}$$

Implicitly via (2.3), in section 2 we assumed that the wavelets have $\tilde d$ vanishing moments. A rough translation of this property to manifolds is that a wavelet integrated against a function vanishes when on each patch the preimage of this function is a polynomial of degree less than $\tilde d$. More precisely, we assume that the wavelets have the so-called *cancellation property of order* $\tilde d\in\mathbb{N}$, saying that there exists a constant $\eta>0$ such that for any $p\in[1,\infty]$, for all continuous, patchwise smooth functions v and $\lambda\in\Lambda$,

$$|\langle v,\psi_\lambda\rangle| \lesssim 2^{-|\lambda|(\frac{n}{2}-\frac{n}{p}+t+\tilde d)} \max_{1\le q\le M} |v|_{W_p^{\tilde d}(B(\operatorname{supp}\psi_\lambda;2^{-|\lambda|}\eta)\cap\Gamma_q)}. \tag{3.2}$$

Remark 3.1. Proofs of (3.2) for $p=\infty$ given in [DS99c, Prop. 4.7], [NS03, Prop. 3.4] can easily be generalized to yield (3.2) for any $p\in[1,\infty]$.

Furthermore, for some $k\in\mathbb{N}_0\cup\{-1\}$, with $k<\mu$ and

$$\gamma := k+\tfrac{3}{2} > t, \tag{3.3}$$

we assume that all $\psi_\lambda\in C^k(\Gamma)$, where $k=-1$ means no global continuity condition, and, similar to (2.3), that for all $r\in[-\tilde d,\gamma)$, $s<\gamma$, necessarily with $|s|,|r|\le\mu$,

$$\|\cdot\|_{H^r(\Gamma)} \lesssim 2^{\ell(r-s)}\|\cdot\|_{H^s(\Gamma)} \qquad \text{on } W_\ell := \operatorname{span}\{\psi_\lambda : |\lambda|=\ell\}. \tag{3.4}$$

Inside a patch, a similar property can be required for larger ranges: For all $1\le q\le M$, and for $r\in[-\tilde d,\gamma)$, $s<\gamma$, we assume that

$$\|\cdot\|_{H^r(\Gamma_q)} \lesssim 2^{\ell(r-s)}\|\cdot\|_{H^s(\Gamma_q)} \quad \text{on } \operatorname{span}\{\psi_\lambda : |\lambda|=\ell, B(\operatorname{supp}\psi_\lambda;2^{-\ell}\eta)\subset\overline{\Gamma_q}\}. \tag{3.5}$$

Remark 3.2. For the case that each parametrization κ_q has a constant Jacobian, in [DS99c] a simple construction is given of continuous finite element wavelets, i.e., $k=0$ and so $\gamma=\frac{3}{2}$, that in principle for any $\tilde d$ and order d satisfies the above assumptions. This restriction on the Jacobians was removed in [NS03]; however, here wavelets with supports that extend to more than one patch have the cancellation property of only order 1. In a forthcoming paper, we will remove this inconvenience for the application in adaptive methods and construct wavelets that all have the cancellation property of order $\tilde d$.

As in the domain case, wavelets that satisfy the assumptions for, in principle, any d, $\tilde d$ and smoothness permitted by both d and the regularity of the manifold were constructed in [DS99b]. As will be shown by Theorem 3.3, via several parameters this regularity, however, seems to impose a principal barrier to the compressibility. In case of differential operators, we could reduce conditions concerning smoothness and cancellation properties to corresponding patchwise conditions. Yet, the arguments used for that do not carry over to the case of nonlocal, integral operators.

With the constructions from [DS99a, CTU99, CM00], biorthogonality was realized with respect to a modified $L_2(\Gamma)$-scalar product. As a consequence, with the interpretation of functions as functionals via the Riesz mapping with respect to the

standard $L_2(\Gamma)$-scalar product, for negative t the wavelets only generate a Riesz basis for $H^t(\Gamma)$ when $t > -\frac{1}{2}$ (cf. [NS03, sect. 4]), and likewise wavelets with supports that extend to more than one patch generally have no cancellation properties in the sense of (3.2).

The frame approach discussed in the previous section seems to be even more attractive in the compact manifold case. Because of the absence of a boundary, all wavelet bases on the overlapping patches may satisfy periodic boundary conditions. One may verify that Theorem 3.3 below formulated for the Riesz basis case, as well as the verification of the estimates (3.6) and (3.7), extends to any of the matrices $\langle \omega_q \Psi_q, L\omega_{q'}\Psi_{q'}\rangle$. It is easy to construct collections Ψ_q with any smoothness permitted by the regularity of the manifold.

For the differential operator case, we made use of the fact that $\langle \psi_\lambda, L\psi_{\lambda'}\rangle = 0$ whenever $\mathrm{vol}(\operatorname{supp}\psi_\lambda \cap \operatorname{supp}\psi_{\lambda'}) = 0$. As a replacement, for the next theorem we will assume estimate (3.6), saying that the size of $|\langle \psi_\lambda, L\psi_{\lambda'}\rangle|$ decreases rapidly with increasing distance between the supports of ψ_λ and $\psi_{\lambda'}$. In section 3.3, we will verify this estimate for an important class of singular integral operators that have Schwarz kernels that decay rapidly away from the diagonal, and which therefore act mainly local. For ψ_λ and $\psi_{\lambda'}$ with $\mathrm{vol}(\operatorname{supp}\psi_\lambda \cap \operatorname{supp}\psi_{\lambda'}) > 0$, in the differential operator case we made use of the faster decay of $|\langle \psi_\lambda, L\psi_{\lambda'}\rangle|$ for $|\lambda| - |\lambda'| \to \infty$ whenever $\operatorname{supp}\psi_\lambda$ does not intersect the singular support of $\psi_{\lambda'}$. For the aforementioned class of singular integral operators, as a replacement in section 3.4 we will prove estimate (3.7), showing that $|\langle \psi_\lambda, L\psi_{\lambda'}\rangle|$ decreases for increasing $|\lambda| - |\lambda'|$ and increasing distance between $\operatorname{supp}\psi_\lambda$ and $\operatorname{sing\,supp}\psi_\lambda$. To obtain (3.7) we will need the additional condition that $\tilde{d} > \gamma - 2t$, which will be needed anyway in Remark 3.4 to conclude s^*-compressibility with $s^* > \frac{d-t}{n}$.

THEOREM 3.3. *Let $L : H^t(\Gamma) \to H^{-t}(\Gamma)$ be bounded, and for some $\sigma \in (0, \mu - |t|]$, let both L and its adjoint L' be bounded from $H^{t+\sigma}(\Gamma) \to H^{-t+\sigma}(\Gamma)$. For Ψ a Riesz basis for $H^t(\Gamma)$ as described above with $t + \tilde{d} > 0$, let $\mathbf{M} = \langle \Psi, L\Psi\rangle$.*

For any $\lambda, \lambda' \in \Lambda$, let

$$|\langle \psi_\lambda, L\psi_{\lambda'}\rangle| \lesssim \left(\frac{2^{-||\lambda|-|\lambda'||/2}}{\delta(\lambda,\lambda')}\right)^{n+2t+2\tilde{d}} \quad \text{when } \delta(\lambda,\lambda') \geq 3\eta, \tag{3.6}$$

where

$$\delta(\lambda,\lambda') := 2^{\min\{|\lambda|,|\lambda'|\}}\operatorname{dist}(\operatorname{supp}\psi_\lambda, \operatorname{supp}\psi_{\lambda'}),$$

and η is from (3.2).

For some $\tau \geq \sigma$ and all $\lambda' \in \Lambda$, $\ell > |\lambda'|$, $A_\ell \subset \{\lambda \in \Lambda : |\lambda| = \ell\}$ with

$$1 \gtrsim \tilde{\delta} := 2^{|\lambda'|}\operatorname{dist}(\cup_{\lambda\in A_\ell}\operatorname{supp}\psi_\lambda, \operatorname{sing\,supp}\psi_{\lambda'}) \geq 2\eta 2^{|\lambda'|-\ell},$$
$$\operatorname{diam}(\cup_{\lambda\in A_\ell}\operatorname{supp}\psi_\lambda) \lesssim 2^{-|\lambda'|},$$

let

$$\left.\begin{matrix}|\langle w_\ell, L\psi_{\lambda'}\rangle| \\ |\langle L'\psi_{\lambda'}, w_\ell\rangle|\end{matrix}\right\} \lesssim \|w_\ell\|_{H^t(\Gamma)} \max\left\{\frac{2^{(|\lambda'|-\ell)(t+\tilde{d})}}{\tilde{\delta}^{2t+\tilde{d}-\gamma}}, 2^{(|\lambda'|-\ell)\tau}\right\} \tag{3.7}$$

for any $w_\ell \in \operatorname{span}\{\psi_\lambda : \lambda \in A_\ell\}$.

Let $\alpha \in (\frac{1}{2}, 1)$ and $b_i := (1+i)^{-1-\varepsilon}$ for some $\varepsilon > 0$. Then for $j \in \mathbb{N}$, and with

$$k(j,n) := \frac{j}{n-1} \quad \text{when } n > 1,$$

and $j\frac{\min\{t+\tilde{d},\tau\}}{\min\{t+\mu,t+\tilde{d},\sigma\}} \le k(j,1) \le 2^j$, $k(j,1) > j\frac{\min\{t+\tilde{d},\tau\}}{\gamma-t}$, *we define the infinite matrix* $\mathbf{M}_j$ *by replacing all entries* $\mathbf{M}_{\lambda,\lambda'} = \langle \psi_\lambda, L\psi_{\lambda'}\rangle$ *by zeros when*

(3.8)
$$\big||\lambda| - |\lambda'|\big| > k(j,n), \quad \textit{or}$$

(3.9)
$$\big||\lambda| - |\lambda'|\big| \le \tfrac{j}{n} \quad \textit{and} \quad \delta(\lambda,\lambda') \ge \max\{3\eta, 2^{\alpha(\frac{j}{n} - \big||\lambda|-|\lambda'|\big|)}\}, \quad \textit{or}$$

(3.10)
$$\big||\lambda| - |\lambda'|\big| > \tfrac{j}{n} \quad \textit{and} \quad \tilde{\delta}(\lambda,\lambda') \ge \max\{2^{n(\frac{j}{n} - \big||\lambda|-|\lambda'|\big|)} b_{||\lambda|-|\lambda'||-\frac{j}{n}}, 2\eta 2^{-||\lambda|-|\lambda'||}\},$$

where

$$\tilde{\delta}(\lambda,\lambda') := 2^{\min\{|\lambda|,|\lambda'|\}} \times \begin{cases} \operatorname{dist}(\operatorname{supp}\psi_\lambda, \operatorname{sing\,supp}\psi_{\lambda'}) & \text{when} |\lambda| > |\lambda'|, \\ \operatorname{dist}(\operatorname{sing\,supp}\psi_\lambda, \operatorname{supp}\psi_{\lambda'}) & \text{when} |\lambda| < |\lambda'|. \end{cases}$$

Then the number of nonzero entries in each row and column of $\mathbf{M}_j$ *is of order* 2^j, *and for any*

$$s \le \min\left\{\tfrac{t+\tilde{d}}{n}, \tfrac{\tau}{n}\right\}, \ \textit{with } s < \tfrac{\gamma-t}{n-1},\ s \le \tfrac{\sigma}{n-1} \ \textit{and } s \le \tfrac{t+\mu}{n-1} \ \textit{when } n > 1,$$

it holds that $\|\mathbf{M} - \mathbf{M}_j\| \lesssim 2^{-sj}$.

Remark 3.4. As in Remark 2.4, we infer that $\mathbf{M}$ is s^*-compressible with

$$s^* = \min\left\{\tfrac{t+\tilde{d}}{n}, \tfrac{\tau}{n}, \tfrac{\gamma-t}{n-1}, \tfrac{\sigma}{n-1}, \tfrac{t+\mu}{n-1}\right\}$$

($s^* = \min\{t+\tilde{d}, \tau\}$ when $n = 1$). So, if $\tilde{d} > d - 2t$, $\tau > d - t$, and when $n > 1$, $\frac{\min\{\gamma-t,\sigma,t+\mu\}}{n-1} > \frac{d-t}{n}$, then $s^* > \frac{d-t}{n}$.

Proof. Let λ be some given index. Since Γ is a Lipschitz manifold, by the locality of the wavelets, the number of indices λ' with fixed $|\lambda'| \ge |\lambda|$ and $\operatorname{dist}(\operatorname{supp}\psi_\lambda, \operatorname{supp}\psi_{\lambda'}) \le R$ is of order $(2^{|\lambda'|}(2^{-|\lambda|}+R))^n$. By using in addition the piecewise smoothness of the wavelets, the number of indices λ' with fixed $|\lambda'| > |\lambda|$ and $\operatorname{dist}(\operatorname{sing\,supp}\psi_\lambda, \operatorname{supp}\psi_{\lambda'}) \le R$, where $2^{-|\lambda'|} \lesssim R \lesssim 2^{-|\lambda|}$, is of order $2^{(|\lambda'|-|\lambda|)(n-1)}2^{|\lambda'|}R$. From this, one may infer that the number of nonzero entries in the λth row or column of $\mathbf{M}_j$ is of order

$$\begin{aligned} &\sum_{-k(j,n)\le|\lambda'|-|\lambda|<0} 1 \quad + \sum_{0\le|\lambda'|-|\lambda|\le\frac{j}{n}} (2^{|\lambda'|}(2^{-|\lambda|} + 2^{-|\lambda|}\max\{3\eta, 2^{\alpha(\frac{j}{n}-|\lambda'|+|\lambda|}\}))^n \\ &+ \sum_{\frac{j}{n}<|\lambda'|-|\lambda|\le k(j,n)} 2^{(|\lambda'|-|\lambda|)(n-1)}2^{|\lambda'|}2^{-|\lambda|}\max\{2^{n(\frac{j}{n}-|\lambda'|+|\lambda|)}b_{|\lambda'|-|\lambda|-\frac{j}{n}}, 2\eta 2^{|\lambda|-|\lambda'|}\} \\ &\eqsim 2^j \end{aligned}$$

because of $\alpha < 1$ and $\sum_i b_i < \infty$.

Let $\hat{\mathbf{M}}_j$ be defined by

$$(\mathbf{M} - \hat{\mathbf{M}}_j)_{\lambda,\lambda'} = \begin{cases} \mathbf{M}_{\lambda,\lambda'} & \text{when } \big||\lambda| - |\lambda'|\big| > k(j,n), \\ 0 & \text{otherwise.} \end{cases}$$

The continuity assumptions on L, L', together with (3.4), show that for

$$r \in (0, t+\tilde{d}] \cap (0, t+\mu] \cap (0,\sigma] \cap (0, \gamma - t)$$

and $w_\ell \in W_\ell$, $w_{\ell'} \in W_{\ell'}$,

$$\begin{aligned}|\langle w_\ell, Lw_{\ell'}\rangle| &\lesssim \|w_\ell\|_{H^{t-r}(\Gamma)}\|Lw_{\ell'}\|_{H^{-t+r}(\Gamma)}\\ &\lesssim \|w_\ell\|_{H^{t-r}(\Gamma)}\|w_{\ell'}\|_{H^{t+r}(\Gamma)} \lesssim 2^{r(\ell'-\ell)}\|w_\ell\|_{H^t(\Gamma)}\|w_{\ell'}\|_{H^t(\Gamma)}\end{aligned}$$

and, analogously, $|\langle w_\ell, Lw_{\ell'}\rangle| = |\langle L'w_\ell, w_{\ell'}\rangle| \lesssim 2^{r(\ell-\ell')}\|w_\ell\|_{H^t(\Gamma)}\|w_{\ell'}\|_{H^t(\Gamma)}$. As in the proof of Theorem 2.3, we conclude that $\|\mathbf{M}-\hat{\mathbf{M}}_j\| \lesssim 2^{-k(j,n)r}$.

As a second step, let $\tilde{\mathbf{M}}_j$ be defined by

$$(\hat{\mathbf{M}}_j - \tilde{\mathbf{M}}_j)_{\lambda,\lambda'} = \begin{cases} \mathbf{M}_{\lambda,\lambda'} & \text{when } \delta(\lambda,\lambda') \geq \max\{1, 3\eta, 2^{\alpha(\frac{j}{n}-||\lambda|-|\lambda'||)}\} \text{ and } \big||\lambda|-|\lambda'|\big| \leq k(j,n),\\ 0 & \text{otherwise.}\end{cases}$$

Let us recall the Schur lemma: If, for some positive scalars ω_λ, $\sum_{\lambda'} \frac{\omega_\lambda |\mathbf{b}_{\lambda,\lambda'}|}{\omega_{\lambda'}} \leq c$ and $\sum_\lambda \frac{\omega_{\lambda'}|\mathbf{b}_{\lambda,\lambda'}|}{\omega_\lambda} \leq c$, then $\|(\mathbf{b}_{\lambda,\lambda'})_{\lambda,\lambda'\in\Lambda}\| \leq c$. We apply this lemma to $\hat{\mathbf{M}}_j - \tilde{\mathbf{M}}_j$ with $\omega_\lambda = 2^{|\lambda|\frac{n}{2}}$. By the locality of the wavelets, for each λ, ℓ', $R \gtrsim 1$, and $\beta > n$, one has

$$\sum_{\{\lambda':|\lambda'|=\ell',\, \delta(\lambda,\lambda')>R\}} \delta(\lambda,\lambda')^{-\beta} \lesssim R^{-\beta+n}2^{n\max\{0,\ell'-|\lambda|\}}.$$

Because of $t+\tilde{d} > 0$, from the decay estimate (3.6) we obtain that

$$\begin{aligned}&\sum_{\lambda'} \frac{\omega_\lambda |(\hat{\mathbf{M}}_j - \tilde{\mathbf{M}}_j)_{\lambda,\lambda'}|}{\omega_{\lambda'}}\\ &\lesssim \sum_{|\lambda'|} 2^{(|\lambda|-|\lambda'|)\frac{n}{2}} 2^{-||\lambda|-|\lambda'||(\frac{n}{2}+t+\tilde{d})}(\max\{1, 2^{\alpha(\frac{j}{n}-||\lambda|-|\lambda'||)}\})^{-(2\tilde{d}+2t)}2^{n\max\{0,|\lambda'|-|\lambda|\}}\\ &\eqsim 2^{-\frac{j}{n}(t+\tilde{d})}\end{aligned}$$

by $\alpha > \frac{1}{2}$. By the symmetry of the right-hand side of (3.6) in λ, λ', analogously we have $\sum_\lambda \frac{\omega_\lambda|(\hat{\mathbf{M}}_j-\tilde{\mathbf{M}}_j)_{\lambda,\lambda'}|}{\omega_{\lambda'}} \lesssim 2^{-\frac{j}{n}(t+\tilde{d})}$, and so $\|\hat{\mathbf{M}}_j - \tilde{\mathbf{M}}_j\| \lesssim 2^{-\frac{j}{n}(t+\tilde{d})}$.

Given λ' and $\ell > |\lambda'|$, let

$$A_{\ell,\lambda'} = \{|\lambda| = \ell : \tilde{\delta}(\lambda,\lambda') \geq \max\{2^{n(\frac{j}{n}-\ell+|\lambda'|)}b_{\ell-|\lambda'|-\frac{j}{n}}, 2\eta 2^{|\lambda'|-\ell}\},\ \delta(\lambda,\lambda') < \max\{1,3\eta\}\}.$$

Since for $\big||\lambda|-|\lambda'|\big| > \frac{j}{n}$, entries $\mathbf{M}_{\lambda,\lambda'}$ with $\delta(\lambda,\lambda') \geq \max\{1,3\eta\}$ were already removed from $\tilde{\mathbf{M}}_j$, we have

$$(\tilde{\mathbf{M}}_j - \mathbf{M}_j)_{\lambda,\lambda'} = \begin{cases} \mathbf{M}_{\lambda,\lambda'} & \begin{cases} \text{when} & \frac{j}{n} < |\lambda|-|\lambda'| \leq k(j,n) \quad \text{and} \quad \lambda \in A_{|\lambda|,\lambda'},\\ \text{or} & \frac{j}{n} < |\lambda'|-|\lambda| \leq k(j,n) \quad \text{and} \quad \lambda' \in A_{|\lambda'|,\lambda},\end{cases}\\ 0 & \text{otherwise.}\end{cases}$$

So from the decay estimate (3.7), for any $\mathbf{c},\mathbf{d}\in\ell_2(\Lambda)$ we have

$$\begin{aligned}
&\Big|\sum_{\frac{j}{n}<\ell-\ell'\le k(j,n)}\sum_{|\lambda'|=\ell'}\mathbf{d}_{\lambda'}\Big\langle\sum_{\lambda\in A_{\ell,\lambda'}}\mathbf{c}_\lambda\psi_\lambda,L\psi_{\lambda'}\Big\rangle\Big|\\
&\lesssim\sum_{\frac{j}{n}<\ell-\ell'\le k(j,n)}\sum_{|\lambda'|=\ell'}|\mathbf{d}_{\lambda'}|\max\Big\{\frac{2^{(\ell'-\ell)(t+\tilde d)}}{(2^{n(\frac{j}{n}-\ell+\ell')}b_{\ell-\ell'-\frac{j}{n}})^{2t+\tilde d-\gamma}},2^{(\ell'-\ell)\tau}\Big\}\Big\|\sum_{\lambda\in A_{\ell,\lambda'}}\mathbf{c}_\lambda\psi_\lambda\Big\|_{H^t(\Gamma)}\\
&\lesssim\sum_{\frac{j}{n}<\ell-\ell'\le k(j,n)}\max\Big\{\frac{2^{(\ell'-\ell)(t+\tilde d)}}{(2^{n(\frac{j}{n}-\ell+\ell')}b_{\ell-\ell'-\frac{j}{n}})^{2t+\tilde d-\gamma}},2^{(\ell'-\ell)\tau}\Big\}\sqrt{\sum_{|\lambda'|=\ell'}|\mathbf{d}_{\lambda'}|^2}\sqrt{\sum_{|\lambda'|=\ell'}\sum_{\lambda\in A_{\ell,\lambda'}}|\mathbf{c}_\lambda|^2}\\
&\lesssim\sum_{\frac{j}{n}<\ell-\ell'\le k(j,n)}\max\Big\{\frac{2^{(\ell'-\ell)(t+\tilde d)}}{(2^{n(\frac{j}{n}-\ell+\ell')}b_{\ell-\ell'-\frac{j}{n}})^{2t+\tilde d-\gamma}},2^{(\ell'-\ell)\tau}\Big\}\sqrt{\sum_{|\lambda'|=\ell'}|\mathbf{d}_{\lambda'}|^2}\sqrt{\sum_{|\lambda|=\ell}|\mathbf{c}_\lambda|^2},
\end{aligned}$$

where for the last line we have used that for fixed $|\lambda'|$, each λ is contained in at most a uniformly bounded number of sets $A_{|\lambda|,\lambda'}$. Since the analogous estimate is valid with interchanged roles of ℓ and ℓ', and for $s\le\min\{\frac{t+\tilde d}{n},\frac{\tau}{n}\}$, with $s<\frac{\gamma-t}{n-1}$ when $n>1$,

$$\sum_{m=1}^{k(j,n)-\frac{j}{n}}\max\{2^{-(m+\frac{j}{n})(t+\tilde d)}(2^{-mn}b_m)^{\gamma-2t-\tilde d},2^{-(m+\frac{j}{n})\tau}\}\lesssim 2^{-sj},\tag{3.11}$$

we conclude that for such s, $\|\tilde{\mathbf{M}}_j-\mathbf{M}_j\|\lesssim 2^{-sj}$.

A combination of the estimates for $\mathbf{M}-\hat{\mathbf{M}}_j$, $\hat{\mathbf{M}}_j-\tilde{\mathbf{M}}_j$, and $\tilde{\mathbf{M}}_j-\mathbf{M}_j$ shows that for $s\le\min\{\frac{t+\tilde d}{n},\frac{\tau}{n}\}$, with for $n>1$, $s<\frac{\gamma-t}{n-1}$, $s\le\frac{\sigma}{n-1}$, and $s\le\frac{t+\mu}{n-1}$, it holds that $\|\mathbf{M}-\mathbf{M}_j\|\lesssim 2^{-sj}$. ∎

3.2. Singular integral operators. In section 3.3 and 3.4, we verify the decay estimates (3.6) and (3.7) for operators

$$Lu(x)=\int_\Gamma K(x,y)u(y)d\mu(y)\qquad(x\in\Gamma),$$

with kernels that satisfy, for all $1\le q,q'\le M$, $\xi\in\Omega_q$, $\eta\in\Omega_{q'}$,

$$|\partial_\xi^\alpha\partial_\eta^\beta K(\kappa_q(\xi),\kappa_{q'}(\eta))|\lesssim\operatorname{dist}(\kappa_q(\xi),\kappa_{q'}(\eta)))^{-(n+2t+|\alpha|+|\beta|)}\quad(n+2t+|\alpha|+|\beta|>0)\tag{3.12}$$

(cf. [DHS02, Def. 2.1]). Following [DHS02], we emphasize that (3.12) requires patchwise smoothness but no global smoothness of Γ. Assuming only global Lipschitz continuity of Γ, the kernel of a boundary integral operator of order $2t$ can be shown to satisfy (3.12).

If Γ is a C^∞-manifold, then these boundary integral operators are known to be pseudodifferential operators, meaning that for any $\sigma\in\mathbb{R}$ they define bounded mappings from $H^{t+\sigma}(\Gamma)\to H^{-t+\sigma}(\Gamma)$. In this case we may conclude that $\mathbf{M}$ is s^*-compressibility with $s^*=\min\{\frac{t+\tilde d}{n},\frac{\gamma-t}{n-1}\}$. For Γ being only Lipschitz continuous, for the classical boundary integral equations it is known that $L:H^{t+\sigma}(\Gamma)\to H^{-t+\sigma}(\Gamma)$ is bounded for the, in this case, maximum possible value $\sigma=1-|t|$ (cf. [Cos88]). With increasing smoothness of Γ one may expect this boundedness for larger values of σ. Few results in this direction seem yet available.

3.3. Decay estimate (3.6). This estimate for singular integral operators with wavelets that satisfy the cancellation property (3.2) of order $\tilde{d}$ was first proved in [Sch98] for C^∞-manifolds. In [DS99c], it was shown for Lipschitz manifolds for a specific wavelet construction. For convenience, in this subsection we recall the arguments used there and show that they also apply to the general setting discussed in this paper.

With $\eta > 0$ from (3.2), let $\lambda, \lambda' \in \Lambda$ with $\delta(\lambda, \lambda') \geq 3\eta$. Then with $\Gamma_{\lambda,\eta} := B(\operatorname{supp}\psi_\lambda; 2^{-|\lambda|}\eta)$, it holds that

$$2^{\min\{|\lambda|,|\lambda'|\}}\operatorname{dist}(\Gamma_{\lambda,\eta}, \Gamma_{\lambda',\eta}) \geq \tfrac{1}{3}\delta(\lambda,\lambda') > 0.$$

Because of

$$n + 2t + 2\tilde{d} > 0,$$

from (3.2) and (3.12), we infer that

$$\begin{aligned}
|\langle \psi_\lambda, L\psi_{\lambda'}\rangle| &\lesssim 2^{-|\lambda|(\frac{n}{2}+t+\tilde{d})} \sup_{\substack{|\alpha|=\tilde{d},\, 1\leq q\leq M,\\ \xi\in\kappa_q^{-1}(\Gamma_{\lambda,\eta}\cap\Gamma_q)}} \left|\int_\Gamma \partial_\xi^\alpha K(\kappa_q(\xi), y)\psi_{\lambda'}(y)d\mu(y)\right| \\
&\lesssim 2^{-(|\lambda|+|\lambda'|)(\frac{n}{2}+t+\tilde{d})} \sup_{\substack{|\alpha|=\tilde{d},\, 1\leq q\leq M,\\ \xi\in\kappa_q^{-1}(\Gamma_{\lambda,\eta}\cap\Gamma_q)}} \sup_{\substack{|\beta|=\tilde{d},\, 1\leq q'\leq M,\\ \zeta\in\kappa_{q'}^{-1}(\Gamma_{\lambda',\eta}\cap\Gamma_{q'})}} |\partial_\xi^\alpha\partial_\zeta^\beta K(\kappa_q(\xi),\kappa_{q'}(\zeta))| \\
&\lesssim 2^{-(|\lambda|+|\lambda'|)(\frac{n}{2}+t+\tilde{d})}(2^{-\min\{|\lambda|,|\lambda'|\}}\delta(\lambda,\lambda'))^{-(n+2t+2\tilde{d})} \\
&= \left(\frac{2^{-||\lambda|-|\lambda'||/2}}{\delta(\lambda,\lambda')}\right)^{n+2t+2\tilde{d}}.
\end{aligned}$$

3.4. Decay estimate (3.7). Let $\Gamma = \cup_{q=1}^M \overline{\Gamma_q}$ be a compact n-dimensional, globally $C^{\mu-1,1}$-manifold in $\mathbb{R}^{n+1}$, where the Γ_q are C^∞-manifolds as described in section 3.1. For some $|t| \leq \mu$, let L be a singular integral operator of order $2t$ as described in section 3.2, which is bounded from $H^t(\Gamma) \to H^{-t}(\Gamma)$, and for which there exists a $\sigma > 0$ such that $L, L' : H^{t+\sigma}(\Gamma) \to H^{-t+\sigma}(\Gamma)$ are bounded. Let Ψ be a Riesz basis for $H^t(\Gamma)$ as described in section 3.1, consisting of local and piecewise smooth $C^k(\Gamma)$ wavelets, that have cancellation properties of order $\tilde{d}$, where $k \in \mathbb{N}_0 \cup \{-1\}$, $k < \mu$, and $\gamma := k + \frac{3}{2} > t$.

In addition, in this subsection we assume that

$$\tilde{d} > \gamma - 2t. \tag{3.13}$$

Furthermore, with

$$\tilde{H}^s(\Gamma_q) := \begin{cases} H^s(\Gamma_q) & \text{when } s \geq 0, \\ (H_0^{-s}(\Gamma_q))' & \text{when } s < 0, \end{cases}$$

we assume that there exists a $\tau \in (0, \mu - |t|]$ such that for all $1 \leq q \leq M$,

$$L : H^{t+\tau}(\Gamma) \to \tilde{H}^{-t+\tau}(\Gamma_q) \quad \text{is bounded.} \tag{3.14}$$

Remark 3.5. Since for any $|s| \leq \mu$, the restriction of functions on Γ to Γ_q is a bounded mapping from $H^s(\Gamma)$ to $\tilde{H}^s(\Gamma_q)$, from the boundedness of $L : H^{t+\sigma}(\Gamma) \to$

$H^{-t+\sigma}(\Gamma)$, it follows that in any case (3.14) is valid for $\tau=\sigma$. So, for example, for Γ a C^∞-manifold, (3.14) is valid for any $\tau\in\mathbb{R}$. Yet, in particular when $t<0$, for Γ less smooth, it might happen that (3.14) is valid for a τ that is strictly larger that any σ for which $L: H^{t+\sigma}(\Gamma)\to H^{-t+\sigma}(\Gamma)$ is bounded.

PROPOSITION 3.6. *In the above setting, for all $\lambda'\in\Lambda$, $\ell>|\lambda'|$, and $A_\ell\subset\{\lambda\in\Lambda: |\lambda|=\ell\}$ with $\tilde\delta:=2^{|\lambda'|}\mathrm{dist}(\cup_{\lambda\in A_\ell}\mathrm{supp}\,\psi_\lambda, \mathrm{sing\,supp}\,\psi_{\lambda'})\geq 2\eta 2^{|\lambda'|-\ell}$ and $\rho:=2^{|\lambda'|}\mathrm{diam}(\cup_{\lambda\in A_\ell}\mathrm{supp}\,\psi_\lambda)$, it holds that*

$$\left.\begin{array}{l}|\langle w_\ell, L\psi_{\lambda'}\rangle|\\ |\langle L'\psi_{\lambda'}, w_\ell\rangle|\end{array}\right\}$$
$$\lesssim \|w_\ell\|_{H^t(\Gamma)}\max\left\{2^{(|\lambda'|-\ell)(t+\tilde d)}\tilde\delta^{-2t-\tilde d+k+1}\rho^{\frac{n-1}{2}}\min\{\tilde\delta,\rho\}^{\frac12}, 2^{(|\lambda'|-\ell)\min\{\tau,t+\tilde d\}}\right\}$$

for any $w_\ell\in\mathrm{span}\{\psi_\lambda:\lambda\in A_\ell\}$. By substituting $\tilde\delta,\rho\lesssim 1$ and by using the fact that $k+\frac32=\gamma$, the decay estimate (3.7) *is obtained.*

The proof of Proposition 3.6 will largely rely on techniques developed in [Sch98, DHS02]. We start with a lemma that will be used to estimate contributions from pairs of functions with disjoint supports.

LEMMA 3.7. *For $\lambda'\in\Lambda$, $J\subset\{1,\dots,M\}$, let either $E=\cup_{i\in J}\overline{\Xi_{\lambda',i}}$ or $E=\overline{\Gamma\setminus\cup_{i\in J}\overline{\Xi_{\lambda',i}}}$, and for $\ell>|\lambda'|$, let $A_\ell\subset\{\lambda\in\Lambda:|\lambda|=\ell\}$ with $\bar\delta:=2^{|\lambda'|}\mathrm{dist}(\cup_{\lambda\in A_\ell}\mathrm{supp}\,\psi_\lambda, E)\geq 2\eta 2^{|\lambda'|-\ell}$ and $\rho:=2^{|\lambda'|}\mathrm{diam}(\cup_{\lambda\in A_\ell}\mathrm{supp}\,\psi_\lambda)$. Then for any $w_\ell\in\mathrm{span}\{\psi_\lambda:\lambda\in A_\ell\}$, $v\in L_\infty(\Gamma)$ with $\mathrm{supp}\,v\subset E$, and*

$$|v(y)|\lesssim 2^{(k+1+\frac{n}{2}-t)|\lambda'|}\mathrm{dist}(y,\partial E)^{k+1},$$

it holds that

$$|\langle w_\ell, Lv\rangle|\lesssim \|w_\ell\|_{H^t(\Gamma)}2^{(|\lambda'|-\ell)(t+\tilde d)}\bar\delta^{-2t-\tilde d+k+1}\rho^{\frac{n-1}{2}}\min\{\bar\delta,\rho\}^{\frac12}.$$

Proof. By (3.2), for any $\lambda\in A_\ell$, it holds that

$$|\langle\psi_\lambda, Lv\rangle|\lesssim 2^{-\ell(\frac{n}{2}+t+\tilde d)}\max_{1\leq q\leq M}|Lv|_{W^{\tilde d}_\infty(B(\mathrm{supp}\,\psi_\lambda;2^{-\ell}\eta)\cap\Gamma_q)}.$$

For any $1\leq q\leq M$, $x=\kappa_q(\xi)\in B(\mathrm{supp}\,\psi_\lambda;2^{-\ell}\eta)\cap\Gamma_q$, and $|\alpha|=\tilde d$, by (3.12) it holds that

$$\begin{aligned}|\partial^\alpha_\xi((Lv)\circ\kappa_q)(\xi)|&\lesssim\int_E|\partial^\alpha_\xi K(\kappa_q(\xi),y)|2^{(k+1+\frac{n}{2}-t)|\lambda'|}\mathrm{dist}(y,\partial E)^{k+1}d\mu(y)\\&\lesssim 2^{(k+1+\frac{n}{2}-t)|\lambda'|}\int_E|x-y|^{-(n+2t+\tilde d)}|x-y|^{k+1}d\mu(y)\\&\lesssim 2^{(k+1+\frac{n}{2}-t)|\lambda'|}\int_{z\in\mathbb{R}^n,|z|\geq\mathrm{dist}(x,E)}|z|^{k+1-(n+2t+\tilde d)}dz\\&\eqsim 2^{(k+1+\frac{n}{2}-t)|\lambda'|}\mathrm{dist}(x,E)^{-2t-\tilde d+k+1}\end{aligned}$$

because of $-2t-\tilde d+k+1<0$ by (3.13).

Writing $w_\ell=\sum_{\lambda\in A_\ell}c_\lambda\psi_\lambda$, by the locality of the wavelets and because of $\bar\delta\geq$

$2\eta 2^{|\lambda'|-\ell}$ we find that

$$\begin{aligned}|\langle w_\ell, Lv\rangle| &\lesssim \left(\sum_{\lambda\in A_\ell}|c_\lambda|^2\right)^{\frac12} 2^{-\ell(\frac{n}{2}+t+\tilde d)}2^{(k+1+\frac{n}{2}-t)|\lambda'|}\\ &\quad\times\left(\sum_{\lambda\in A_\ell}\sup_{x\in B(\operatorname{supp}\psi_\lambda;2^{-\ell}\eta)}\operatorname{dist}(x,E)^{-4t-2\tilde d+2k+2}\right)^{\frac12}\\ &\eqsim \|w_\ell\|_{H^t(\Gamma)}2^{-\ell(\frac{n}{2}+t+\tilde d)}2^{(k+1+\frac{n}{2}-t)|\lambda'|}\\ &\quad\times\left(2^{\ell n}\int_{\cup_{\lambda\in A_\ell}\operatorname{supp}\psi_\lambda}\operatorname{dist}(x,E)^{-4t-2\tilde d+2k+2}d\mu(x)\right)^{\frac12}. \qquad (3.15)\end{aligned}$$

Since $-4t-2\tilde d+2k+2<-1$ by (3.13), because of the geometry of E we may estimate

$$\begin{aligned}&\int_{\cup_{\lambda\in A_\ell}\operatorname{supp}\psi_\lambda}\operatorname{dist}(x,E)^{-4t-2\tilde d+2k+2}d\mu(x)\\ &\lesssim (2^{-|\lambda'|}\rho)^{n-1}\int_{2^{-|\lambda'|}\bar\delta}^{2^{-|\lambda'|}\bar\delta+2^{-|\lambda'|}\rho} z^{-4t-2\tilde d+2k+2}dz\\ &\eqsim (2^{-|\lambda'|}\rho)^{n-1}[(2^{-|\lambda'|}\bar\delta)^{-4t-2\tilde d+2k+3}-(2^{-|\lambda'|}(\bar\delta+\rho))^{-4t-2\tilde d+2k+3}]\\ &\eqsim 2^{|\lambda'|(4t+2\tilde d-2k-2-n)}\rho^{n-1}\bar\delta^{-4t-2\tilde d+2k+2}\min\{\bar\delta,\rho\}.\end{aligned}$$

By substituting this result into (3.15) the proof is completed. □

Proof of Proposition 3.6. (I) Let $\lambda'\in\Lambda$, $\ell>|\lambda'|$, and $A_\ell\subset\{\lambda\in\Lambda:|\lambda|=\ell\}$ with $\tilde\delta:=2^{|\lambda'|}\operatorname{dist}(\cup_{\lambda\in A_\ell}\operatorname{supp}\psi_\lambda,\operatorname{sing\,supp}\psi_{\lambda'})\ge 2\eta 2^{|\lambda'|-\ell}$ and $\rho:=2^{|\lambda'|}\operatorname{diam}(\cup_{\lambda\in A_\ell}\operatorname{supp}\psi_\lambda)$ and let $w_\ell\in\operatorname{span}\{\psi_\lambda:\lambda\in A_\ell\}$. It suffices to prove the bound for $\langle w_\ell, L\psi_{\lambda'}\rangle$, since the proof for $\langle L'\psi_{\lambda'}, w_\ell\rangle$ is similar.

For $1\le i\le m$, we define $w_\ell^{(i)}$ by

$$w_\ell^{(i)}(x)=\begin{cases} w_\ell(x) & \text{when } x\in\Xi_{\lambda',i},\\ 0 & \text{elsewhere}\end{cases}$$

and set $w_\ell^{(0)}=w_\ell-\sum_{i=1}^m w_\ell^{(i)}$, meaning that $\operatorname{supp}w_\ell^{(0)}\cap\operatorname{supp}\psi_{\lambda'}=\emptyset$.

We assume that $\operatorname{dist}(\operatorname{supp}\psi_\lambda,\operatorname{sing\,supp}\psi_{\lambda'})\ge 2\eta 2^{-|\lambda|}$ implies that either $\operatorname{supp}\psi_\lambda\subset\overline{\Xi_{\lambda',i}}$ for some $1\le i\le m$ or $\operatorname{supp}\psi_\lambda\cap\operatorname{supp}\psi_{\lambda'}=\emptyset$. In the very unlikely situation that this does not hold "automatically," we can always increase the parameter η such that this is true, since $\operatorname{diam}(\operatorname{supp}\psi_\lambda)\lesssim 2^{-|\lambda|}$. Under this assumption, for all i we have $w_\ell^{(i)}\in W_\ell$ and so $\|w_\ell^{(i)}\|_{H^t(\Gamma)}\lesssim\|w_\ell\|_{H^t(\Gamma)}$.

(II) We consider $\langle w_\ell^{(0)}, L\psi_{\lambda'}\rangle$. Let $E=\operatorname{supp}\psi_{\lambda'}$. If i is such that $\overline{\Xi_{\lambda',i}}\cap\partial E\ne\emptyset$, then because of $\psi_{\lambda'}\in C^k(\Gamma)$ and (3.1) it follows that

$$|\psi_{\lambda'}(y)|\lesssim 2^{(k+1+\frac{n}{2}-t)|\lambda'|}\operatorname{dist}(y,\partial E)^{k+1}\qquad(y\in\Xi_{\lambda',i}). \qquad (3.16)$$

If $\overline{\Xi_{\lambda',i}}\cap\partial E=\emptyset$, then by the "shape regularity" of all sets $\Xi_{\lambda',i'}$ for $i'\ne i$, we have $\operatorname{dist}(\Xi_{\lambda',i},E)\gtrsim 2^{-|\lambda'|}$, and so (3.16) follows from $|\psi_{\lambda'}(y)|\lesssim 2^{(\frac{n}{2}-t)|\lambda'|}$. From an application of Lemma 3.7 with "w"$=w_\ell^{(0)}$ and "v"$=\psi_{\lambda'}$, we conclude that

$$|\langle w_\ell^{(0)}, L\psi_{\lambda'}\rangle|\lesssim\|w_\ell\|_{H^t(\Gamma)}2^{(|\lambda'|-\ell)(t+\tilde d)}\tilde\delta^{-2t-\tilde d+k+1}\rho^{\frac{n-1}{2}}\min\{\tilde\delta,\rho\}^{\frac12}. \qquad (3.17)$$

(III) Let $1 \le i \le m$, and let $1 \le q \le M$ such that $\Xi_{\lambda',i} \subset \Gamma_q$. Note that $\operatorname{supp} w_\ell^{(i)} \subset \operatorname{supp} \psi_\lambda$. From (3.5), for $r \in [-\tilde{d},\gamma)$, $s < \gamma$, we have $\|w_\ell^{(i)}\|_{H^r(\Gamma_q)} \lesssim 2^{\ell(r-s)}\|w_\ell^{(i)}\|_{H^s(\Gamma_q)}$. We have to handle the less interesting case $\tau < \gamma - t$ separately.

(a) If $\tau < \gamma - t$, and so $\tau \le t + \tilde{d}$ by (3.13), then, by $\tau \le \mu - t$, the continuity of L as stated in (3.14), and when $t - \tau > 0$ additionally by $w_\ell^{(i)} \in H_0^{t-\tau}(\Gamma_q)$, we have

$$\begin{aligned}|\langle w_\ell^{(i)}, L\psi_{\lambda'}\rangle| &\lesssim \|w_\ell^{(i)}\|_{H^{t-\tau}(\Gamma_q)}\|L\psi_{\lambda'}\|_{\tilde{H}^{\tau-t}(\Gamma_q)}\\ &\lesssim 2^{(t-\tau)\ell}\|w_\ell^{(i)}\|\|\psi_{\lambda'}\|_{H^{\tau+t}(\Gamma)} \lesssim 2^{(|\lambda'|-\ell)\tau}\|w_\ell\|_{H^t(\Gamma)},\end{aligned}$$

which completes the proof in this case.

(b) Now let $\tau + t \ge \gamma \ge 0$. By assumption, $\psi_{\lambda'} \circ \kappa_q$ is smooth on $\kappa_q^{-1}(\Xi_{\lambda',i})$, which is a uniformly Lipschitz domain. From (3.1) and Remark 2.1, we learn that $(\psi_{\lambda'} \circ \kappa_q)|_{\kappa_q^{-1}(\Xi_{\lambda',i})}$ has an extension to a smooth function $\varphi_{\lambda',i}$, with, for $s \ge 0$ and $p \in [1,\infty]$, $\|\varphi_{\lambda',i}\|_{W_p^s(\mathbb{R}^n)} \lesssim 2^{(s-t+\frac{n}{2}-\frac{n}{p})|\lambda'|}$. By multiplying $\varphi_{\lambda',i}$ by a smooth function that is one on Ω_q and has support inside the "extended" domain $\hat{\Omega}_q$, we may assume that $\operatorname{supp}\varphi_{\lambda',i} \subset \hat{\Omega}_q$, so that $\varphi_{\lambda',i} \circ \hat{\kappa}_q^{-1} \in H^\mu(\Gamma)$ with for $s \in [0,\mu]$, $\|\varphi_{\lambda',i} \circ \hat{\kappa}_q^{-1}\|_{H^s(\Gamma)} \lesssim 2^{(s-t)|\lambda'|}$. With $\max\{0,-t\} \le s := \min\{\tau, t+\tilde{d}\} \le \mu - t$, the same arguments as applied in (a) show that

$$\begin{aligned}|\langle w_\ell^{(i)}, L(\varphi_{\lambda',i} \circ \hat{\kappa}_q^{-1})\rangle| &\lesssim \|w_\ell^{(i)}\|_{H^{t-s}(\Gamma_q)}\|L(\varphi_{\lambda',i} \circ \hat{\kappa}_q^{-1})\|_{\tilde{H}^{s-t}(\Gamma_q)}\\ &\lesssim 2^{(t-s)\ell}\|w_\ell^{(i)}\|\|\varphi_{\lambda',i} \circ \hat{\kappa}_q^{-1}\|_{H^{s+t}(\Gamma)} \lesssim 2^{(|\lambda'|-\ell)s}\|w_\ell\|_{H^t(\Gamma)}.\end{aligned} \tag{3.18}$$

It remains to estimate $|\langle w_\ell^{(i)}, L(\psi_{\lambda'} - \varphi_{\lambda',i} \circ \hat{\kappa}_q^{-1})\rangle|$. Recall that $\operatorname{supp} w_\ell^{(i)} \subset \Xi_{\lambda',i}$, whereas $\psi_{\lambda'} - \varphi_{\lambda',i} \circ \hat{\kappa}_q^{-1}$ vanishes on $\Xi_{\lambda',i}$. The global smoothness of $\psi_{\lambda'}$ will ensure that directly outside $\Xi_{\lambda',i}$, $\psi_{\lambda'}$ is sufficiently close to the smooth extension $\varphi_{\lambda',i} \circ \hat{\kappa}_q^{-1}$ of $\psi_{\lambda'}|_{\Xi_{\lambda',i}}$ so that an application of Lemma 3.7 will lead to the desired bound. We have to distinguish between a number of cases.

Suppose $i' \ne i$ with $\overline{\Xi_{\lambda',i'}} \cap \overline{\Xi_{\lambda',i}} \ne \emptyset$, and let q' be such that $\Xi_{\lambda',i'} \subset \Gamma_{q'}$. Then from

(a) $\sup_{\xi \in \kappa_{q'}^{-1}(\Xi_{\lambda',i'})} |\partial^\beta(\psi_{\lambda'} \circ \kappa_{q'})(\xi)| \lesssim 2^{(|\beta|+\frac{n}{2}-t)|\lambda'|}$, $(\beta \in \mathbb{N}_0)$ ((3.1));

(b) $\kappa_{q'}^{-1} \circ \hat{\kappa}_q \in C^{\mu-1,1}(\hat{\kappa}_q^{-1}(\Gamma_{q'} \cap \operatorname{Im}\hat{\kappa}_q))$, where $\mu > k$;

(c) $\sup_{\xi \in \hat{\Omega}_q} |\partial^\beta \varphi_{\lambda',i}(\xi)| \lesssim 2^{(|\beta|+\frac{n}{2}-t)|\lambda'|}$, $(\beta \in \mathbb{N}_0)$;

(d) $\psi_{\lambda'} \circ \hat{\kappa}_q - \varphi_{\lambda',i} \in C^k(\hat{\Omega}_q)$, where it vanishes on $\hat{\kappa}_q^{-1}(\Xi_{\lambda',i})$,

one infers that $|(\psi_{\lambda'} \circ \hat{\kappa}_q - \varphi_{\lambda',i})(\xi)| \lesssim 2^{(k+1+\frac{n}{2}-t)|\lambda'|}\operatorname{dist}(\xi, \hat{\kappa}_q^{-1}(\Xi_{\lambda',i}))^{k+1}$ when $\xi \in \hat{\kappa}_q^{-1}(\Xi_{\lambda',i'} \cap \operatorname{Im}\hat{\kappa}_q)$, so that for $y \in \Xi_{\lambda',i'} \cap \operatorname{Im}\hat{\kappa}_q$,

$$|(\psi_{\lambda'} - \varphi_{\lambda',i} \circ \hat{\kappa}_q^{-1})(y)| \lesssim 2^{(k+1+\frac{n}{2}-t)|\lambda'|}\operatorname{dist}(y, \Xi_{\lambda',i})^{k+1}. \tag{3.19}$$

If $\overline{\Gamma\backslash\operatorname{supp}\psi_{\lambda',i}} \cap \overline{\Xi_{\lambda',i}} \ne \emptyset$, then (a), (b) show that (3.19) is also valid for $y \in (\Gamma\backslash\operatorname{supp}\psi_{\lambda',i}) \cap \operatorname{Im}\hat{\kappa}_q$. For the remaining cases that either $y \in \Xi_{\lambda',i'}$ with $\overline{\Xi_{\lambda',i'}} \cap \overline{\Xi_{\lambda',i}} = \emptyset$ or $y \in (\Gamma\backslash\operatorname{supp}\psi_{\lambda',i}) \cap \operatorname{Im}\hat{\kappa}_q$ when $\overline{\Gamma\backslash\operatorname{supp}\psi_{\lambda',i}} \cap \overline{\Xi_{\lambda',i}} = \emptyset$ or $y \notin \operatorname{Im}\hat{\kappa}_q$, the "shape regularity" of all sets $\Xi_{\lambda',\check{\imath}}$ show that $\operatorname{dist}(y, \Xi_{\lambda',i}) \gtrsim 2^{-|\lambda'|}$, and so from $|(\psi_{\lambda'} - \varphi_{\lambda',i} \circ \hat{\kappa}_q^{-1})(y)| \lesssim 2^{(\frac{n}{2}-t)|\lambda'|}$, we conclude that (3.19) is valid for *all* $y \in \Gamma\backslash\Xi_{\lambda',i}$. An application of Lemma 3.7 with $E = \overline{\Gamma\backslash\Xi_{\lambda',i}}$, "$w$"$= w_\ell^{(i)}$, and "$v$"$= \psi_{\lambda'} - \varphi_{\lambda',i} \circ \hat{\kappa}_q^{-1}$ now shows that

$$|\langle w_\ell^{(i)}, L(\psi_{\lambda'} - \varphi_{\lambda',i} \circ \hat{\kappa}_q^{-1})\rangle| \lesssim \|w_\ell\|_{H^t(\Gamma)} 2^{(|\lambda'|-\ell)(t+\tilde{d})}\tilde{\delta}^{-2t-\tilde{d}+k+1}\rho^{\frac{n-1}{2}}\min\{\tilde{\delta},\rho\}^{\frac{1}{2}},$$

which, together with (3.17) and (3.18), completes the proof for the case that $\tau + t \geq \gamma$. ∎

Below we discuss some extensions of Proposition 3.6.

In the case $\operatorname{supp} w_\ell \cap \operatorname{supp} \psi_{\lambda'} = \emptyset$, then step (II) of the proof shows that

$$\left.\begin{array}{l} |\langle w_\ell, L\psi_{\lambda'}\rangle| \\ |\langle L'\psi_{\lambda'}, w_\ell\rangle| \end{array}\right\} \lesssim \|w_\ell\|_{H^t(\Gamma)} 2^{(|\lambda'|-\ell)(t+\tilde{d})} \tilde{\delta}^{-2t-\tilde{d}+k+1} \rho^{\frac{n-1}{2}} \min\{\tilde{\delta}, \rho\}^{\frac{1}{2}}. \tag{3.20}$$

In the case $\operatorname{supp} \psi_{\lambda'}$ is contained in one patch $\overline{\Gamma_q}$, then

$$\begin{aligned} &\left.\begin{array}{l} |\langle w_\ell, L\psi_{\lambda'}\rangle| \\ |\langle L'\psi_{\lambda'}, w_\ell\rangle| \end{array}\right\} \\ &\quad \lesssim \|w_\ell\|_{H^t(\Gamma)} \max\left\{ 2^{(|\lambda'|-\ell)(t+\tilde{d})} \tilde{\delta}^{-2t-\tilde{d}+k+1} \rho^{\frac{n-1}{2}} \min\{\tilde{\delta}, \rho\}^{\frac{1}{2}}, 2^{(|\lambda'|-\ell)(t+\tilde{d})} \right\}. \end{aligned} \tag{3.21}$$

To see this, consider $\langle w_\ell^{(i)}, L\psi_{\lambda'}\rangle$ as in step (III) of the proof. The functions $w_\ell^{(i)}$ and $\psi_{\lambda'}$ both live on the same patch $\overline{\Gamma_q}$, and so $\langle w_\ell^{(i)}, L\psi_{\lambda'}\rangle$ is independent of $\Gamma \backslash \Gamma_q$. That is, $\langle w_\ell^{(i)}, L\psi_{\lambda'}\rangle$ gets the same value if Γ is replaced by some extension of Γ_q to a compact C^∞-manifold. If we now follow the proof of Proposition 3.6 using this smooth manifold, then $\tau \geq \sigma$ can be any positive number, with which we conclude (3.21).

The estimate (3.21) is also valid when w_ℓ vanishes on all $\Xi_{\lambda',i}$ that do not have distance $\gtrsim 2^{-|\lambda'|}$ to an interface between different patches. Indeed, for the other $\Xi_{\lambda',i}$ the smooth extension of $\psi_\lambda|_{\Xi_{\lambda',i}}$ can always be chosen to be supported in one patch. So only in the remaining situation that $\operatorname{supp} \psi_{\lambda'}$ is not contained in one patch $\overline{\Gamma_q}$ *and* w_ℓ does not vanish on all $\Xi_{\lambda',i}$ adjacent to a patch interface, the parameter τ from (3.14) enters the upper bound for $|\langle w_\ell, L\psi_{\lambda'}\rangle|$ and $|\langle L'\psi_{\lambda'}, w_\ell\rangle|$. Nevertheless, unfortunately the value of τ does enter the estimate for s^*-compressibility, and although this can be expected also for less smooth manifolds, only for C^∞-manifolds have we shown that the value of s^* is never limited by τ.

As we have said, our proof of Proposition 3.6 is a modification of the approach from [DHS02]. A direct application of the technique from that paper yields the estimate

$$\left.\begin{array}{l} |\langle w_\ell, L\psi_{\lambda'}\rangle| \\ |\langle L'\psi_{\lambda'}, w_\ell\rangle| \end{array}\right\} \lesssim \|w_\ell\|_{H^t(\Gamma)} \max\left\{ 2^{(|\lambda'|-\ell)(t+\tilde{d})} \tilde{\delta}^{-2t-\tilde{d}} \rho^{\frac{n-1}{2}} \min\{\tilde{\delta}, \rho\}^{\frac{1}{2}}, 2^{(|\lambda'|-\ell)(t+\tilde{d})} \right\}, \tag{3.22}$$

which is less sharp when $k \geq 0$, except possibly, dependent on the value of τ from (3.14), in cases where (3.20) or (3.21) do not apply. If instead of Proposition 3.6 we apply (3.22) in the proof of Theorem 3.3, then by only adapting (3.11) we find that $\mathbf{M}$ is s^*-compressible with $s^* = \min\{\frac{t+\tilde{d}}{n}, \frac{1/2-t}{n-1}, \frac{\sigma}{n-1}, \frac{t+\mu}{n-1}\}$, and $s^* = t + \tilde{d}$ when $n = 1$.

The idea to obtain (3.22) is to write, for $1 \leq i \leq m$,

$$\langle w_\ell^{(i)}, L\psi_{\lambda'}\rangle = \langle w_\ell^{(i)}, L(\psi_{\lambda'}|_{\Xi_{\lambda',i}})\rangle + \langle w_\ell^{(i)}, L(\psi_{\lambda'} - \psi_{\lambda'}|_{\Xi_{\lambda',i}})\rangle.$$

Since $\psi_{\lambda'} - \psi_{\lambda'}|_{\Xi_{\lambda',i}}$ is generally discontinuous over $\partial\Xi_{\lambda',i}$, the second term can be bounded only by $2^{(|\lambda'|-\ell)(t+\tilde{d})} \tilde{\delta}^{-2t-\tilde{d}} \rho^{\frac{n-1}{2}} \min\{\tilde{\delta}, \rho\}^{\frac{1}{2}}$. On the other hand, $w_\ell^{(i)}$ and $\psi_{\lambda'}|_{\Xi_{\lambda',i}}$ both live on the same patch $\overline{\Gamma_q}$, so for estimating the first term we can always

think of Γ as being a C^∞-manifold, so that for any $\sigma>0$, $L: H^{t+\sigma}(\Gamma)\to H^{-t+\sigma}(\Gamma)$ is bounded. Writing $\psi_{\lambda'}|_{\Xi_{\lambda',i}}$ as the sum of a smooth function on this manifold plus their difference, which is zero on $\Xi_{\lambda',i}$ but which generally is discontinuous over $\partial\Xi_{\lambda',i}$, we can obtain the estimate (3.22).

The upper bounds for $|\langle w_\ell, L\psi_{\lambda'}\rangle|$ and $|\langle L'\psi_{\lambda'}, w_\ell\rangle|$ from Proposition 3.6, (3.21), and (3.22) are all given as a minimum of two terms. The first term decreases when $\rho:=2^{|\lambda'|}\mathrm{diam}(\cup_{\lambda\in A_\ell}\mathrm{supp}\,\psi_\lambda)$ decreases. The second term is independent of ρ. Responsible for this term is the part of w_ℓ spanned by those ψ_λ with $\lambda\in\tilde A_\ell:=\{\lambda\in A_\ell:\mathrm{supp}\,\psi_\lambda\subset\mathrm{supp}\,\psi_{\lambda'}\}$. Since for our application of Proposition 3.6 in Theorem 3.3 it held that $1\gtrsim\rho\geq\tilde\rho:=2^{|\lambda'|}\mathrm{diam}(\cup_{\lambda\in\tilde A_\ell}\mathrm{supp}\,\psi_\lambda)\gtrsim 1$, there was no need to construct an estimate that improves with decreasing diameter. However, to give, for example, sharp estimates of the individual entries of the infinite stiffness matrix $\langle\Psi, L\Psi\rangle$, such a need does exist.

For simplicity, let us consider a situation where we may pretend that Γ is a compact C^∞-manifold, as for (3.21) and (3.22). By the Sobolev embedding theorem, boundedness for any $\sigma\geq 0$ of $L: H^{t+\sigma}(\Gamma)\to H^{-t+\sigma}(\Gamma)$ implies boundedness for any $\sigma\geq 0$ and $p\in[2,\infty)$ of $L: H^{t+\sigma}(\Gamma)\to W_p^{-t+\sigma-n(\frac12-\frac1p)}(\Gamma)$. With $\tilde A_\ell^{(i)}:=\{\lambda\in A_\ell:\mathrm{supp}\,\psi_\lambda\subset\Xi_{\lambda',i}\}$ and $\frac{1}{p'}:=1-\frac1p$, using (3.2) and the locality of the wavelets, for $w_\ell=\sum_{\lambda\in A_\ell}c_\lambda\psi_\lambda$ we can replace (3.18) by

$$
\begin{aligned}
&\Big|\Big\langle\sum_{\lambda\in\tilde A_\ell^{(i)}}c_\lambda\psi_\lambda, L(\varphi_{\lambda',i}\circ\hat\kappa_q^{-1})\Big\rangle\Big|\\
&\lesssim 2^{-\ell(\frac{n}{2}-\frac{n}{p}+t+\tilde d)}\Big(\sum_{\lambda\in\tilde A_\ell^{(i)}}|c_\lambda|^{p'}\Big)^{\frac{1}{p'}}\Big(\sum_{\lambda\in\tilde A_\ell^{(i)}}|L(\varphi_{\lambda',i}\circ\hat\kappa_q^{-1})|^p_{W_p^{\tilde d}(B(\mathrm{supp}\,\psi_\lambda;2^{-|\lambda|}\eta)}\Big)^{\frac1p}\\
&\lesssim 2^{-\ell(\frac{n}{2}-\frac{n}{p}+t+\tilde d)}(\#\tilde A_\ell^{(i)})^{\frac12-\frac1p}\Big(\sum_{\lambda\in\tilde A_\ell^{(i)}}|c_\lambda|^2\Big)^{\frac12}|L(\varphi_{\lambda',i}\circ\hat\kappa_q^{-1})|_{W_p^{\tilde d}(\Gamma_q)}\\
&\lesssim 2^{-\ell(\frac{n}{2}-\frac{n}{p}+t+\tilde d)}(2^{\ell n}\mathrm{vol}(\cup_{\lambda\in\tilde A_\ell^{(i)}}\mathrm{supp}\,\psi_\lambda))^{\frac12-\frac1p}\|w_\ell\|_{H^t(\Gamma)}\|\varphi_{\lambda',i}\circ\hat\kappa_q^{-1}\|_{H^{\tilde d+2t+n(\frac12-\frac1p)}(\Gamma)}\\
&\lesssim 2^{-\ell(\frac{n}{2}-\frac{n}{p}+t+\tilde d)}2^{(\ell-|\lambda'|)n(\frac12-\frac1p)}\tilde\rho^{n(\frac12-\frac1p)}\|w_\ell\|_{H^t(\Gamma)}2^{(t+\tilde d+n(\frac12-\frac1p))|\lambda'|}\\
&\leq\rho^{n(\frac12-\frac1p)}\|w_\ell\|_{H^t(\Gamma)}2^{(|\lambda'|-\ell)(t+\tilde d)}.
\end{aligned}
$$

We conclude that for any $p\in[2,\infty)$, we may replace (3.21) by

$$
\begin{aligned}
&\left.\begin{matrix}|\langle w_\ell, L\psi_{\lambda'}\rangle|\\ |\langle L'\psi_{\lambda'}, w_\ell\rangle|\end{matrix}\right\}\\
&\lesssim\|w_\ell\|_{H^t(\Gamma)}\max\Big\{2^{(|\lambda'|-\ell)(t+\tilde d)}\tilde\delta^{-2t-\tilde d+k+1}\rho^{\frac{n-1}{2}}\min\{\tilde\delta,\rho\}^{\frac12}, 2^{(|\lambda'|-\ell)(t+\tilde d)}\rho^{n(\frac12-\frac1p)}\Big\}
\end{aligned}
\tag{3.23}
$$

and (3.22) by

$$
\begin{aligned}
&\left.\begin{matrix}|\langle w_\ell, L\psi_{\lambda'}\rangle|\\ |\langle L'\psi_{\lambda'}, w_\ell\rangle|\end{matrix}\right\}\\
&\lesssim\|w_\ell\|_{H^t(\Gamma)}\max\Big\{2^{(|\lambda'|-\ell)(t+\tilde d)}\tilde\delta^{-2t-\tilde d}\rho^{\frac{n-1}{2}}\min\{\tilde\delta,\rho\}^{\frac12}, 2^{(|\lambda'|-\ell)(t+\tilde d)}\rho^{n(\frac12-\frac1p)}\Big\}.
\end{aligned}
\tag{3.24}
$$

With $w_\ell=\psi_\lambda$ for some $|\lambda|=\ell$ and thus $\tilde\delta\gtrsim\rho\eqsim 2^{|\lambda'|-\ell}$, the estimate (3.24) can already be found in [DHS02] (however, with $p=\infty$, which seems to be a mistake).

The above technique can also be applied in the general setting of Proposition 3.6, thus in situations where (3.20) or (3.21) are not necessarily valid. However, in that case the resulting estimate will depend on continuity properties of L defined on a $C^{\mu-1,1}$-manifold with respect to $W_p^s(\Gamma)$-norms for $p \neq 2$.

Remark 3.8. In contrast to the derivation of the decay estimate from section 3.3, in the present subsection we never used the fact that $\psi_{\lambda'}$ is a wavelet, i.e., that it has cancellation properties. In other words, all results derived in this subsection are also valid when the wavelet on the lowest of the two involved levels is replaced by an arbitrary "scaling function" on that level, assuming that its $H^t(\Gamma)$-norm is of order 1.

To illustrate (3.23) and (3.24) we end with an example.

Example 3.9. Let Γ be the boundary of the unit square $[0,1]^2$, and let L be the single-layer operator, i.e., $(Lu)(x) = \int_\Gamma \log|x-y| u(y) d\mu(y)$. Let $\phi_{\lambda'}$ be the standard continuous piecewise linear "hat"-function with respect to the mesh $\Gamma \cap 2^{-|\lambda'|}\mathbb{N}_0^2$, attaining its maximum in the point denoted by $m(\lambda')$. Let ψ_λ be a continuous piecewise linear function with respect to $\Gamma \cap 2^{-|\lambda|}\mathbb{N}_0^2$, with $\operatorname{diam}(\operatorname{supp}\psi_\lambda) \eqsim 2^{-|\lambda|}$, $\operatorname{supp}\psi_\lambda \subset \operatorname{supp}\phi_{\lambda'}$, $\tilde{\delta} := 2^{|\lambda'|}\operatorname{dist}(\operatorname{supp}\psi_\lambda, \operatorname{sing\,supp}\phi_{\lambda'}) = 2^{|\lambda'|}\operatorname{dist}(\operatorname{supp}\psi_\lambda, m(\lambda'))$, having the cancellation property of order 3, and which has support contained in one of the edges of $[0,1]^2$. Note that $n=1$, $\mu=1$, $t=-\frac{1}{2}$, $k=0$, $d=2$, and $\tilde{d}=3$.

Using Maple, we computed $e := \frac{|\langle \psi_\lambda, L\phi_{\lambda'}\rangle|}{\|\phi_{\lambda'}\|_{H^t(\Gamma)}\|\psi_\lambda\|_{H^t(\Gamma)}}$. Regardless of the position of $m(\lambda')$, for $\tilde{\delta} \eqsim 1$ we found $e \eqsim 2^{3(|\lambda'|-|\lambda|)}$, which is in correspondence with (3.24) if we could take $p=\infty$. For $\tilde{\delta} = 2^{|\lambda'|-|\lambda|}$ and $m(\lambda')$ a corner of Γ, we found $e \eqsim 2^{(|\lambda'|-|\lambda|)}$, as predicted by (3.24). For $\tilde{\delta} = 2^{|\lambda'|-|\lambda|}$ and $m(\lambda')$ not being a corner, and so $\operatorname{supp}\phi_{\lambda'}$ contained in one of the edges, we found $e \eqsim 2^{2(|\lambda'|-|\lambda|)}$, which matches our new estimate (3.23), and which is an order better than predicted by (3.24).

REFERENCES

[BBC+01] A. Barinka, T. Barsch, P. Charton, A. Cohen, S. Dahlke, W. Dahmen, and K. Urban, *Adaptive wavelet schemes for elliptic problems—Implementation and numerical experiments*, SIAM J. Sci. Comput., 23 (2001), pp. 910–939.

[BCR91] G. Beylkin, R. Coifman, and V. Rokhlin, *The fast wavelet transform and numerical algorithms*, Comm. Pure Appl. Math., 44 (1991), pp. 141–183.

[CDD01] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods for elliptic operator equations—Convergence rates*, Math. Comp., 70 (2001), pp. 27–75.

[CDD02] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods.* II. *Beyond the elliptic case*, Found. Comput. Math., 2 (2002), pp. 203–245.

[CES00] A. Cohen, L. M. Echeverry, and Q. Sun, *Finite Element Wavelets*, technical report, Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie, 2000.

[CM00] A. Cohen and R. Masson, *Wavelet adaptive method for second order elliptic problems: Boundary conditions and domain decomposition*, Numer. Math., 86 (2000), pp. 193–238.

[Coh00] A. Cohen, *Wavelet methods in numerical analysis*, in Handbook of Numerical Analysis. Vol. 7, P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, 2000, pp. 417–711.

[Cos88] M. Costabel, *Boundary integral operators on Lipschitz domains: Elementary results*, SIAM J. Numer. Anal., 19 (1988), pp. 613–626.

[CTU99] C. Canuto, A. Tabacco, and K. Urban, *The wavelet element method part* I: *Construction and analysis*, Appl. Comput. Harmon. Anal., 6 (1999), pp. 1–52.

[Dah96] W. Dahmen, *Stability of multiscale transformations*, J. Fourier Anal. Appl., 4 (1996), pp. 341–362.

[Dah97] W. Dahmen, *Wavelet and multiscale methods for operator equations*, Acta Numer., 6 (1997), pp. 55–228.

[Dah99] S. Dahlke, *Besov regularity for elliptic boundary value problems on polygonal domains*, Appl. Math. Lett., 12 (1999), pp. 31–36.

[DD97] S. Dahlke and R. DeVore, *Besov regularity for elliptic boundary value problems*, Comm. Partial Differential Equations, 22 (1997), pp. 1–16.

[DDU02] S. Dahlke, W. Dahmen, and K. Urban, *Adaptive wavelet methods for saddle point problems—Optimal convergence rates*, SIAM J. Numer. Anal., 40 (2002), pp. 1230–1262.

[DeV98] R. DeVore, *Nonlinear approximation*, Acta Numer., 7 (1998), pp. 51–150.

[DHS02] W. Dahmen, H. Harbrecht, and R. Schneider, *Compression Techniques for Boundary Integral Equations—Optimal Complexity Estimates*, IGPM report, RWTH Aachen, Germany, 2002.

[DS98] W. Dahmen and R. Schneider, *Wavelets with complementary boundary conditions—Function spaces on the cube*, Results Math., 34 (1998), pp. 255–293.

[DS99a] W. Dahmen and R. Schneider, *Composite wavelet bases for operator equations*, Math. Comp., 68 (1999), pp. 1533–1567.

[DS99b] W. Dahmen and R. Schneider, *Wavelets on manifolds* I: *Construction and domain decomposition*, SIAM J. Math. Anal., 31 (1999), pp. 184–230.

[DS99c] W. Dahmen and R. Stevenson, *Element-by-element construction of wavelets satisfying stability and moment conditions*, SIAM J. Numer. Anal., 37 (1999), pp. 319–352.

[GR87] L. Greengard and V. Rokhlin, *A fast algorithm for particle simulation*, J. Comput. Phys., 73 (1987), pp. 325–348.

[HN89] W. Hackbusch and Z. P. Nowak, *On the fast matrix multiplication in the boundary element by panel clustering*, Numer. Math., 54 (1989), pp. 463–491.

[NS03] H. Nguyen and R. P. Stevenson, *Finite element wavelets on manifolds*, IMA J. Numer. Math, 23 (2003), pp. 149–173.

[Sch98] R. Schneider, *Multiskalen- und Wavelet-Matrixkompression: Analysisbasierte Methoden zur Lösung großer vollbesetzter Gleigungssysteme*, Adv. Numer. Math., Teubner, Stuttgart, 1998.

[Ste70] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

[Ste00] R. P. Stevenson, *Locally supported, piecewise polynomial biorthogonal wavelets on nonuniform meshes*, Constr. Approx., 19 (2003), pp. 477–508.

[Ste02] R. Stevenson, *Adaptive solution of operator equations using wavelet frames*, SIAM J. Numer. Anal., 41 (2003), pp. 1074–1100.

[vPS97] T. von Petersdorff and C. Schwab, *Fully discrete multiscale Galerkin BEM*, in Multiscale Wavelet Methods for Partial Differential Equations, Wavelet Anal. Appl. 6, W. Dahmen, P. Kurdila, and P. Oswald, eds., Academic Press, San Diego, CA, 1997, pp. 287–346.

SIAM J. MATH. ANAL.
Vol. 35, No. 5, pp. 1133–1159

ON THE OPTIMIZATION OF THE FUEL DISTRIBUTION IN A NUCLEAR REACTOR*

LAURENT THEVENOT†

Abstract. In this paper we give an optimality condition for the optimization problem of the distribution of fuel assemblies in a nuclear reactor by using the homogenization method. This study deals with purely fissile fuels and is based on the neutron transport equation modeling for continuous models. In particular, we prove the differentiability of the leading eigenvalue of the neutron transport operator with respect to the design parameter, the configuration of the fuels.

Key words. neutron transport, nuclear reactor, optimization, optimality condition, homogenization, leading eigenvalue, eigenvalue derivative

AMS subject classifications. 49K20, 49J50, 35F15

DOI. 10.1137/S0036141002415062

1. Introduction. This paper deals with the optimization problem of the assemblies distribution in a nuclear reactor. The core of a nuclear reactor is composed of a large number of different fuel and control rods which are immersed in a medium called a moderator (water, in most reactors). Because of the fission process that consumes the fuel, the old fuel must be replaced periodically by fresh fuel. However, only part of the fuel rods is removed since the fuel depletion is not spatially uniform in the reactor. So, to increase the lifetime of the assemblies, one is faced with the so-called optimal fuel reloading problem: optimize the assemblies distribution by maximizing the reactivity of the reactor. Although this is, in fact, a discrete optimization problem, it is beyond the scope of combinatorial methods due to the huge number of permutations of the rods and to the complexity of each computation. Many numerical methods have been proposed for solving this problem. For more details, we refer to [L], [AC1], [AC2], and the references therein. To date, these studies have been based on the modeling of the fission process by the multigroup neutron diffusion equation approximation. Our aim in this paper is to give an optimality condition for a simplified model of the continuous (in velocity) neutron transport equation accounting for a medium consisting only of (almost) purely fissile fuels. On the other hand, we deal with the continuous optimization problem; namely, we consider the shape of the domains occupied by the fuels to be free; a quasi-optimal distribution could be recovered by means of a penalization method. We follow the strategy developed in [AC1] for the one-group neutron diffusion equation (and generalized in [AC2] for the two-group neutron diffusion equation), inspired itself by the homogenization method introduced in particular in [MT]. Of course a natural follow-up of this work would be to implement numerically the optimality condition obtained herein.

The reactivity of the reactor core is measured by the real simple leading eigenvalue λ of the transport operator, which is associated with a positive eigenfunction, the only one to have a physical meaning. If $\lambda < 0$ (the reactor is said to be subcritical), the leakage of the neutrons at the boundary and the absorption of neutrons by the media

*Received by the editors September 28, 2002; accepted for publication (in revised form) June 13, 2003; published electronically January 6, 2004. This research was supported by European Research Training Network HMS-2000 (Project RTMI-1999-00040).
http://www.siam.org/journals/sima/35-5/41506.html
†Department of Mathematics, Technical University of Denmark, Building 303, 2800 Kgs. Lyngby, Denmark (thevenot@descartes.univ-fcomte.fr).

dominate the fission process, and the nuclear chain reaction dies out. If $\lambda > 0$ (the reactor is said to be supercritical), the phenomenon of fission leads and the nuclear chain reaction escalates. Lastly, if $\lambda = 0$ (the reactor is said to be critical), there is a perfect balance between the phenomenon of fission and that of depletion of neutrons by leakage or absorption, and the reactor can safely be operated. Nevertheless, the reactor can produce energy only if the criticality eigenvalue λ is slightly greater than 0; the reaction is then controlled by rods of absorbing matter.

Because of the disappearance phenomenon of the leading eigenvalue for small bodies with small velocities ($v \to 0$), before dealing with the optimization problem one has to ensure the existence of the leading eigenvalue, whatever the configuration of the assemblies, by finding practical criteria related to the physical parameters (space and velocity domains, cross sections which model the collisions). Indeed, one proves that the leading eigenvalue can disappear if the measure of the space domain Ω (not necessarily convex) becomes too small. However, for the models without small velocities ($v \not\to 0$) there exist some general existence results (see [M2, Chap. 5.4, p. 112]). To get an optimality condition, we follow the strategy of [AC1]. Namely, first we find a relaxed formulation of the optimization problem by using the homogenization method (see [DG], [M3], and [JT]); second, since the relaxed admissible subset is convex, we can take advantage of the direct method of calculus of variations to derive optimality conditions. But first, we prove the differentiability of the leading eigenvalue with respect to the design parameter, the (relaxed) configuration of the fuels. The proof is inspired by [CHM] and based on the implicit function theorem and the regularity of some convolution operators.

Here, the method for proving the differentiability of the leading eigenvalue, consisting in integrating with respect to the velocities the eigenvalue equation in order to get a more tractable equation, imposes to deal with separable transfer cross sections. This means we consider, in this paper, a medium composed only of (almost) purely fissile fuels for which the velocities before and after the collisions are not related. Let us point out that the spectral theory and the relaxation procedure rely on the compactness of the velocity averages, which is valid only for a class of transfer cross sections close to a separable one (see [M2, sect. 4.2]).

This paper is organized as follows. In section 2 we state the optimization problem and give a brief survey of the spectral theory of the neutron transport operator. In particular, we give some criteria of the existence of the leading eigenvalue independently of the configurations of the fuels for models with $v \to 0$, whose proof is postponed until Appendix B. Section 3 is devoted to the relaxed formulation. In section 4 we study the differentiability of the leading eigenvalue. The optimality condition is derived in section 5. Lastly, Appendix A deals with the disappearance phenomenon of the leading eigenvalue.

2. Statement of the problem. Let $\Omega \subset \mathbb{R}^N$ be a bounded open convex subset, the domain occupied by the core of the reactor. Let $V \subset \mathbb{R}^N$ be a bounded closed subset symmetric about 0, the velocity admissible space of the neutrons, endowed with the Lebesgue measure.

A nuclear fuel is characterized by two physical parameters called cross sections, which are bounded and nonnegative: the collision frequency $\sigma(x, v)$, which is the collision probability for neutrons located at position x and with velocity v, and the transfer probability $k(x, v, v')$, which is the probability that a neutron at position x and with velocity v' gives rise, after a collision with a nucleus of the fuel, to a neutron with velocity v. The collision takes into account not only the fission phenomenon but

also the scattering phenomenon when the neutron collides with a nonfissile nucleus and is just scattered into another direction (the energy of the velocity can also change if one considers neutrons with high velocity). In this study we will assume that the fuels are (almost) purely fissile, and we will neglect the scattering phenomenon. In the fission process, the velocities before and after the collision are not related, and thus the transfer cross section k, due to the fission, is a function with separated velocity variables. We refer to [BG] for a more complete introduction to reactor physics.

As explained in the introduction, the physically relevant parameter which measures the reactivity of the nuclear reaction is the leading eigenvalue of the transport operator. The transport operator on $L^p(\Omega \times V)$ $(1 \leq p < \infty)$ is given by $T + K$; the unbounded penetration operator T is defined by

$$\begin{cases} T\varphi(x,v) = -v \cdot \nabla_x \varphi(x,v) - \sigma(x,v)\varphi(x,v), \\ D(T) = W_0^p(\Omega \times V) \end{cases} \tag{1}$$

and the (bounded) transfer operator K by

$$K\varphi(x,v) = \int_V k(x,v,v')\varphi(x,v')dv', \tag{2}$$

where

$$W_0^p(\Omega \times V) = \left\{ \varphi \in L^p(\Omega \times V),\ v \cdot \nabla_x \varphi \in L^p(\Omega \times V),\ \varphi = 0 \text{ on } \Gamma_- \right\}$$

and $\Gamma_- = \{(x,v) \in \partial\Omega \times V \setminus v \cdot n(x) < 0\}$ denotes the incoming flux, $n(x)$ being the outward normal at $x \in \partial\Omega$.

The *existence theory* of the *leading eigenvalue* is based on the compactness of the velocity averages in transport theory (see [GLPS], [G], and [M2]) and the strict positivity of the transport semigroup $\{V(t);\ t \geq 0\}$ generated by $T + K$, due to the nonnegativity of the cross sections σ and k. Both arguments are given, respectively, by the two following results.

THEOREM 1 (see [M2, Thm. 3.2(ii), p. 37]). *Let Ω be a bounded and not necessarily convex open subset and let $d\mu$ be the measure on V. Let $1 < p < \infty$. If the hyperplanes have zero $d\mu$-measure, then the following operator is compact:*

$$\varphi \in W_0^p(\Omega \times V) \longmapsto \int_V \varphi(x,v)d\mu(v) \in L^p(\Omega).$$

THEOREM 2 (see [Vo1, Thm. 3.2]). *Let Ω be connected and let V be an open subset endowed with the Lebesgue measure. Let there exist $0 \leq c_1 < c_2 < \infty$ such that*

$$V_0 := \left\{ v \in \mathbb{R}^N;\ c_1 \leq |v| \leq c_2 \right\} \subset V$$

and

$$k(x,v,v') > 0 \text{ a.e. on } (\Omega \times V_0 \times V) \cup (\Omega \times V \times V_0); \tag{3}$$

then the semigroup $\{V(t);\ t \geq 0\}$ is irreducible.

The leading eigenvalue is related to the "asymptotic" spectrum of the transport operator

$$\sigma_{as}(T+K) = \sigma(T+K) \cap \left\{ \lambda \in \mathbb{C};\ \mathrm{Re}\lambda > s(T) \right\},$$

where $s(T) = \sup\{\mathrm{Re}\lambda, \lambda \in \sigma(T)\}$ is the spectral bound of T. The compactness of the velocity averages ensures that the transfer operator K is T-compact (i.e., $K(\lambda - T)^{-1}$-compact for $\mathrm{Re}\lambda > s(T)$) in $L^p(\Omega \times V)$ for $1 < p < \infty$ if the transfer cross section k is a separable or bounded function. Thus, from a classical stability result of essential spectra and from [Vo1], one deduces that, with our assumptions (σ bounded and nonnegative and Ω bounded), the half-plane $\mathrm{Re}\lambda \leq s(T)$ belongs to the essential spectrum of $T+K$; this result remains true in L^1 setting (see [M4]), even though only a power of $K(\lambda - T)^{-1}$ is compact in $L^1(\Omega \times V)$ (see [M2, Chap. 4.2]). As a consequence of the so-called Gohberg–Shmulyan theorem, it follows that the asymptotic spectrum consists at most of isolated eigenvalues with finite algebraic multiplicity (see [Vi], [KLH, Thm. 13.13, p. 277], or [GMP, Chap. XII.5]). If the semigroup $\{V(t);\ t \geq 0\}$ is irreducible and if its spectral bound $s(T+K)$ is a pole of the resolvent $(\lambda - T - K)^{-1}$, then $\omega(T+K)(= s(T+K))$ is a simple eigenvalue (possibly cyclic) associated with a positive eigenfunction (see [C, p. 209]). Lastly, if k is a separable or bounded function, it is known that the essential type of T is stable with respect to the perturbation K, namely, $\omega_e(T+K) = \omega_e(T)$. (The proof is related also to a compactness argument of some velocity averages; see [M2, sect. 4.3, p. 65, and Thm. 2.10, p. 24].) In our setting we have $\omega_e(T) = s(T)$ (see [Vo1] and [Vo2]). It is now obvious to conclude, under the previous assumptions, that if $\sigma_{as}(T+K) \neq \varnothing$, then $\omega(T+K)$ is the only element of the spectrum of $T+K$ located on the line $\mathrm{Re}\lambda = \omega(T+K)$, and $\omega(T+K)$ is a simple eigenvalue associated with a positive eigenfunction ($\omega(T+K)$ is also the only eigenvalue to be associated with a nonnegative eigenfunction). Moreover, there exists $\varepsilon > 0$ such that

$$\sigma(T+K) \cap \{\lambda \in \mathbb{C};\ \mathrm{Re}\lambda \geq \omega(T+K) - \varepsilon\} = \{\omega(T+K)\}.$$

Thus $\omega(T+K)$ is called the leading eigenvalue, and the eigenfunction associated with $\omega(T+K)$ is the only one which has physical meaning because it is positive. The question is therefore to find practical criteria for the existence of this leading eigenvalue. Since the function

$$\lambda \in\ ]s(T), \infty[\ \longmapsto r_\sigma\big((\lambda - T)^{-1}K\big)$$

is continuous and strictly decreasing (see [M2, sect. 5.7]), then $\sigma_{as}(T+K)$ is nonempty if and only if

$$\lim_{\lambda \to s(T)} r_\sigma\big((\lambda - T)^{-1}K\big) > 1,$$

and then the leading eigenvalue is a solution of the equation

$$r_\sigma\big((\lambda - T)^{-1}K\big) = 1.$$

For a more complete survey of the theory, we refer to [M2, Chap. 2, 3, 4, and 5] (and [M5] and [M6] for recent developments). Finally, we recall (see [Vo2]) that T generates an explicit c_0-semigroup

$$\mathrm{e}^{tT}\varphi = \mathrm{e}^{-\int_0^t \sigma(x-sv,v)ds}\varphi(x-tv,v)\chi_{[0,s(x,v)]}(t), \tag{4}$$

where the stay time $s(x,v)$ is given by

$$s(x,v) = \inf\{s > 0\ /\ x - vs \notin \Omega\},$$

and

$$s(T) = -\lim_{t\to\infty}\left\{\operatorname{ess}\inf_{t<s(x,v)}\ \frac{1}{t}\int_0^t \sigma(x-sv,v)ds\right\}. \tag{5}$$

Let us note, if $0 \in V$ and if the collision frequency σ is space homogeneous, that

$$s(T) = -\liminf_{v\to 0}\sigma(v). \tag{6}$$

We are now in position to state the *optimization problem* with which we are concerned. We consider that the matter consists of I different fuels, which are characterized by the (space homogeneous and measurable) cross sections $\sigma_i(v)$ and $k_i(v,v')$ $(i=1,\dots,I)$. Let $\omega_i \subset \Omega$ be the domain occupied by each fuel i such that

$$\omega_i \cap \omega_j = \varnothing \text{ if } i \neq j \quad \text{and} \quad \bigcup_{i=1}^{I}\omega_i = \Omega. \tag{7}$$

One imposes the volume constraint

$$\int_\Omega \chi_i(x)dx = \mu_i, \tag{8}$$

where χ_i is the characteristic function of the domain ω_i, and μ_i are fixed positive constants such that

$$\sum_{i=1}^{I}\mu_i = |\Omega|.$$

We define the cross sections of the reactor, which depend on the partition defined by $\chi = (\chi_i)_{i=1,\dots,I}$:

$$\sigma_\chi(x,v) = \sum_{i=1}^{I}\chi_i(x)\sigma_i(v),\ \ k_\chi(x,v,v') = \sum_{i=1}^{I}\chi_i(x)k_i(v,v').$$

We denote by λ_χ the leading eigenvalue corresponding to the partition $\chi = (\chi_i)_{i=1,\dots,I}$; namely, λ_χ is the leading eigenvalue of the transport operator $T_\chi + K_\chi$ on $L^p\big(\Omega\times V\big)$, where

$$\begin{cases} T_\chi\varphi(x,v) = -v\cdot\nabla_x\varphi(x,v) - \sigma_\chi(x,v)\varphi(x,v), \\ D(T_\chi) = W_0^p\big(\Omega\times V\big) \end{cases}$$

and

$$K_\chi\varphi(x,v) = \int_V k_\chi(x,v,v')\varphi(x,v')dv'.$$

The aim is to find the supremum of the leading eigenvalue λ_χ on the admissible subset of all configurations

$$\mathcal{U}_{ad} = \left\{\chi = (\chi_i)_{1\le i\le I} \in L^\infty(\Omega;\{0,1\});\ \sum_{I=1}^{I}\chi_i(x) = 1 \text{ and } \int_\Omega \chi_i(x)dx = \mu_i\right\}.$$

However, to avoid some peaks of the power distribution (defined as the energy released by fission) which can locally yield a strong increase of the temperature and cause damage to the reactor, one has to minimize also the power peak given by the maximal value of $\int_V K_\chi \varphi_\chi(x,v)dv$, where φ_χ is the positive leading eigenfunction associated with λ_χ. Since such a criterion is nondifferentiable, we search to minimize the averaging of $K_\chi\varphi_\chi$ on $\Omega \times V$, normalized by $\left(\sum_{i=1}^I \| \int_V \varphi_\chi(.,v) g_i(v)dv\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}$ because the eigenfunction is not uniquely defined. The functions g_i are introduced below. For more details we refer to [AC1] and [AC2]. Let us point out that in this paper we get only the differentiability of the velocity averages (with the weights g_i) of the leading eigenfunction and the differentiability of $K_\chi\varphi_\chi$, but not the differentiability of the eigenfunction φ_χ. Therefore, the problem is written as follows:

$$\text{Find} \quad \inf_{\chi\in\mathcal{U}_{ad}} \left\{ -\lambda_\chi + \frac{(K_\chi\varphi_\chi, 1)_{L^2(\Omega\times V)}}{\left(\sum_{i=1}^I \| \int_V \varphi_\chi(.,v) g_i(v)dv\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}} \right\}. \tag{9}$$

However, because of the disappearance phenomenon of the leading eigenvalue for small domains (see Appendix A), one has to prove the existence of λ_χ for all the configurations in $\mathcal{U}_{ad}$ before dealing with the optimization problem.

We state now the technical *assumptions* made throughout this paper. First, we assume there exist some nonnegative functions $\underline{\sigma}$ and $\overline{\sigma}$ in $L^\infty(V)$ such that $\underline{\sigma} \le \sigma_i \le \overline{\sigma}$ for all $i = 1, \dots, I$. Second, as we consider only purely fissile fuels, the transfer cross sections are separable functions

$$k_i(v,v') = f_i(v)g_i(v'), \text{ where } f_i \in L^\infty(V),\ g_i \in L^\infty(V)\ (i = 1,\dots,I).$$

The functions f_i, which denote the fission spectrum, are normalized by

$$\int_V f_i(v)dv = 1, \tag{10}$$

while the functions g_i represent the product of the absorption spectrum with the average number of neutrons produced by fission. We assume there exist some functions f_m, f_M, g_m, g_M in $L^\infty(V)$ such that $f_m \le f_i \le f_M$, $g_m \le g_i \le g_M$ for all $i = 1, \dots, I$. To be in the setting of Theorem 2 and to ensure the irreducibility of the semigroup generated by $T_\chi + K_\chi$ for all configurations, we assume furthermore that the functions f_m and g_m are positive.

On the other hand, since the transfer cross sections k_χ are the sum of separable functions, the compactness of velocity averages (Theorem 1) implies the compactness of $K_\chi(\lambda - T_\chi)^{-1}$ in $L^p(\Omega \times V)$ for $\mathrm{Re}\lambda > s(T_\chi)$. Moreover, as the results in the following sections are true in full generality, we assume that the asymptotic spectrum $\sigma_{as}(T_\chi + K_\chi)$ is nonempty regardless of the configuration (that is, the leading eigenvalue exists regardless of the configuration). More precisely, let us consider the unbounded penetration operators $\underline{T}$ and $\overline{T}$ defined as in (1) but with the collision frequencies $\underline{\sigma}$ and $\overline{\sigma}$, respectively, as well as the bounded integral operators $\underline{K}$ and $\overline{K}$ defined as in (2) but with the kernels $f_m(v)g_m(v')$ and $f_M(v)g_M(v')$, respectively. We assume there exist $\overline{\lambda} > \underline{\lambda} \ge s(\underline{T})$ such that $r_\sigma\big((\underline{\lambda} - \overline{T})^{-1}\underline{K}\big) > 1$ and $r_\sigma\big((\overline{\lambda} - \underline{T})^{-1}\overline{K}\big) < 1$. With these assumptions the leading eigenvalue λ_χ exists regardless of the configuration χ in $\mathcal{U}_{ad}$ and satisfies $\underline{\lambda} \le \lambda_\chi \le \overline{\lambda}$, and problem (9) is well-posed, as is shown in the following proof.

Proof 1. First, according to (5), we have

$$s(\overline{T}) \leq s(T_\chi) \leq s(\underline{T}).$$

On the other hand, it is easy to see, due to (4) (see also the proof of Proposition 1) that

$$(\lambda - \overline{T})^{-1}\underline{K} \leq (\lambda - T_\chi)^{-1}K_\chi \leq (\lambda - \underline{T})^{-1}\overline{K} \quad \forall \lambda > s(\underline{T}).$$

Then

$$r_\sigma\big((\underline{\lambda} - T_\chi)^{-1}K_\chi\big) \geq r_\sigma\big((\underline{\lambda} - \overline{T})^{-1}\underline{K}\big) > 1 \quad \forall \chi \in \mathcal{U}_{ad},$$

and therefore the leading eigenvalue λ_χ exists regardless of the configuration χ in $\mathcal{U}_{ad}$ and $\underline{\lambda} \leq \lambda_\chi$. Moreover,

$$r_\sigma\big((\overline{\lambda} - T_\chi)^{-1}K_\chi\big) \leq r_\sigma\big((\overline{\lambda} - \underline{T})^{-1}\overline{K}\big) < 1 \quad \forall \chi \in \mathcal{U}_{ad},$$

so that $\lambda_\chi \leq \overline{\lambda}$. Finally problem (9) is well-posed because, using the definition of the characterization function, we obtain

$$\|K_\chi \varphi_\chi\|^2_{L^2(\Omega\times V)} \leq \sum_{i=1}^{I} \|f_i\|^2_{L^2(V)} \left\| \int_V \varphi_\chi(.,v')g_i(v')dv' \right\|^2_{L^2(\Omega)}. \qquad \square$$

We end this section by giving some *practical criteria*, which ensure those for latter assumptions, and by posing an open question. By adapting the method in [M1, Lem. 2], introduced for space homogeneous cross sections, we get some practical criteria related to the size of the domain Ω. We detail the proof in Appendix B.

PROPOSITION 1. *Let $1 < p < \infty$. Assume that $0 \in V$, the collision frequency σ_i are positive constants, $f_i \in L^p(V)$, $g_i \in L^{p'}(V)$, and there exist some functions f_m, f_M in $L^p(V)$ and g_m, g_M in $L^{p'}(V)$ such that $f_m \leq f_i \leq f_M$ and $g_m \leq g_i \leq g_M$ for all $i = 1, \ldots, I$. Assume furthermore that $f_m(-v)g_m(-v) = f_m(v)g_m(v)$ for all $v \in V$ and*

$$\int_V f_m(v)g_m(v)dv > \overline{\sigma} - \underline{\sigma},$$

where $\underline{\sigma} = \min\{\sigma_i,\ i = 1, \ldots, I\}$ and $\overline{\sigma} = \max\{\sigma_i,\ i = 1, \ldots, I\}$.

If Ω contains a large enough ball, independent of the configurations $\chi \in \mathcal{U}_{ad}$, then

$$\sigma_{as}(T_\chi + K_\chi) \neq \varnothing \quad \forall \chi \in \mathcal{U}_{ad}.$$

Moreover, there exists a positive constant α such that the leading eigenvalue λ_χ satisfies

$$-\overline{\sigma} + \alpha \leq \lambda_\chi \leq -\underline{\sigma} + \|\overline{K}\| \quad \forall \chi \in \mathcal{U}_{ad},$$

where $\overline{K}$ denotes the integral operator on $L^p(\Omega \times V)$ defined as in (2) with kernel $f_M(v)g_M(v')$.

Remark 1. If $0 \notin V$ (i.e., V bounded away from 0), Proposition 1 is still valid. The only difference is that $s(T_\chi) = -\infty$ and the leading eigenvalue always exists independently of the size of Ω (see [M2, Thm. 5.17 and 5.18, p. 113]).

Open question. Assuming that the leading eigenvalue exists for each fuel i (for the whole space domain Ω), does the leading eigenvalue exist for any possible mixture of fuels?

3. The relaxed problem. Problem (9) has two disadvantages. First, the set $\mathcal{U}_{ad}$ is not closed in the weak star topology of $L^\infty(\Omega)$. Indeed minimizing sequences may converge to limits which are not within $\mathcal{U}_{ad}$, that is, which are not characteristic functions. Second, we cannot differentiate the cost functions in problem (9) because the set $\mathcal{U}_{ad}$ is not stable with respect to convex combinations. To overcome these two difficulties, we carry out the so-called relaxation procedure. We determine the relaxed or generalized optimization problem by applying the homogenization methods introduced in [DG] and [M3] (and used also in [JT]), which rely on arguments of compactness of the velocity averages. From a physical point of view, the relaxation procedure means one deals with composite fuels obtained by mixing microscopically the original fuels, instead of dealing with a juxtaposition of different fuels.

The following result is inspired by [M3] and [JT]. For the relaxation procedure, the separability assumption of the transfer cross section k_i is not necessary, and the boundedness of the functions k_i is sufficient, as it is stated below.

LEMMA 1. *Let $\chi_n = (\chi_i^n)_{i=1,\dots,I}$ be a sequence in $\mathcal{U}_{ad}$. Let σ_n and k_n be the cross sections defined from the configurations χ_n,*

$$\sigma_n(x,v) = \sum_{i=1}^{I} \chi_i^n(x)\sigma_i(v) \quad \textit{and} \quad k_n(x,v,v') = \sum_{i=1}^{I} \chi_i^n(x) k_i(v,v'),$$

where the functions $k_i \in L^\infty(V^2)$ $(i = 1,\dots,I)$, and let K_n be the integral operator defined as in (2) with the kernel k_n. Let $\lambda_n \in \mathbb{R}$ and $\varphi_n \in W_0^2(\Omega \times V)$ be the leading eigenelements satisfying

$$-v \cdot \nabla_x \varphi_n - \sigma_n \varphi_n + K_n \varphi_n = \lambda_n \varphi_n. \tag{11}$$

The eigenfunctions φ_n are normalized by

$$\sum_{i=1}^{I} \Big\| \int_V \varphi_n(.,v) g_i(v) dv \Big\|^2_{L^2(\Omega)} = 1. \tag{12}$$

Then the following hold:

(i) *There exist some functions $\theta_i \in L^\infty(\Omega)$ $(i = 1,\dots,I)$ satisfying*

$$0 \le \theta_i(x) \le 1, \ \sum_{i=1}^{I} \theta_i(x) = 1, \ \int_\Omega \theta_i(x) dx = \mu_i$$

such that, possibly on a subsequence,

$$\chi_i^n \longrightarrow \theta_i \ \textit{weakly} \ \star \ \textit{in} \ L^\infty(\Omega).$$

(ii) *There exist a real λ_θ and a positive function $\varphi_\theta \in W_0^2(\Omega \times V)$ such that, possibly on a subsequence, $\lambda_n \longrightarrow \lambda_\theta$ and*

$$\varphi_n \longrightarrow \varphi_\theta \ \textit{weakly in} \ L^2(\Omega \times V).$$

Moreover, $(\lambda_\theta, \varphi_\theta)$ satisfies

$$-v \cdot \nabla_x \varphi_\theta - \sigma_\theta \varphi_\theta + K_\theta \varphi_\theta = \lambda_\theta \varphi_\theta, \tag{13}$$

where

$$\sigma_\theta(x,v)=\sum_{i=1}^{I}\theta_i(x)\sigma_i(v) \quad \textit{and} \quad k_\theta(x,v,v')=\sum_{i=1}^{I}\theta_i(x)k_i(v,v'), \tag{14}$$

and K_θ is the integral operator defined as in (2) *with the kernel k_θ. Thus λ_θ is the leading eigenvalue of the transport operator $T_\theta + K_\theta$, where T_θ is defined as in* (1) *with the collision frequency σ_θ. Moreover, possibly on a subsequence,*

$$K_n\varphi_n \longrightarrow K_\theta\varphi_\theta \textit{ weakly in } L^2(\Omega\times V) \tag{15}$$

and

$$\int_V \varphi_n(.,v)g_i(v)dv \longrightarrow \int_V \varphi_\theta(.,v)g_i(v)dv \textit{ in } L^2(\Omega). \tag{16}$$

Proof. The point (i) is obvious. We split the proof of (ii) into several parts.

Step 1. *Estimate of the eigenelements.* We already know from the assumptions that the leading eigenvalue λ_n is bounded; more precisely, $\underline{\lambda}\le\lambda_n\le\overline{\lambda}$. So there exists $\lambda_\theta\in\mathbb{R}$ such that, possibly on a subsequence, $\lambda_n\longrightarrow\lambda_\theta$. The leading eigenfunction φ_n satisfies $\varphi_n=(\lambda_n-T_n)^{-1}K_n\varphi_n$, that is,

$$\varphi_n(x,v)=\int_0^{s(x,v)}\mathrm{e}^{-\lambda_n t-\int_0^t\sum_i\chi_i^n(x-sv)\sigma_i(v)ds}\int_V\sum_{i=1}^{I}\chi_i^n(x-tv)k_i(v,v')\varphi_n(x-tv,v')dv'dt.$$

Then

$$\varphi_n(x,v)\le C\int_0^{s(x,v)}\mathrm{e}^{-(\underline{\lambda}+\underline{\sigma}(v))t}dt\int_V\sum_{i=1}^{I}f_i(v)g_i(v')\varphi_n(x-tv,v')dv'$$

and

$$\big\|\varphi_n\big\|_{L^2(\Omega\times V)}\le C\big\|(\underline{\lambda}-\underline{T})^{-1}\big\|\sum_{i=1}^{I}\|f_i\|^2_{L^2(V)}\Big\|\int_V\varphi_n(.,v')g_i(v')dv'\Big\|_{L^2(\Omega)}.$$

According to (12), it follows that φ_n is bounded in $L^2(\Omega\times V)$. Moreover, one easily deduces from (11) that $v\cdot\nabla_x\varphi_n$ is bounded in $L^2(\Omega\times V)$. Thus φ_n is bounded in $W_0^2(\Omega\times V)$. Then there exists a nonnegative function $\varphi_\theta\in W_0^2(\Omega\times V)$ such that, possibly on a subsequence,

$$\varphi_n\longrightarrow\varphi_\theta \text{ weakly in } L^2(\Omega\times V),$$

$$v\cdot\nabla_x\varphi_n\longrightarrow v\cdot\nabla_x\varphi_\theta \text{ weakly in } L^2(\Omega\times V).$$

Step 2. *Homogenization procedure.* First, we show that, possibly on a subsequence,

$$\sigma_n\varphi_n\longrightarrow\sigma_\theta\varphi_\theta \text{ weakly in } L^2(\Omega\times V). \tag{17}$$

Let q be a test function in $L^2(\Omega\times V)$. Since $\sigma_n\varphi_n$ is bounded in $L^2(\Omega\times V)$, by density one can assume that $q(x,v)=q_1(x)q_2(v)$ and even that q_1 and q_2 are continuous functions with compact support. Thus

$$\big(\sigma_n\varphi_n,q\big)_{L^2(\Omega\times V)}=\sum_{i=1}^{I}\int_\Omega\chi_i^n(x)q_1(x)dx\int_V\varphi_n(x,v)\sigma_i(v)q_2(v)dv.$$

Let $d\beta(v) = q_2(v)\sigma_i(v)dv$. By decomposing q_2 into positive and negative parts, we can assume that $d\beta$ is a positive bounded measure such that the hyperplanes have zero $d\beta$-measure. Then Theorem 1 implies

$$\int_V \varphi_n(x,v)\sigma_i(v)q_2(v)dv \longrightarrow \int_V \varphi_\theta(x,v)\sigma_i(v)q_2(v)dv \text{ strongly in } L^2(\Omega).$$

Because Ω is bounded, the previous convergence occurs also in $L^1(\Omega)$. Therefore (i) yields assertion (17), namely, possibly on a subsequence,

$$\big(\sigma_n\varphi_n, q\big)_{L^2(\Omega\times V)} \longrightarrow \big(\sigma_\theta\varphi_\theta, q\big)_{L^2(\Omega\times V)}.$$

Second, we prove (15). Since the transfer cross sections k_i are bounded functions, a domination argument implies that the operators in velocity

$$\varphi \in L^2(V) \longmapsto \int_V k_i(v,v')\varphi(v')dv' \in L^2(V)$$

are compact, and thus can be approximated in the norm operator by finite rank operators. Therefore, because φ_n is bounded in $L^2(\Omega\times V)$, one can assume that k_n is a combination of separable functions $k_n(x,v,v') = \sum_{i=1}^I \chi_i^n(x)f_i(v)g_i(v')$, with f_i and g_i some functions in $L^2(V)$. On the other hand, since the sequence $K_n\varphi_n$ is bounded in $L^2(\Omega\times V)$, one can still choose a test function $q(x,v) = q_1(x)q_2(v)$ with q_1 and q_2 some continuous functions with compact support. Thus we have

$$\big(K_n\varphi_n, q\big)_{L^2(\Omega\times V)} = \sum_{i=1}^I \int_\Omega \chi_i^n(x)q_1(x)dx \int_V \varphi_n(x,v')g_i(v')dv' \int_V f_i(v)q_2(v)dv.$$

The same arguments as before yield

$$\int_V \varphi_n(x,v')g_i(v')dv' \longrightarrow \int_V \varphi_\theta(x,v')g_i(v')dv' \text{ strongly in } L^1(\Omega)$$

and, possibly on a subsequence,

$$\big(K_n\varphi_n, q\big)_{L^2(\Omega\times V)} \longrightarrow \big(K_\theta\varphi_\theta, q\big)_{L^2(\Omega\times V)}.$$

Step 3. *Limit equation.* Passing to the limit in (11), according to the previous steps, it follows that $(\lambda_\theta, \varphi_\theta)$ satisfies (13). Moreover, Theorem 1 implies the convergence (16) and

$$\sum_{i=1}^I \Big\| \int_V \varphi_\theta(.,v)g_i(v)dv \Big\|^2_{L^2(\Omega)} = 1,$$

showing that φ_θ is not null. On the other hand, as the semigroup generated by $T_\theta + K_\theta$ is irreducible, then the only eigenvalue of $T_\theta + K_\theta$ associated with a nonnegative eigenfunction is the leading eigenvalue. Therefore λ_θ is the leading eigenvalue of $T_\theta + K_\theta$. □

We can state now the relaxed or generalized admissible subset of configurations

$$\mathcal{U}_{ad}^\star = \left\{ \theta = (\theta_i)_{1\le i\le I} \in L^\infty(\Omega);\ 0 \le \theta_i(x) \le 1,\ \sum_{i=1}^I \theta_i(x) = 1 \text{ and } \int_\Omega \theta_i(x)dx = \mu_i \right\}.$$

Let $T_\theta + K_\theta$ be the transport operator and let $(\lambda_\theta, \varphi_\theta)$ be the leading eigenelements of the transport operator $T_\theta + K_\theta$ defined as in Lemma 1 and which depend of θ via the cross sections (14). The arguments in Proof 1 are still valid if we replace $\mathcal{U}_{ad}$ by $\mathcal{U}^\star_{ad}$. In particular the leading eigenvalue λ_θ exists regardless of the configuration in $\mathcal{U}^\star_{ad}$ and is bounded. We are now in position to state the relaxation result.

THEOREM 3. *Let*

$$J(\theta) = -\lambda_\theta + \frac{(K_\theta \varphi_\theta, 1)_{L^2(\Omega\times V)}}{\left(\sum_{i=1}^I \| \int_V \varphi_\theta(.,v) g_i(v) dv \|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}},$$

where $\theta \in \mathcal{U}^\star_{ad}$. *Then*

$$\inf_{\chi\in\mathcal{U}_{ad}} J(\chi) = \min_{\theta\in\mathcal{U}^\star_{ad}} J(\theta).$$

Proof. Since for all θ in $\mathcal{U}^\star_{ad}$ there exists a sequence $\chi_n \subset \mathcal{U}_{ad}$ such that

$$\chi_n \longrightarrow \theta \text{ weakly } \star \text{ in } L^\infty(\Omega)$$

(see [MT, Rem. 7]), we deduce from Lemma 1 that

$$\inf_{\chi\in\mathcal{U}_{ad}} J(\chi) = \inf_{\theta\in\mathcal{U}^\star_{ad}} J(\theta).$$

Let us choose a minimizing sequence $\chi_n \subset \mathcal{U}_{ad}$ of the previous infimum. Then, by passing to the limit in $J(\chi_n)$ thanks to Lemma 1, we deduce that the infimum is attained in $\mathcal{U}^\star_{ad}$. ∎

4. Differentiability of the leading eigenvalue. We are concerned in this section with the differentiability of the leading eigenvalue with respect to the design parameter, namely, the configurations in the relaxed admissible subset $\mathcal{U}^\star_{ad}$. We noted in section 2 that the leading eigenelements $(\lambda_\theta, \varphi_\theta)$, depending on $\theta \in \mathcal{U}^\star_{ad}$, are solutions of the implicit eigenvalue equations

$$\varphi = (\lambda - T_\theta)^{-1} K_\theta \varphi,\ \lambda > s(T_\theta),\ \varphi \geq 0, \tag{18}$$

and

$$r_\sigma\big((\lambda - T_\theta)^{-1} K_\theta\big) = 1,\ \lambda > s(T_\theta).$$

Thus the implicit function theorem seems to be a natural approach for getting the differentiability of the leading eigenelements. Nevertheless, the resolvent $(\lambda - T_\theta)^{-1}$ is the Laplace transform of the semigroup generated by T_θ, and differentiating under the time integral turns out to be an important obstacle. So we apply a now-classic method in neutron transport (since [LW]) which consists in transforming the latter eigenvalue problem by integrating it with respect to the velocities. Thus, using the convexity of the space domain Ω and the separability of the transfer cross sections k_i, we obtain a more tractable eigenvalue problem (20) for an integral operator in the space variable whose kernel, given by a time integral, can be more easily estimated than the Laplace transform. We were inspired by the method of [CHM] in which the authors obtain the derivative of the leading eigenvalue with respect to the domain for a simplified model.

The eigenvalue equation (18) for a function φ in $W_0^2(\Omega \times V)$ is written

(19)
$$\varphi(x,v) = \int_0^{s(x,v)} \mathrm{e}^{-\lambda t - \int_0^t \sum_i \theta_i(x-sv)\sigma_i(v)ds} dt \int_V \sum_{i=1}^I \theta_i(x-tv) f_i(v) g_i(v') \varphi(x-tv,v') dv'.$$

Let $\widetilde{\varphi}_i(x) = \int_V g_i(v')\varphi(x,v')dv'$. Multiplying (19) by g_j and integrating over V, it follows that

$$\widetilde{\varphi}_j(x) = \int_V g_j(v) dv \int_0^{s(x,v)} \mathrm{e}^{-\lambda t - \int_0^t \sum_i \theta_i(x-sv)\sigma_i(v)ds} \left(\sum_{i=1}^I \theta_i(x-tv) f_i(v) \widetilde{\varphi}_i(x-tv) \right) dt.$$

Thanks to the convexity of Ω, the change of variable $x' = x - tv$ yields

$$\begin{aligned} (20) \qquad \widetilde{\varphi}_j(x) = \int_\Omega dx' \int_0^\infty & \mathrm{e}^{-\lambda t - \int_0^t \sum_i \theta_i(\frac{s}{t}x + (1-\frac{s}{t})x')\sigma_i(\frac{x-x'}{t})ds} \\ & \times \left(\sum_{i=1}^I \theta_i(x') f_i\left(\frac{x-x'}{t}\right) g_j\left(\frac{x-x'}{t}\right) \widetilde{\varphi}_i(x') \right) \chi_V\left(\frac{x-x'}{t}\right) \frac{dt}{t^N}. \end{aligned}$$

Thus

$$\widetilde{\varphi}_j = \sum_{i=1}^I N_{ij}(\theta,\lambda)\widetilde{\varphi}_i,$$

where $N_{ij}(\theta,\lambda)$ denotes the integral operators on $L^2(\Omega)$,

$$N_{ij}(\theta,\lambda)\varphi(x) = \int_\Omega E_{ij}(\theta,\lambda,x,x')\theta_i(x')\varphi(x')dx',$$

and

$$\begin{aligned} & E_{ij}(\theta,\lambda,x,x') \\ & \quad = \int_0^\infty \mathrm{e}^{-\lambda t - \int_0^t \sum_l \theta_l(\frac{s}{t}x + (1-\frac{s}{t})x')\sigma_l(\frac{x-x'}{t})ds} f_i\left(\frac{x-x'}{t}\right) g_j\left(\frac{x-x'}{t}\right) \chi_V\left(\frac{x-x'}{t}\right) \frac{dt}{t^N}. \end{aligned}$$

Let us denote by $N(\theta,\lambda)$ the bounded operator

$$\widetilde{\varphi} = \left(\widetilde{\varphi}_i\right)_{i=1,\dots,I} \in \left(L^2(\Omega)\right)^I \longmapsto N(\theta,\lambda)\widetilde{\varphi} = \left(\sum_{i=1}^I N_{ij}(\theta,\lambda)\widetilde{\varphi}_i \right)_{j=1,\dots,I} \in \left(L^2(\Omega)\right)^I.$$

Then the eigenvalue problem (18) is equivalent to the problem $\widetilde{\varphi} = N(\theta,\lambda)\widetilde{\varphi}$ for $\lambda > s(T_\theta)$ and $\widetilde{\varphi} \geq 0$. Indeed, let $\widetilde{\varphi}$ be a solution of $\widetilde{\varphi} = N(\theta,\lambda)\widetilde{\varphi}$; then we can check that the function

$$\varphi(x,v) = \int_0^{s(x,v)} \mathrm{e}^{-\lambda t - \int_0^t \sum_i \theta_i(x-sv)\sigma_i(v)ds} dt \sum_{i=1}^I \theta_i(x-tv) f_i(v) \widetilde{\varphi}_i(x-tv)$$

satisfies (18). On the other hand, let us notice that $(\lambda - T_\theta)^{-1}$ and $N(\theta,\lambda)$ are well defined for $\lambda > \underline{\lambda}$ if $\sum_{l=1}^I \theta_l(x) = 1$ and $\theta_l \geq 0$ $(l = 1,\dots,I)$; we refer to Proof 1.

Thus, we are led to replace θ_I by $1-\sum_{l=1}^{I-1}\theta_l$ in the expressions of $N_{ij}(\theta,\lambda)$, and in particular the kernel of $N_{Ij}(\theta,\lambda)$ gets

$$E_{Ij}(\theta,\lambda,x,x')\theta_I(x')=\int_0^\infty \frac{dt}{t^N}\left(1-\sum_{l=1}^{I-1}\theta_l(x')\right) f_I\Big(\frac{x-x'}{t}\Big)g_j\Big(\frac{x-x'}{t}\Big)\chi_V\Big(\frac{x-x'}{t}\Big)$$
$$\times\exp\left(-\lambda t-t\sigma_I\left(\frac{x-x'}{t}\right)\right.$$
$$\left.-\int_0^t\sum_{l=1}^{I-1}\theta_l\left(\frac{s}{t}x+\left(1-\frac{s}{t}\right)x'\right)\left(\sigma_l\left(\frac{x-x'}{t}\right)-\sigma_I\left(\frac{x-x'}{t}\right)\right)ds\right).$$

Let us consider the mapping

$$\Psi:\mathcal{U}\times]\underline{\lambda},\infty[\times\left(L^2(\Omega)\right)^I\longrightarrow\left(L^2(\Omega)\right)^I\times\mathbb{R},$$

$$(\theta,\lambda,\widetilde{\varphi})\longmapsto\left(\Psi_1\left(\theta,\lambda,\widetilde{\varphi}\right),\Psi_2\left(\theta,\lambda,\widetilde{\varphi}\right)\right),$$

where

$$\mathcal{U}=\left\{\theta=(\theta_i)_{1\le i\le I-1}\in\left(L^\infty(\Omega)\right)^{I-1};\ 0\le\theta_i(x)\le 1,\ 0\le\sum_{i=1}^{I-1}\theta_i(x)\le 1\right\}$$

and

$$\Psi_1\left(\theta,\lambda,\widetilde{\varphi}\right)=N(\theta,\lambda)\widetilde{\varphi}-\widetilde{\varphi}\quad\text{and}\quad\Psi_2\left(\theta,\lambda,\widetilde{\varphi}\right)=\sum_{i=1}^{I}\|\widetilde{\varphi}_i\|^2_{L^2(\Omega)}-1.$$

The operator $N(\theta,\lambda)$ is well defined on $\mathcal{U}\times]\underline{\lambda},\infty[$. Thus the eigenelements $(\lambda_\theta,\varphi_\theta)$ satisfy $\Psi\left(\theta,\lambda_\theta,\widetilde{\varphi_\theta}\right)=0$ for all $\theta\in\mathcal{U}$, where $\widetilde{\varphi_\theta}=((\widetilde{\varphi_\theta})_i)_{i=1,\dots,I}$.

We introduce the notation

$$\frac{\partial E_{ij}}{\partial\lambda}(\theta,\lambda,x,x')=-\int_0^\infty\frac{dt}{t^{N-1}}f_i\left(\frac{x-x'}{t}\right)g_j\left(\frac{x-x'}{t}\right)\chi_V\left(\frac{x-x'}{t}\right)$$
$$\times\exp\left(-\lambda t-t\sigma_I\left(\frac{x-x'}{t}\right)\right.$$
$$\left.-\int_0^t\sum_{l=1}^{I-1}\theta_l\left(\frac{s}{t}x+\left(1-\frac{s}{t}\right)x'\right)\left(\sigma_l\left(\frac{x-x'}{t}\right)-\sigma_I\left(\frac{x-x'}{t}\right)\right)ds\right),$$

and for an increment $\delta\theta=(\delta\theta_l)_{l=1,\dots,I-1}$ in $\left(L^\infty(\Omega)\right)^{I-1}$ such that $\theta+\delta\theta\in\mathcal{U}$,

$$\frac{\partial E_{ij}}{\partial\theta_k}(\theta,\lambda,x,x')(\delta\theta_k)=-\int_0^\infty\frac{dt}{t^N}f_i\left(\frac{x-x'}{t}\right)g_j\left(\frac{x-x'}{t}\right)\ \chi_V\left(\frac{x-x'}{t}\right)$$
$$\times\int_0^t\delta\theta_k\left(\frac{s}{t}x+\left(1-\frac{s}{t}\right)x'\right)\left(\sigma_k\left(\frac{x-x'}{t}\right)-\sigma_I\left(\frac{x-x'}{t}\right)\right)ds$$

$$\times \exp\left(-\lambda t - t\sigma_I\left(\frac{x-x'}{t}\right) - \int_0^t \sum_{l=1}^{I-1} \theta_l \left(\frac{s}{t}x + \left(1-\frac{s}{t}\right)x'\right)\left(\sigma_l\left(\frac{x-x'}{t}\right) - \sigma_I\left(\frac{x-x'}{t}\right)\right) ds\right).$$

Lastly, let us denote by $E_{ij}(\theta,\lambda)$, $\frac{\partial E_{ij}}{\partial\lambda}(\theta,\lambda)$, and $\frac{\partial E_{ij}}{\partial\theta_k}(\theta,\lambda)(\delta\theta_k)$ the integral operators on $L^2(\Omega)$ defined, respectively, with the kernels $E_{ij}(\theta,\lambda,x,x')$, $\frac{\partial E_{ij}}{\partial\lambda}(\theta,\lambda,x,x')$, and $\frac{\partial E_{ij}}{\partial\theta_k}(\theta,\lambda,x,x')(\delta\theta_k)$.

THEOREM 4. *The map $\theta\in\mathcal{U}\longmapsto(\lambda_\theta,\widetilde{\varphi_\theta})$ is Fréchet-differentiable on $\mathcal{U}$. The adjoint operator $N^\star(\theta,\lambda_\theta)$ of $N(\theta,\lambda_\theta)$ admits 1 as a simple eigenvalue, associated with the normalized positive eigenfunction $\widetilde{\varphi_\theta}^\star$ in $L^2(\Omega)^I$. Let $\theta\in\mathcal{U}$ and $\delta\theta$ be an increment in $(L^\infty(\Omega))^{I-1}$ such that $\theta+\delta\theta\in\mathcal{U}$. We have*

$$\frac{\partial\lambda_\theta}{\partial\theta}(\delta\theta) = \frac{1}{\gamma(\lambda_\theta,\theta)}\sum_{j=1}^{I}\int_\Omega dx\ \widetilde{\varphi_\theta}_j^\star(x)\left[\sum_{i=1}^{I-1}\sum_{k=1}^{I-1}\frac{\partial E_{ij}}{\partial\theta_k}(\theta,\lambda_\theta)(\delta\theta_k)\theta_i\widetilde{\varphi_\theta}_i(x)\right.$$
$$+\sum_{k=1}^{I-1}\frac{\partial E_{Ij}}{\partial\theta_k}(\theta,\lambda_\theta)(\delta\theta_k)\left(1-\sum_{l=1}^{I-1}\theta_l\right)\widetilde{\varphi_\theta}_I(x)$$
$$\left.+\sum_{i=1}^{I-1}E_{ij}(\theta,\lambda_\theta)\delta\theta_i\widetilde{\varphi_\theta}_i(x)-\sum_{l=1}^{I-1}E_{Ij}(\theta,\lambda_\theta)\delta\theta_l\widetilde{\varphi_\theta}_I(x)\right],$$

where the function

$$\gamma(\lambda_\theta,\theta) = \sum_{j=1}^{I}\int_\Omega \widetilde{\varphi_\theta}_j^\star(x)\left[\int_\Omega\sum_{i=1}^{I-1}\frac{\partial E_{ij}}{\partial\lambda}(\theta,\lambda_\theta)\theta_i\widetilde{\varphi_\theta}_i(x)\right.$$
$$\left.+\frac{\partial E_{Ij}}{\partial\lambda}(\theta,\lambda_\theta)\left(1-\sum_{l=1}^{I-1}\theta_l\right)\widetilde{\varphi_\theta}_I(x)\right]dx \tag{21}$$

is negative. The derivative $\frac{\partial\widetilde{\varphi_\theta}}{\partial\theta}(\delta\theta)$ is the unique solution of the system

$$\sum_{i=1}^{I}\int_\Omega\widetilde{\varphi_\theta}_i(x)\frac{\partial\widetilde{\varphi_\theta}_i}{\partial\theta}(\delta\theta)(x)dx = 0 \tag{22}$$

and

$$N(\theta,\lambda_\theta)\frac{\partial\widetilde{\varphi_\theta}}{\partial\theta}(\delta\theta) - \frac{\partial\widetilde{\varphi_\theta}}{\partial\theta}(\delta\theta) = G\left(\frac{\partial\lambda_\theta}{\partial\theta}(\delta\theta)\right), \tag{23}$$

where $G\left(\frac{\partial\lambda_\theta}{\partial\theta}(\delta\theta)\right)=\left(G_j\left(\frac{\partial\lambda_\theta}{\partial\theta}(\delta\theta)\right)\right)_{j=1,\dots,I}$ *is defined by*

(24)

$$\begin{aligned}G_j\Big(\frac{\partial\lambda_\theta}{\partial\theta}(\delta\theta)\Big)=&\sum_{i=1}^{I-1}\sum_{k=1}^{I-1}\frac{\partial E_{ij}}{\partial\theta_k}(\theta,\lambda_\theta)(\delta\theta_k)\theta_i\widetilde{\varphi_{\theta i}}+\sum_{k=1}^{I-1}\frac{\partial E_{Ij}}{\partial\theta_k}(\theta,\lambda_\theta)(\delta\theta_k)\left(1-\sum_{l=1}^{I-1}\theta_l\right)\widetilde{\varphi_{\theta I}}\\&+\sum_{i=1}^{I-1}E_{ij}(\theta,\lambda_\theta)\delta\theta_i\widetilde{\varphi_{\theta i}}-\sum_{l=1}^{I-1}E_{Ij}(\theta,\lambda_\theta)\delta\theta_l\widetilde{\varphi_{\theta I}}+\frac{\partial\lambda_\theta}{\partial\theta}(\delta\theta)\sum_{i=1}^{I-1}\frac{\partial E_{ij}}{\partial\lambda}(\theta,\lambda_\theta)\theta_i\widetilde{\varphi_{\theta i}}\\&+\frac{\partial\lambda_\theta}{\partial\theta}(\delta\theta)\frac{\partial E_{Ij}}{\partial\lambda}(\theta,\lambda_\theta)\left(1-\sum_{l=1}^{I-1}\theta_l\right)\widetilde{\varphi_{\theta I}}.\end{aligned}$$

LEMMA 2. *The operators* $N_{ij}(\theta,\lambda)$ *are compact on* $L^2(\Omega)$ *for all* $\lambda>\underline{\lambda}$ *and* $\theta\in\mathcal{U}$. *Moreover, the operators* $N_{ij}(\theta+\delta\theta,\lambda+\delta\lambda)$ *are uniformly bounded on* $L^2(\Omega)$ *for every small enough increment* $\delta\lambda\in\mathbb{R}$ *and every small enough increment* $\delta\theta\in\left(L^\infty(\Omega)\right)^{I-1}$ *such that* $\theta+\delta\theta\in\mathcal{U}$.

Proof. Let $\lambda>\underline{\lambda}$ and $\theta\in\mathcal{U}$. Let ω be such that $\lambda>\omega>\underline{\lambda}$. Let $\alpha>0$ be small enough such that $\lambda>\omega+\alpha$. Let $\delta\lambda$ be small enough such that $\lambda+\delta\lambda>\omega+\alpha$. Let $\delta\theta\in\left(L^\infty(\Omega)\right)^{I-1}$ such that $\theta+\delta\theta\in\mathcal{U}$. Since $\omega>\underline{\lambda}$, it follows that $\omega>s(T_{\theta+\delta\theta})$, namely,

$$\omega>-\lim_{t\to\infty}\left\{\operatorname{ess}\inf_{t<s(x,v)}\ \frac{1}{t}\int_0^t\sigma_{\theta+\delta\theta}(x-sv,v)ds\right\}. \tag{25}$$

If $\delta\theta$ is small enough, then there exists $\varepsilon>0$ such that for all $t>0$

$$-\operatorname{ess}\inf_{t<s(x,v)}\ \frac{1}{t}\int_0^t\sigma_{\theta+\delta\theta}(x-sv,v)ds\le-\operatorname{ess}\inf_{t<s(x,v)}\ \frac{1}{t}\int_0^t\sigma_\theta(x-sv,v)ds+\varepsilon.$$

Then, according to (25), there exists $t_0>0$ such that, for ε small enough,

$$-\operatorname{ess}\inf_{t<s(x,v)}\ \frac{1}{t}\int_0^t\sigma_\theta(x-sv,v)ds+\varepsilon<\omega\quad\forall t\ge t_0.$$

Therefore, for all $\delta\theta$ small enough such that $\theta+\delta\theta\in\mathcal{U}$, we have

$$-\operatorname{ess}\inf_{t<s(x,v)}\ \frac{1}{t}\int_0^t\sigma_{\theta+\delta\theta}(x-sv,v)ds<\omega\quad\forall t\ge t_0$$

and

$$\exp\left(-\int_0^t\sigma_{\theta+\delta\theta}\left(\frac{s}{t}x+\left(1-\frac{s}{t}\right)x',\frac{x-x'}{t}\right)ds\right)\chi_V\left(\frac{x-x'}{t}\right)\le\mathrm{e}^{\omega t}\quad\forall t\ge t_0.$$

On the other hand, for $\varphi\ge0$ and $i,j=1,\dots,I$,

$$\begin{aligned}&N_{ij}(\theta+\delta\theta,\lambda+\delta\lambda)\varphi(x)\\&\le\int_\Omega\left(\int_{t_0}^\infty\frac{dt}{t^N}f_M\left(\frac{x-x'}{t}\right)g_M\left(\frac{x-x'}{t}\right)\ \chi_V\left(\frac{x-x'}{t}\right)\mathrm{e}^{-(\lambda+\delta\lambda-\omega)t}\right.\\&\quad+\left.\int_0^{t_0}\frac{dt}{t^N}f_M\left(\frac{x-x'}{t}\right)g_M\left(\frac{x-x'}{t}\right)\ \chi_V\left(\frac{x-x'}{t}\right)\mathrm{e}^{-(\lambda+\delta\lambda)t}\right)\varphi(x')dx'.\end{aligned}$$

Let us denote by $\overline{N}$ the integral operator defined by the right-hand side of the above inequality. Then $N_{ij}(\theta+\delta\theta,\lambda+\delta\lambda)\leq\overline{N}$. By a domination argument, it is sufficient to show that $\overline{N}$ is compact on $L^2(\Omega)$ and uniformly bounded for $\delta\lambda$ small enough. Then, by a convolution argument (see [B, p. 74]), it is enough to show that the function

$$\int_{t_0}^{\infty}\frac{dt}{t^N}f_M\left(\frac{x}{t}\right)g_M\left(\frac{x}{t}\right)\chi_V\left(\frac{x}{t}\right)\mathrm{e}^{-(\lambda+\delta\lambda-\omega)t}+\int_0^{t_0}\frac{dt}{t^N}f_M\left(\frac{x}{t}\right)g_M\left(\frac{x}{t}\right)\chi_V\left(\frac{x}{t}\right)\mathrm{e}^{-(\lambda+\delta\lambda)t}$$

is bounded in $L^1(\Omega)$ by a constant independent of $\delta\lambda$. The change of variables $x=tv$ leads to

$$\begin{aligned}\int_\Omega dx\int_0^{t_0}\frac{dt}{t^N}f_M\left(\frac{x}{t}\right)g_M\left(\frac{x}{t}\right)\ \chi_V\left(\frac{x}{t}\right)\mathrm{e}^{-(\lambda+\delta\lambda)t}&\leq\int_V dv\int_0^{t_0}dt f_M(v)g_M(v)\mathrm{e}^{-(\lambda+\delta\lambda)t}\\&\leq C\|f_Mg_M\|_{L^1(V)}\end{aligned}$$

for all $\delta\lambda$ small enough. The same change of variables as before yields

$$\begin{aligned}\int_\Omega dx\int_{t_0}^{\infty}\frac{dt}{t^N}f_M\left(\frac{x}{t}\right)g_M\left(\frac{x}{t}\right)\ \chi_V\left(\frac{x}{t}\right)\mathrm{e}^{-(\lambda+\delta\lambda-\omega)t}&\leq\int_V dv\int_{t_0}^{\infty}dt f_M(v)g_M(v)\mathrm{e}^{-\alpha t}\\&\leq\frac{\mathrm{e}^{-\alpha t_0}}{\alpha}\|f_Mg_M\|_{L^1(V)}.\qquad\square\end{aligned}$$

Proof of Theorem 4. We split the proof into several parts.

Step 1. *Differentiability of the operators* E_{ij}. We show that the operators E_{ij} are differentiable on $\mathcal{U}\times]\underline{\lambda},\infty[$ and that the differential of E_{ij} at point (θ,λ) with the increments $\delta\lambda\in\mathbb{R}$ and $\delta\theta=(\delta\theta_k)_{k=1,\dots,I-1}\in(L^\infty(\Omega))^{I-1}$ such that $\theta+\delta\theta\in\mathcal{U}$ is

$$DE_{ij}(\lambda,\theta)(\delta\theta,\delta\lambda)=\delta\lambda\frac{\partial E_{ij}}{\partial\lambda}(\theta,\lambda)+\sum_{k=1}^{I-1}\frac{\partial E_{ij}}{\partial\theta_k}(\theta,\lambda)(\delta\theta_k).$$

First, the arguments of the proof of Lemma 2 can be applied to show that the operators $\frac{\partial E_{ij}}{\partial\lambda}(\theta,\lambda)$ and $\frac{\partial E_{ij}}{\partial\theta_k}(\theta,\lambda)(\delta\theta_k)$ are uniformly bounded (and even compact) on $L^2(\Omega)$ for every small enough increments $\delta\lambda\in\mathbb{R}$ and $\delta\theta\in(L^\infty(\Omega))^{I-1}$. Let $\varphi\in L^2(\Omega)$. We split

$$A=E_{ij}(\theta+\delta\theta,\lambda+\delta\lambda)\varphi-E_{ij}(\theta,\lambda)\varphi-\delta\lambda\frac{\partial E_{ij}}{\partial\lambda}(\theta,\lambda)\varphi-\sum_{k=1}^{I-1}\frac{\partial E_{ij}}{\partial\theta_k}(\theta,\lambda)(\delta\theta_k)\varphi$$

as $A=A_1+A_2+A_3$, where

$$A_1=E_{ij}(\theta+\delta\theta,\lambda+\delta\lambda)\varphi-E_{ij}(\theta+\delta\theta,\lambda)\varphi-\delta\lambda\frac{\partial E_{ij}}{\partial\lambda}(\theta+\delta\theta,\lambda)\varphi,$$

$$A_2=E_{ij}(\theta+\delta\theta,\lambda)\varphi-E_{ij}(\theta,\lambda)\varphi-\sum_{k=1}^{I-1}\frac{\partial E_{ij}}{\partial\theta_k}(\theta,\lambda)(\delta\theta_k)\varphi,$$

$$A_3=\delta\lambda\frac{\partial E_{ij}}{\partial\lambda}(\theta+\delta\theta,\lambda)\varphi-\delta\lambda\frac{\partial E_{ij}}{\partial\lambda}(\theta,\lambda)\varphi.$$

Let ω be such that $\lambda > \omega > \underline{\lambda}$. Let $\alpha > 0$ such that $\lambda > \omega + \alpha$. We choose $\delta\lambda$ small enough such that $\lambda + \delta\lambda > \omega + \alpha$. In the proof of Lemma 2, we show there exists $t_0 > 0$ (which only depends on θ) such that

$$\exp\left(-\int_0^t \sigma_{\theta+\delta\theta}\left(\frac{s}{t}x + \left(1-\frac{s}{t}\right)x', \frac{x-x'}{t}\right)\,ds\right)\chi_V\left(\frac{x-x'}{t}\right) \le \mathrm{e}^{\omega t} \quad \forall t \ge t_0$$

for every small enough increment $\delta\theta \in (L^\infty(\Omega))^{I-1}$. The inequality

$$\left|\mathrm{e}^{-\delta\lambda t} - 1 + \delta\lambda t\right| \le |\delta\lambda|^2 t^2\ \mathrm{e}^{|\delta\lambda| t}$$

implies

$$\|A_1\|_{L^2} \le |\delta\lambda|^2 \left\| \int_\Omega dx' \int_0^{t_0} \frac{dt}{t^{N-2}} \mathrm{e}^{-(\lambda+|\delta\lambda|)t} f_M\left(\frac{x-x'}{t}\right) g_M\left(\frac{x-x'}{t}\right)\chi_V\left(\frac{x-x'}{t}\right)|\varphi(x')| \right.$$
$$\left. + \int_\Omega dx' \int_{t_0}^{\infty} \frac{dt}{t^{N-2}} \mathrm{e}^{-(\lambda+|\delta\lambda|-\omega)t} f_M\left(\frac{x-x'}{t}\right) g_M\left(\frac{x-x'}{t}\right)\chi_V\left(\frac{x-x'}{t}\right)|\varphi(x')| \right\|_{L^2(\Omega)}.$$

For $N = 1$ or $N = 2$ we have

$$\begin{aligned}\|A_1\|_{L^2(\Omega)} &\le |\delta\lambda|^2 \left\| \int_\Omega dx' \int_0^{t_0} \frac{dt}{t^{N-2}}\ \mathrm{e}^{-(\omega+\alpha)t}|\varphi(x')| \right. \\ &\qquad \left. + \int_\Omega dx' \int_{t_0}^{\infty} \frac{dt}{t^{N-2}}\ \mathrm{e}^{-\alpha t}|\varphi(x')| \right\|_{L^2(\Omega)} \\ &\le C|\delta\lambda|^2 \|\varphi\|_{L^2(\Omega)}.\end{aligned}$$

Let $v_M = \sup\{|v|, v \in V\}$. Since

$$\frac{x-x'}{t} \in V \Longrightarrow \frac{|x-x'|}{t} \le v_M,$$

we get, for $N \ge 3$,

$$\begin{aligned}\|A_1\|_{L^2(\Omega)} &\le v_M^{N-2}|\delta\lambda|^2 \left\| \int_\Omega dx' \int_0^{t_0} dt\ \mathrm{e}^{-(\omega+\alpha)t} \frac{|\varphi(x')|}{|x-x'|^{N-2}} \right. \\ &\qquad \left. + \int_\Omega dx' \int_{t_0}^{\infty} dt\ \mathrm{e}^{-\alpha t} \frac{|\varphi(x')|}{|x-x'|^{N-2}} \right\|_{L^2(\Omega)} \\ &\le C|\delta\lambda|^2 \|\varphi\|_{L^2(\Omega)}.\end{aligned}$$

Then the inequality

$$\left|\mathrm{e}^{-\beta} - 1 + \beta\right| \le |\beta|^2\ \mathrm{e}^{|\beta|},$$

with

$$\beta = \int_0^t \delta\theta_k\left(\frac{s}{t}x + \left(1-\frac{s}{t}\right)x'\right)\left(\sigma_k\left(\frac{x-x'}{t}\right) - \sigma_I\left(\frac{x-x'}{t}\right)\right)ds,$$

leads to, for $\delta\theta$ small enough,

$$\|A_2\|_{L^2} \le C\|\delta\theta\|_\infty^2 \left\| \int_\Omega dx' \int_0^{t_0} \frac{dt}{t^{N-2}}\ \mathrm{e}^{-\lambda t} f_M\left(\frac{x-x'}{t}\right) g_M\left(\frac{x-x'}{t}\right)\chi_V\left(\frac{x-x'}{t}\right)|\varphi(x')| \right.$$
$$\left. + \int_\Omega dx' \int_{t_0}^{\infty} \frac{dt}{t^{N-2}}\ \mathrm{e}^{-(\lambda-\omega)t} f_M\left(\frac{x-x'}{t}\right) g_M\left(\frac{x-x'}{t}\right)\chi_V\left(\frac{x-x'}{t}\right)|\varphi(x')| \right\|_{L^2(\Omega)}.$$

Above, we used the boundedness of the frequency collisions σ_i and an argument of the proof of Lemma 2 quoted just before. And, as previously, we get

$$\|A_2\|_{L^2(\Omega)} \leq C\|\delta\theta\|_\infty^2 \|\varphi\|_{L^2(\Omega)}.$$

Lastly, the inequality

$$\left|\mathrm{e}^{-\delta\lambda t} - 1\right| \leq |\delta\lambda| t \ \mathrm{e}^{|\delta\lambda| t}$$

yields

$$\|A_3\|_{L^2(\Omega)} \leq |\delta\lambda|^2 \left\| \int_\Omega dx' \int_0^{t_0} \frac{dt}{t^{N-2}} \mathrm{e}^{-\lambda t} f_M\left(\frac{x-x'}{t}\right) g_M\left(\frac{x-x'}{t}\right) \chi_V\left(\frac{x-x'}{t}\right) |\varphi(x')| \right.$$
$$\left. + \int_\Omega dx' \int_{t_0}^\infty \frac{dt}{t^{N-2}} \mathrm{e}^{-(\lambda-\omega)t} f_M\left(\frac{x-x'}{t}\right) g_M\left(\frac{x-x'}{t}\right) \chi_V\left(\frac{x-x'}{t}\right) |\varphi(x')| \right\|_{L^2(\Omega)}.$$

And, as previously, we get

$$\|A_3\|_{L^2(\Omega)} \leq C|\delta\lambda|^2 \|\varphi\|_{L^2(\Omega)}.$$

We can then conclude with

$$\|A\|_{L^2(\Omega)} = o\left(|\delta\lambda|, \|\delta\theta\|_\infty\right) \|\varphi\|_{L^2(\Omega)}.$$

Step 2. *Differentiability of the function* Ψ. The differentiability of Ψ_2 is obvious, and

$$\frac{\partial \Psi_2}{\partial \theta}(\theta, \lambda, \widetilde{\varphi}) = 0, \quad \frac{\partial \Psi_2}{\partial \lambda}(\theta, \lambda, \widetilde{\varphi}) = 0, \quad \text{and} \quad \frac{\partial \Psi_2}{\partial \widetilde{\varphi}}(\theta, \lambda, \widetilde{\varphi})(\delta\widetilde{\varphi}) = 2\sum_{i=1}^I \int_\Omega \widetilde{\varphi}_i(x)\delta\widetilde{\varphi}_i(x)dx.$$

On the other hand, by using the differentiability of the operator E_{ij}, as well as the mean-value theorem with E_{ij}, we get easily

$$E_{ij}(\theta + \delta\theta, \lambda + \delta\lambda)\left((\theta_l + \delta\theta_l)(\widetilde{\varphi}_i + \delta\widetilde{\varphi}_i)\right) - E_{ij}(\theta, \lambda)\theta_l\widetilde{\varphi}_i - \delta\lambda \frac{\partial E_{ij}}{\partial \lambda}(\theta, \lambda)\theta_l\widetilde{\varphi}_i$$
$$- \sum_{k=1}^{I-1} \frac{\partial E_{ij}}{\partial \theta_k}(\theta, \lambda)(\delta\theta_k)\theta_l\widetilde{\varphi}_i - E_{ij}(\theta, \lambda)\delta\theta_l\widetilde{\varphi}_i - E_{ij}(\theta, \lambda)\theta_l\delta\widetilde{\varphi}_i = o\left(\delta\theta, \delta\lambda, \delta\widetilde{\varphi}_j\right).$$

Thus the function Ψ_1 is differentiable and

$$\frac{\partial (\Psi_1)_j}{\partial \lambda}(\theta, \lambda, \widetilde{\varphi})(\delta\lambda) = \sum_{i=1}^{I-1} \delta\lambda \frac{\partial E_{ij}}{\partial \lambda}(\theta, \lambda)\theta_i\widetilde{\varphi}_i + \delta\lambda \frac{\partial E_{Ij}}{\partial \lambda}(\theta, \lambda)\left(1 - \sum_{l=1}^{I-1}\theta_l\right)\widetilde{\varphi}_I,$$

$$\frac{\partial (\Psi_1)_j}{\partial \theta}(\theta, \lambda, \widetilde{\varphi})(\delta\theta) = \sum_{i=1}^{I-1}\sum_{k=1}^{I-1} \frac{\partial E_{ij}}{\partial \theta_k}(\theta, \lambda)(\delta\theta_k)\theta_i\widetilde{\varphi}_i$$
$$+ \sum_{k=1}^{I-1} \frac{\partial E_{Ij}}{\partial \theta_k}(\theta, \lambda)(\delta\theta_k)\left(1 - \sum_{l=1}^{I-1}\theta_l\right)\widetilde{\varphi}_I + \sum_{i=1}^{I-1} E_{ij}(\theta, \lambda)\delta\theta_i\widetilde{\varphi}_i - \sum_{l=1}^{I-1} E_{Ij}(\theta, \lambda)\delta\theta_l\widetilde{\varphi}_I,$$

$$\frac{\partial (\Psi_1)_j}{\partial \widetilde{\varphi}}(\theta, \lambda, \widetilde{\varphi})(\delta\widetilde{\varphi}) = \sum_{i=1}^{I-1} E_{ij}(\theta, \lambda)\theta_i\delta\widetilde{\varphi}_i + E_{Ij}(\theta, \lambda)\left(1 - \sum_{l=1}^{I-1}\theta_l\right)\delta\widetilde{\varphi}_I - \delta\widetilde{\varphi}_j.$$

Step 3. *Differentiability of the leading eigenelements.* First, it is easy to see that the function

$$(\theta, \lambda, \widetilde{\varphi}) \longmapsto \frac{\partial \Psi}{\partial(\lambda, \widetilde{\varphi})}$$

is continuous. Second, let λ_θ and φ_θ be the leading eigenelements for a configuration $\theta \in \mathcal{U}$. We now show that the operator $\frac{\partial \Psi}{\partial(\lambda,\widetilde{\varphi})}(\theta, \lambda_\theta, \widetilde{\varphi_\theta})$ is an isomorphism from $\mathbb{R} \times \left(L^2(\Omega)\right)^I$ onto $\mathbb{R} \times \left(L^2(\Omega)\right)^I$. It follows then from a variant of the implicit function theorem (see [S, Thm. 25, p. 278, and Thm. 26, p. 283]) that the map $\theta \in \mathcal{U} \longmapsto (\lambda_\theta, \widetilde{\varphi_\theta})$ is Fréchet-differentiable.

Let $(\lambda', \phi') \in \mathbb{R} \times \left(L^2(\Omega)\right)^I$. We show there exists a unique $(\lambda, \phi) \in \mathbb{R} \times \left(L^2(\Omega)\right)^I$ such that

$$\frac{\partial \Psi}{\partial(\lambda, \widetilde{\varphi})}(\theta, \lambda_\theta, \widetilde{\varphi_\theta})(\lambda, \phi) = (\lambda', \phi').$$

The above equality is written as

$$2 \sum_{i=1}^{I} \int_\Omega \widetilde{\varphi_\theta}_i(x) \phi_i(x) dx = \lambda' \tag{26}$$

and, for $j = 1, \ldots, I$,

$$\begin{aligned} &\sum_{i=1}^{I-1} \lambda \frac{\partial E_{ij}}{\partial \lambda}(\theta, \lambda_\theta) \theta_i \widetilde{\varphi_\theta}_i + \lambda \frac{\partial E_{Ij}}{\partial \lambda}(\theta, \lambda_\theta) \left(1 - \sum_{l=1}^{I-1} \theta_l\right) \widetilde{\varphi_\theta}_I \\ &+ \sum_{i=1}^{I-1} E_{ij}(\theta, \lambda_\theta) \theta_i \phi_i + E_{Ij}(\theta, \lambda_\theta) \left(1 - \sum_{l=1}^{I-1} \theta_l\right) \phi_I - \phi_j = \phi_j'. \end{aligned} \tag{27}$$

Let $F(\lambda) = (F_j(\lambda))_{j=1,\ldots,I}$, where

$$F_j(\lambda) = \sum_{i=1}^{I-1} \lambda \frac{\partial E_{ij}}{\partial \lambda}(\theta, \lambda_\theta) \theta_i \widetilde{\varphi_\theta}_i + \lambda \frac{\partial E_{Ij}}{\partial \lambda}(\theta, \lambda_\theta) \left(1 - \sum_{l=1}^{I-1} \theta_l\right) \widetilde{\varphi_\theta}_I.$$

Then (27) is equivalent to

$$N(\theta, \lambda_\theta)\phi - \phi = -F(\lambda) + \phi'. \tag{28}$$

Let $N^\star(\theta, \lambda_\theta)$ be the adjoint operator of $N(\theta, \lambda_\theta)$ which is compact on $L^2(\Omega)^I$. Then, since $r_\sigma\left(N^\star(\theta, \lambda_\theta)\right) = r_\sigma\left(N(\theta, \lambda_\theta)\right) = 1$ and since the operator $N^\star(\theta, \lambda_\theta)$ is strictly positive, we deduce from the Krein–Rutman theorem [B, p. 100] that 1 is a simple eigenvalue of $N^\star(\theta, \lambda_\theta)$ associated with the normalized positive eigenfunction denoted by $\widetilde{\varphi_\theta}^\star$. From the Fredholm alternative, (28) has a solution if and only if

$$\left(-F(\lambda) + \phi', \widetilde{\varphi_\theta}^\star\right)_{L^2(\Omega)^I} = 0,$$

and all solutions are given by $\phi = \phi_0 + \beta \widetilde{\varphi_\theta}$, where β is a constant and ϕ_0 is a solution of (28). Then it follows that

$$\lambda = \frac{\left(\phi', \widetilde{\varphi_\theta}^\star\right)_{L^2(\Omega)^I}}{\mu(\theta, \lambda_\theta)},$$

where

$$\mu(\theta,\lambda_\theta)=\sum_{i=1}^{I-1}\int_\Omega \widetilde{\varphi_\theta}_i^\star(x)\frac{\partial E_{ij}}{\partial\lambda}(\theta,\lambda_\theta)\theta_i\widetilde{\varphi_\theta}_i(x)dx + \int_\Omega \widetilde{\varphi_\theta}_I^\star(x)\frac{\partial E_{Ij}}{\partial\lambda}(\theta,\lambda_\theta)\left(1-\sum_{l=1}^{I-1}\theta_l\right)\widetilde{\varphi_\theta}_I(x)dx.$$

We notice that $\mu(\theta,\lambda_\theta)<0$ because the eigenfunctions $\widetilde{\varphi_\theta}^\star$ and $\widetilde{\varphi_\theta}$, as well as the functions f_m and g_m, are positive. Lastly, ϕ is uniquely determined by (26).

Step 4. *Calculus of the derivatives of* λ_θ *and* $\widetilde{\varphi_\theta}$ *with respect to* θ. To simplify the notation, we denote $\lambda'(\delta\theta)=\frac{\partial\lambda_\theta}{\partial\theta}(\delta\theta)$ and $\widetilde{\varphi}'(\delta\theta)=\frac{\partial\widetilde{\varphi_\theta}}{\partial\theta}(\delta\theta)$. The derivative of $\Psi(\theta,\lambda_\theta,\widetilde{\varphi_\theta})=0$ gives

$$\sum_{i=1}^{I}\int_\Omega \widetilde{\varphi_\theta}_i(x)\widetilde{\varphi}_i'(\delta\theta)(x)dx=0 \tag{29}$$

and, for $j=1,\dots,I$,

$$\begin{aligned}\sum_{i=1}^{I-1}\sum_{k=1}^{I-1}\frac{\partial E_{ij}}{\partial\theta_k}(\theta,\lambda_\theta)(\delta\theta_k)\theta_i\widetilde{\varphi_\theta}_i+\sum_{k=1}^{I-1}\frac{\partial E_{Ij}}{\partial\theta_k}(\theta,\lambda_\theta)(\delta\theta_k)\left(1-\sum_{l=1}^{I-1}\theta_l\right)\widetilde{\varphi_\theta}_I\\ +\sum_{i=1}^{I-1}E_{ij}(\theta,\lambda_\theta)\delta\theta_i\widetilde{\varphi_\theta}_i-\sum_{l=1}^{I-1}E_{Ij}(\theta,\lambda_\theta)\delta\theta_l\widetilde{\varphi_\theta}_I+\lambda'(\delta\theta)\sum_{i=1}^{I-1}\frac{\partial E_{ij}}{\partial\lambda}(\theta,\lambda_\theta)\theta_i\widetilde{\varphi_\theta}_i\\ +\lambda'(\delta\theta)\frac{\partial E_{Ij}}{\partial\lambda}(\theta,\lambda_\theta)\left(1-\sum_{l=1}^{I-1}\theta_l\right)\widetilde{\varphi_\theta}_I+\sum_{i=1}^{I-1}E_{ij}(\theta,\lambda_\theta)\theta_i\widetilde{\varphi}_i'(\delta\theta)\\ +E_{Ij}(\theta,\lambda_\theta)\left(1-\sum_{l=1}^{I-1}\theta_l\right)\widetilde{\varphi}_I'(\delta\theta)-\widetilde{\varphi}_j'(\delta\theta)=0.\end{aligned} \tag{30}$$

Multiplying (30) by $\widetilde{\varphi_\theta}_j^\star$, integrating over Ω, and noticing that

$$\left(N(\theta,\lambda_\theta)\widetilde{\varphi}'(\delta\theta),\widetilde{\varphi_\theta}^\star\right)_{L^2(\Omega)^I}=\left(\widetilde{\varphi}'(\delta\theta),N^\star(\theta,\lambda_\theta)\widetilde{\varphi_\theta}^\star\right)_{L^2(\Omega)^I}=\left(\widetilde{\varphi}'(\delta\theta),\widetilde{\varphi_\theta}^\star\right)_{L^2(\Omega)^I},$$

we get

$$\begin{aligned}\lambda'(\delta\theta)\ \gamma(\lambda_\theta,\theta)=\sum_{j=1}^{I}\int_\Omega\widetilde{\varphi_\theta}_j^\star(x)\Bigg[\sum_{i=1}^{I-1}\sum_{k=1}^{I-1}\frac{\partial E_{ij}}{\partial\theta_k}(\theta,\lambda_\theta)(\delta\theta_k)\theta_i\widetilde{\varphi_\theta}_i(x)\\ +\sum_{k=1}^{I-1}\frac{\partial E_{Ij}}{\partial\theta_k}(\theta,\lambda_\theta)(\delta\theta_k)\left(1-\sum_{l=1}^{I-1}\theta_l\right)\widetilde{\varphi_\theta}_I(x)\ +\sum_{i=1}^{I-1}E_{ij}(\theta,\lambda_\theta)\delta\theta_i\widetilde{\varphi_\theta}_i(x)\\ -\sum_{l=1}^{I-1}E_{Ij}(\theta,\lambda_\theta)\delta\theta_l\widetilde{\varphi_\theta}_I(x)\Bigg]dx,\end{aligned}$$

where $\gamma(\lambda_\theta,\theta)$ is defined by (21). We note that $\gamma(\theta,\lambda_\theta)<0$ since the eigenfunctions $\widetilde{\varphi_\theta}^\star$ and $\widetilde{\varphi_\theta}$, as well as the functions f_m and g_m, are positive. On the other hand, (30) is written as

$$N(\theta,\lambda_\theta)\widetilde{\varphi}'(\delta\theta)-\widetilde{\varphi}'(\delta\theta)=G(\lambda'(\delta\theta)),$$

where $G(\lambda'(\delta\theta)) = (G_j(\lambda'(\delta\theta)))_{j=1,\dots,I}$ is defined by (24). The necessary condition

$$\left(G(\lambda'(\delta\theta)), \widetilde{\varphi_\theta}^\star\right)_{L^2(\Omega)^I} = 0$$

to solve (30) is ensured by the computation of $\lambda'(\delta\theta)$. Lastly, the solution $\widetilde{\varphi}'(\delta\theta)$, defined up to the addition of a multiple of $\widetilde{\varphi_\theta}$, is uniquely determined by (29). □

5. Optimality condition. Let H be the function

$$\theta \in \mathcal{U} \longmapsto \frac{(K_\theta \varphi_\theta, 1)_{L^2(\Omega\times V)}}{\left(\sum_{i=1}^I \|\widetilde{\varphi_{\theta i}}\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}}.$$

As an immediate consequence of Theorem 4 we have the following.

COROLLARY 1. *The norms $\|\widetilde{\varphi_{\theta i}}\|^2_{L^2(\Omega)}$ are positive for all $i = 1, \dots, I$. The function H is Fréchet-differentiable on $\mathcal{U}$ and*

$$\frac{\partial H}{\partial\theta}(\delta\theta) = \frac{1}{\left(\sum_{i=1}^I \|\widetilde{\varphi_{\theta i}}\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}} \left(\int_\Omega \frac{\partial\widetilde{\varphi_{\theta I}}}{\partial\theta}(\delta\theta)(x)dx + \sum_{i=1}^{I-1}\left(\int_\Omega \delta\theta_i(x)\widetilde{\varphi_{\theta i}}(x)dx \right.\right.$$
$$\left.\left. + \int_\Omega \theta_i(x)\frac{\partial\widetilde{\varphi_{\theta i}}}{\partial\theta}(\delta\theta)(x)dx - \int_\Omega \delta\theta_i(x)\widetilde{\varphi_{\theta I}}(x)dx - \int_\Omega \theta_i(x)\frac{\partial\widetilde{\varphi_{\theta I}}}{\partial\theta}(\delta\theta)(x)dx\right)\right)$$
$$- \frac{(K_\theta\varphi_\theta, 1)_{L^2(\Omega\times V)}}{\left(\sum_{i=1}^I \|\widetilde{\varphi_{\theta i}}\|^2_{L^2(\Omega)}\right)^{\frac{3}{2}}} \times \sum_{i=1}^I \int_\Omega \widetilde{\varphi_{\theta i}}(x)\frac{\partial\widetilde{\varphi_{\theta i}}}{\partial\theta}(\delta\theta)(x)dx.$$

We are now in position to give an optimality condition for the minimizers of the cost function J upon the relaxed admissible subset $\mathcal{U}^*_{ad}$. Let us recall, when replacing θ_I by $1 - \sum_{i=1}^{I-1}\theta_i$ for $\theta = (\theta_i)_{1\le i\le I} \in \mathcal{U}^*_{ad}$, that one considers in fact configurations $\theta = (\theta_i)_{1\le i\le I-1}$ satisfying

$$\theta \in \mathcal{U} = \left\{\theta = (\theta_i)_{1\le i\le I-1} \in (L^\infty(\Omega))^{I-1};\ 0 \le \theta_i(x) \le 1,\ 0 \le \sum_{i=1}^{I-1}\theta_i(x) \le 1\right\}$$

as well as the volume constraints $\int_\Omega \theta_i(x)dx = \mu_i$ $(i = 1, \dots, I-1)$. Applying the general Lagrange multipliers theorem [Z, Thm. 48.B, p. 417], we obtain the following.

THEOREM 5. *If θ_0 is a minimizer of J on the subset $\mathcal{U}^*_{ad}$, then there exist $I-1$ constants C_i $(i = 1, \dots, I-1)$ such that*

$$-\frac{\partial\lambda_{\theta_0}}{\partial\theta}(\theta - \theta_0) + \frac{\partial H}{\partial\theta}(\theta_0)(\theta - \theta_0) + \sum_{i=1}^{I-1} C_i \int_\Omega (\theta_i(x) - (\theta_0)_i(x))\,dx \ge 0 \tag{31}$$

for all $\theta \in \mathcal{U}$.

Appendix A. Disappearance of the leading eigenvalue. The disappearance phenomenon, for domains with a small diameter, was first observed by [AM] and was generalized by [M1, Thm. 1]. Here one proves that the leading eigenvalue can disappear if the measure of the space domain Ω (not necessarily convex) becomes too small. This result shows that the disappearance phenomenon can also occur in

the case of a medium containing absorbing matter (modeled by holes) in too large a quantity.

PROPOSITION 2. *Let us assume that* $0 \in V$, *the collision frequency* σ *is space homogeneous, and*

$$\liminf_{v\to 0} \sigma(v) = \inf_{v\in V} \sigma(v). \tag{32}$$

Moreover, let us assume that the kernels $k(x,v,v')$ *and* $|v|^{-1}k(x,v,v')$ *belong to* $L^\infty(\Omega\times V^2)$ *and* $N<p<\infty$.

If the measure $|\Omega|$ *of* Ω *is small enough, then* $\sigma_{as}(T+K)=\varnothing$. *More precisely, there exists a positive constant* M, *which depends on* p, N, $v_M=\sup\{|v|, v\in V\}$, *the measure* $|V|$ *of* V, $\operatorname{ess\,sup}\left(|v|^{-1}k(x,v,v')\right)$, *and the diameter* d *of* Ω *such that* $\sigma_{as}(T+K)=\varnothing$ *if* $|\Omega|<M$.

Proof. The proof consists in showing that $r_\sigma\left((\lambda-T)^{-1}K\right)<1$ for all $\lambda>s(T)$ and relies on the inequality

$$\left(r_\sigma\left((\lambda-T)^{-1}K\right)\right)^2 = r_\sigma\left(\left((\lambda-T)^{-1}K\right)^2\right) \le \left\|\left((\lambda-T)^{-1}K\right)^2\right\|.$$

From (6) and (32), it follows that

$$s(T) = -\inf_{v\in V}\sigma(v).$$

According to the assumptions, the integral operator denoted by $|v|^{-1}K$ of kernel $|v|^{-1}k(x,v,v')$ is bounded on $L^p(\Omega\times V)$; moreover,

$$\left((\lambda-T)^{-1}K\right)^2 = |v|(\lambda-T)^{-1}|v|^{-1}K(\lambda-T)^{-1}K.$$

First, we estimate the norm of the operator $|v|(\lambda-T)^{-1}$ as in [M1, Thm. 1]. We recall that

$$(\lambda-T)^{-1}\varphi(x,v) = \int_0^{s(x,v)} \mathrm{e}^{-(\lambda+\sigma(v))t}\varphi(x-tv,v)\,dt. \tag{33}$$

The change of variable $t\to t|v|$ in (33) leads to

$$|v|(\lambda-T)^{-1}\varphi(x,v) = \int_0^{s(x,\frac{v}{|v|})} \mathrm{e}^{-(\lambda+\sigma(v))\frac{t}{|v|}}\varphi\left(x-t\frac{v}{|v|},v\right)dt.$$

Thus

$$\begin{aligned}
&\left\||v|(\lambda-T)^{-1}\varphi\right\|_{L^p}^p\\
&\le \int_{\Omega\times V}\left(\int_0^{s(x,\frac{v}{|v|})} \mathrm{e}^{-(\lambda+\sigma(v))\frac{tp'}{|v|}}dt\right)^{\frac{p}{p'}}\left(\int_0^{s(x,\frac{v}{|v|})}\left|\varphi\left(x-t\frac{v}{|v|},v\right)\right|^p dt\right)dxdv\\
&\le d^{\frac{p}{p'}}\int_{\Omega\times V}\int_0^d\left|\varphi\left(x-t\frac{v}{|v|},v\right)\right|^p dtdxdv \le d^{\frac{p}{p'}+1}\|\varphi\|_{L^p}^p = d^p\|\varphi\|_{L^p}^p,
\end{aligned}$$

where $\frac{1}{p}+\frac{1}{p'}=1$. Therefore

$$\left\||v|(\lambda-T)^{-1}\right\| \le d. \tag{34}$$

Second, as the operator $|v|^{-1}K(\lambda - T)^{-1}K$ is positive, it is sufficient to estimate its norm on the subset of nonnegative functions. Let φ be a nonnegative function in $L^p(\Omega \times V)$. We have

$$|v|^{-1}K(\lambda - T)^{-1}K\varphi(x,v) = \int_V dv' \ |v|^{-1}k(x,v,v') \int_0^{s(x,v')} dt \ \mathrm{e}^{-(\lambda+\sigma(v'))t} \times \int_V dv'' \ k(x - tv', v', v'')\varphi(x - tv', v'').$$

The change of variable $x' = x - tv'$ yields

$$|v|^{-1}K(\lambda - T)^{-1}K\varphi(x,v) \leq \int_{\Omega\times V} H(x,x',v,v'')\varphi(x',v'')dx'dv'' := H\varphi, \tag{35}$$

where

$$H(x,x',v,v'') = \int_0^\infty |v|^{-1}k\left(x,v,\frac{x-x'}{t}\right) \mathrm{e}^{-\left(\lambda+\sigma(\frac{x-x'}{t})\right)t} \times \ k\left(x,\frac{x-x'}{t},v''\right)\chi_V\left(\frac{x-x'}{t}\right)\frac{dt}{t^N}.$$

Let us note that we have equality in (35) if Ω is a convex subset. By positivity

$$\left\||v|^{-1}K(\lambda - T)^{-1}K\right\| \leq \|H\|.$$

On the other hand

$$\|H\| \leq \left(\int_{\Omega\times V}\left(\int_{\Omega\times V}|H(x,x',v,v'')|^{p'} \ dx'dv''\right)^{\frac{p}{p'}}dxdv\right)^{\frac{1}{p}}. \tag{36}$$

According to the assumptions and since

$$\frac{x-x'}{t} \in V \Longrightarrow \frac{|x-x'|}{v_M} \leq t,$$

we get

$$\int_{\Omega\times V}|H(x,x',v,v'')|^{p'} \ dx'dv'' \leq C\int_{\Omega\times V}\left|\int_{\frac{|x-x'|}{v_M}}^\infty |x-x'|\frac{dt}{t^{N+1}}\right|^{p'}dx'dv'' \leq C \ |V| \ v_M^{Np'} \ N^{-p'}\int_\Omega \frac{dx'}{|x-x'|^{p'(N-1)}}.$$

By a convolution argument, the function

$$g: \ x \longmapsto \int_\Omega \frac{dx'}{|x-x'|^{p'(N-1)}}$$

is continuous on $\mathbb{R}^N$ if $p' < \frac{N}{N-1}$, that is, if $p > N$. Therefore g is bounded on Ω by a constant which depends on p and N. One can conclude, thanks to (34) and (36), with the following:

$$\left\|\left((\lambda - T)^{-1}K\right)^2\right\| \leq C(p,N) \ v_m^N \ |V| \ d \ |\Omega|^{\frac{1}{p}}. \qquad \square$$

Appendix B. Proof of Proposition 1. According to (5) we deduce $s(T_\chi) \leq -\underline{\sigma}$ for all $\chi \in \mathcal{U}_{ad}$. Let the unbounded penetration operators $\overline{T}$ be defined as in (1) but with the collision frequencies $\overline{\sigma}$, and let the bounded integral operator $\underline{K}$ be defined as in (2) but with the kernel $f_m(v)g_m(v')$. Let φ be a nonnegative function in $L^p(\Omega \times V)$ and $\lambda > -\underline{\sigma}$. It is easy to see that

$$(\lambda - T_\chi)^{-1} K_\chi \varphi(x,v) \geq \int_0^{s(x,v)} \mathrm{e}^{-(\lambda+\overline{\sigma})t} \int_V f_m(v) g_m(v') \varphi(x - tv, v') dv'.$$

Therefore

$$(\lambda - T_\chi)^{-1} K_\chi \geq (\lambda - \overline{T})^{-1} \underline{K} \quad \forall \chi \in \mathcal{U}_{ad}$$

and thus

$$r_\sigma \left((\lambda - T_\chi)^{-1} K_\chi \right) \geq r_\sigma \left((\lambda - \overline{T})^{-1} \underline{K} \right) \quad \forall \chi \in \mathcal{U}_{ad}.$$

We search some conditions ensuring there exists some $\underline{\lambda} > -\underline{\sigma}$ such that $r_\sigma \left((\underline{\lambda} - \overline{T})^{-1} \underline{K} \right) \geq 1$. In this case we can deduce that the asymptotic spectrum $\sigma_{as}(T_\chi + K_\chi)$ is nonempty for all χ in $\mathcal{U}_{ad}$ and $\lambda_\chi \geq \underline{\lambda}$.

Let us notice that $(\lambda - \overline{T})^{-1} \underline{K}$ is power compact. On the other hand $(\lambda - \overline{T})^{-1} \underline{K}$ is an irreducible operator; indeed one can show that $\left((\lambda - \overline{T})^{-1} \underline{K} \right)^2$ is strictly positive (see the proof of [M2, Thm. 5.15, p. 109]). It follows that $r_\sigma \left((\lambda - \overline{T})^{-1} \underline{K} \right)$ is the only eigenvalue of $(\lambda - \overline{T})^{-1} \underline{K}$ associated with a positive eigenfunction (see [Ma]). We are thus led to study the eigenvalues of $(\lambda - \overline{T})^{-1} \underline{K}$, and thus we consider the following eigenvalue problem:

$$\gamma \varphi(x,v) = \int_0^{s(x,v)} \mathrm{e}^{-(\lambda+\overline{\sigma})t} \int_V f_m(v) g_m(v') \varphi(x - tv, v') dv'.$$

Multiplying the previous equation by $g_m(v)$ and integrating over V, we get

$$\gamma \psi(x) = \int_V dv \; f_m(v) g_m(v) \int_0^{s(x,v)} \mathrm{e}^{-(\lambda+\overline{\sigma})t} \psi(x - tv) dt,$$

where $\psi(x) = \int_V g_m(v) \varphi(x,v) dv$. Let us extend ψ by 0 outside Ω as well as $f_m g_m$ and σ by 0 outside V. The change of variable $x' = x - tv$ yields, thanks to the convexity of Ω,

$$\gamma \psi(x) = \int_\Omega N_\lambda(x - x') \psi(x') dx' := N_\lambda \psi(x),$$

where

$$N_\lambda(x) = \int_0^\infty f_m \left(\frac{x}{t} \right) g_m \left(\frac{x}{t} \right) \; \mathrm{e}^{-(\lambda+\overline{\sigma})t} \; \frac{dt}{t^N}.$$

As the kernel N_λ is positive, the convolution operator N_λ on $L^p(\Omega)$ is strictly positive. And according to the Krein–Rutman theorem (see [B, p. 100]), $r_\sigma(N_\lambda)$ is the only eigenvalue associated with a positive eigenfunction. Therefore

$$r_\sigma(N_\lambda) = r_\sigma \left((\lambda - \overline{T})^{-1} \underline{K} \right).$$

On the other hand

$$\int_{\mathbb{R}^N} N_\lambda(x)dx = \int_V f_m(v)g_m(v)dv \int_0^\infty \mathrm{e}^{-(\lambda+\overline{\sigma})t}dt \leq \frac{\|f_m g_m\|_{L^1(V)}}{\lambda+\overline{\sigma}};$$

therefore N_λ maps $L^q(\Omega)$ into itself for all $1 \leq q \leq \infty$ and N_λ is compact in $L^q(\Omega)$ for all $1 \leq q < \infty$ (see [B, p. 74]). Consequently the spectrum of N_λ is the same in all $L^q(\Omega)$ for $1 \leq q < \infty$ (see [D, Thm. 1.6.1, p. 35]). So it is sufficient to study its spectrum in $L^2(\Omega)$. By passing in the Fourier variable, $(N_\lambda\psi)^\wedge = \widehat{N_\lambda}\widehat{\psi}$, where

$$\widehat{\psi}(\xi) = (2\pi)^{-\frac{N}{2}} \int_{\mathbb{R}^N} \mathrm{e}^{-ix.\xi}\psi(x)dx$$

and

$$\begin{aligned}
\widehat{N_\lambda}(\xi) &= (2\pi)^{-\frac{N}{2}} \int_{\mathbb{R}^N} dx\ \mathrm{e}^{-ix.\xi} \int_0^\infty f_m\left(\frac{x}{t}\right) g_m\left(\frac{x}{t}\right)\ \mathrm{e}^{-(\lambda+\overline{\sigma})t}\ \frac{dt}{t^N} \\
&= (2\pi)^{-\frac{N}{2}} \int_V dv \int_0^\infty f_m(v)g_m(v)\ \mathrm{e}^{-(\lambda+\overline{\sigma}+iv.\xi)t}dt \\
&= (2\pi)^{-\frac{N}{2}} \int_V \frac{(\lambda+\overline{\sigma})f_m(v)g_m(v)}{(\lambda+\overline{\sigma})^2+(v.\xi)^2}\ dv.
\end{aligned}$$

The last equality is due to the evenness of the function $f_m g_m$. Moreover,

$$(N_\lambda\psi,\psi)_{L^2(\Omega)} = \int_{\mathbb{R}^N} d\xi\ \ |\psi(\xi)|^2 \int_V \frac{(\lambda+\overline{\sigma})f_m(v)g_m(v)}{(\lambda+\overline{\sigma})^2+(v.\xi)^2}\ dv.$$

Therefore N_λ is a positive self-adjoint compact operator in $L^2(\Omega)$. We deduce furthermore that

$$\|N_\lambda\|_{\mathcal{L}(L^2(\Omega))} = r_\sigma(N_\lambda) = r_\sigma\left((\lambda-\overline{T})^{-1}\underline{K}\right).$$

Let B_r be the biggest ball included in Ω, with radius r. By positivity,

$$\|N_\lambda\|_{\mathcal{L}(L^2(\Omega))} \geq \|N_\lambda\|_{\mathcal{L}(L^2(B_r))}.$$

Let B be the unit ball of $\mathbb{R}^N$ and $\phi \in L^2(B)$ such that $\|\phi\|_{L^2(B)} = 1$. Let $\psi(x) = r^{-\frac{N}{2}}\phi\left(\frac{x}{r}\right)$; then $\|\psi\|_{L^2(B_r)} = 1$. Finally, let $R > 0$. We have

$$\begin{aligned}
\|N_\lambda\|_{\mathcal{L}(L^2(B_r))} &\geq (N_\lambda\psi,\psi)_{L^2(B_r)} \\
&= \int_{\mathbb{R}^N} d\xi\ r^N \left|\widehat{\phi}(r\xi)\right|^2 \int_V \frac{(\lambda+\overline{\sigma})f_m(v)g_m(v)}{(\lambda+\overline{\sigma})^2+(v.\xi)^2}\ dv \\
&= \int_{\mathbb{R}^N} d\xi\ \ \left|\widehat{\phi}(\xi)\right|^2 \int_V \frac{(\lambda+\overline{\sigma})f_m(v)g_m(v)}{(\lambda+\overline{\sigma})^2+\left(\frac{v.\xi}{r}\right)^2}\ dv \\
&\geq \int_{|\xi|<R} d\xi\ \ \left|\widehat{\phi}(\xi)\right|^2 \int_V \frac{(\lambda+\overline{\sigma})f_m(v)g_m(v)}{(\lambda+\overline{\sigma})^2+\frac{R^2v_M^2}{r^2}}\ dv \\
&= \varepsilon(R)\ \ \|f_m g_m\|_{L^1(V)}\ \ \frac{(\lambda+\overline{\sigma})}{(\lambda+\overline{\sigma})^2+\frac{R^2v_M^2}{r^2}},
\end{aligned}$$

where

$$\varepsilon(R) = \int_{|\xi|<R} \left|\widehat{\phi}(\xi)\right|^2 d\xi.$$

Let us take now $\alpha > 0$ small enough and $\lambda = -\underline{\sigma} + \alpha$. Thus we have

$$\varepsilon(R)\ \|f_m g_m\|_{L^1(V)}\ \frac{(-\underline{\sigma} + \alpha + \overline{\sigma})}{(-\underline{\sigma} + \alpha + \overline{\sigma})^2 + \dfrac{R^2 v_M^2}{r^2}} > 1$$

if and only if

$$r^2 > \frac{v_M^2 R^2}{\Big(\varepsilon(R)\ \|f_m g_m\|_{L^1(V)} - (-\underline{\sigma} + \alpha + \overline{\sigma})\Big)(-\underline{\sigma} + \alpha + \overline{\sigma})}.$$

We just have to choose R big enough and α small enough to ensure

$$\varepsilon(R)\ \|f_m g_m\|_{L^1(V)} - (-\underline{\sigma} + \alpha + \overline{\sigma}) > 0,$$

which is possible according to the assumptions and because

$$\lim_{R\to 0} \int_{|\xi|<R} \left|\widehat{\phi}(\xi)\right|^2 d\xi = \int_{\mathbb{R}^N} \left|\widehat{\phi}(\xi)\right|^2 d\xi = 1.$$

Thus we have proved the first part of the proposition.

Let φ_χ be a positive eigenfunction associated with the leading eigenvalue λ_χ. From the eigenvalue equation we deduce

$$\varphi_\chi(x,v) \le \int_0^{s(x,v)} \mathrm{e}^{-(\lambda_\chi + \underline{\sigma})t} dt \int_V f_M(v) g_M(v') \varphi_\chi(x - tv, v') dv'.$$

We extend φ_χ by 0 outside $\Omega \times V$ and we denote $\phi_\chi(v) = \int_\Omega \varphi_\chi(x,v)dx$. We get

$$\phi_\chi(v) \le \frac{\overline{K}\phi_\chi(v)}{\lambda_\chi + \underline{\sigma}}.$$

Since ϕ_χ is positive it follows that

$$\|\phi_\chi\|_{L^2(V)} \le \frac{\left\|\overline{K}\right\| \|\phi_\chi\|_{L^2(V)}}{\lambda_\chi + \underline{\sigma}}.$$

Therefore $\lambda_\chi \le -\underline{\sigma} + \left\|\overline{K}\right\|$. ∎

REFERENCES

[AC1] G. Allaire and C. Castro, *A new approach for the optimal distribution of assemblies in a nuclear reactor*, Numer. Math., 89 (2001), pp. 1–29.

[AC2] G. Allaire and C. Castro, *Optimization of nuclear fuel reloading by the homogenization method*, Structural and Multidisciplinary Optim., 24 (2002), pp. 1–22.

[AM] S. Albertoni and B. Montagnini, *On the spectrum of neutron transport equation in finite bodies*, J. Math. Anal. Appl., 13 (1966), pp. 19–48.

[B] H. Brezis, *Analyse fonctionnelle*, Masson, Paris, 1983.

[BG] G. I. Bell and S. Glasstone, *Nuclear Reactor Theory*, Krieger, Melbourne, FL, 1970.

[C] P. Clement, H. J. A. M. Heijmans, S. Angenent, C. J. van Duijn, and B. de Pagter, *One Parameter Semigroups*, North-Holland, Amsterdam, 1987.

[CHM] M. Choulli, A. Henrot, and M. Mokhtar-Kharroubi, *Domain derivative of the leading eigenvalue of a model transport operator*, Transport Theory Statist. Phys., 28 (1999), pp. 403–418.

[D] E. B. Davies, *Heat Kernels and Spectral Theory*, Cambridge Tracts in Math. 92, Cambridge University Press, Cambridge, UK, 1989.

[DG] L. Dumas and F. Golse, *Homogenization of transport equations*, SIAM J. Appl. Math., 60 (2000), pp. 1447–1470.

[G] P. Gerard, *Microlocal defect measures*, Comm. Partial Differential Equations, 16 (1991), pp. 1761–1794.

[GLPS] F. Golse, P. L. Lions, B. Perthame, and R. Sentis, *Regularity of the moments of the solution of a transport equation*, J. Funct. Anal., 76 (1988), pp. 110–125.

[GMP] W. Greenberg, C. van der Mee, and V. Protopopescu, *Boundary Value Problems in Abstract Kinetic Theory*, Birkhäuser Verlag, Basel, 1987.

[JT] K. Jarmouni-Idrissi and L. Thevenot, *Homogenization of a nonlinear neutron transport equation*, Transport Theory Statist. Phys., 31 (2002), pp. 93–123.

[KLH] H. G. Kaper, C. G. Lekkerkerker, and J. Hejtmanek, *Spectral Methods in Linear Transport Theory*, Birkhäuser Verlag, Basel, 1982.

[L] S. Levine, *In-core fuel management of four reactor types*, in Handbook of Nuclear Reactor Calculations, Vol. II, Y. Ronen, ed., CRC Press, Boca Raton, FL, 1986, pp. 87–201.

[LW] J. Lehner and M. Wing, *On the spectrum of an unsymmetric operator arising in the transport theory of neutrons*, Comm. Pure Appl. Math., 8 (1955), pp. 217–234.

[Ma] I. Marek, *Frobenius theory of positive operators: Comparison theorems and applications*, SIAM J. Appl. Math., 19 (1970), pp. 607–628.

[M1] M. Mokhtar-Kharroubi, *Some spectral properties of the neutron transport operator in bounded geometries*, Transport Theory Statist. Phys., 16 (1987), pp. 935–958.

[M2] M. Mokhtar-Kharroubi, *Mathematical Topics in Neutron Transport. New Aspects*, Ser. Adv. Math. Appl. Sci. 46, World Scientific, River Edge, NJ, 1997.

[M3] M. Mokhtar-Kharroubi, *Homogenization of boundary value problems and spectral problems for neutron transport in locally periodic media*, Math. Models Methods Appl. Sci., to appear.

[M4] M. Mokhtar-Kharroubi, *On the Essential Spectrum of Transport Operators in L^1 Spaces*, preprint, Mathematics Laboratory of Besançon, Besançon, France, 2003.

[M5] M. Mokhtar-Kharroubi, *Optimal Spectral Theory of Neutron Transport Models*, preprint, Mathematics Laboratory of Besançon, Besançon, France, 2003.

[M6] M. Mokhtar-Kharroubi, *On L^1 Compactness in Transport Theory*, preprint, Mathematics Laboratory of Besançon, Besançon, France, 2003.

[MT] F. Murat and L. Tartar, *Calculus of variations and homogenization*, in Topics in the Mathematical Modelling of Composite Materials, Progr. Nonlinear Differential Equations Appl. 31, Birkhäuser Boston, Boston, MA, 1997, pp. 139–173.

[S] L. Schwartz, *Cours d'analyse*, Hermann, Paris, 1967.

[Vi] I. Vidav, *Existence and uniqueness of nonnegative eigenfunctions of the Boltzmann operator*, J. Math. Anal. Appl., 22 (1968), pp. 144–155.

[Vo1] J. Voigt, *Positivity in time dependent linear transport theory*, Acta. Appl. Math., 2 (1984), pp. 311–331.

[Vo2] J. Voigt, *Spectral properties of the neutron transport equation*, J. Math. Anal. Appl., 106 (1985), pp. 140–153.

[Z] E. Zeidler, *Nonlinear Functional Analysis and Its Applications.* III, Springer-Verlag, New York, 1987.

SIAM J. MATH. ANAL.
Vol. 35, No. 5, pp. 1160–1176

NONLINEAR TIME-DEPENDENT ONE-DIMENSIONAL SCHRÖDINGER EQUATION WITH DOUBLE-WELL POTENTIAL*

ANDREA SACCHETTI†

Abstract. We consider time-dependent Schrödinger equations in one dimension with double-well potential and an external nonlinear perturbation. If the initial state belongs to the eigenspace spanned by the eigenvectors associated to the two lowest eigenvalues, then, in the semiclassical limit, we show that the reduction of the time-dependent equation to a 2-mode equation gives the dominant term of the solution with a precise estimate of the error. By means of this stability result we are able to prove the absence of the beating motion for large enough nonlinearity.

Key words. nonlinear Schrödinger operator, Gross–Pitaevskii equation, norm estimate of solutions

AMS subject classifications. 35Q40, 35B, 35K55

DOI. 10.1137/S0036141002415438

1. Introduction. Recently, the theoretical analysis of the nonlinear time-dependent Schrödinger equation

$$i\hbar\dot{\psi} = H_0\psi + \epsilon|\psi|^2\psi, \quad \epsilon \in \mathrm{R}, \ \dot{\psi} = \frac{\partial\psi}{\partial t}, \tag{1}$$

where

$$H_0 = -\frac{\hbar^2}{2m}\Delta + V, \quad \Delta = \sum_{j=1}^{d}\frac{\partial^2}{\partial x_j^2}, \ d \geq 1,$$

has attracted an increasing interest (see [15] for a review and [11] for a rigorous derivation of the Gross–Pitaevskii energy functional). When V is a double-well potential, one of the main goals is to understand how the nonlinear perturbation with strength ϵ affects the unperturbed beating motion (see, e.g., the review paper [5] and the paper [19], where (1) is proposed as a model for chiral molecules). To this end, it is crucial to study the solution ψ for times of the order of the beating period; in other words, for practical purposes the unit of time is given by the beating period $T = \pi\hbar/\omega$, where $\hbar$ is the Planck's constant and ω is one-half of the energy splitting between the two lowest energies.

Here, I consider (1) in the semiclassical limit where, by assuming that $d = 1$ and under some generic assumption on the double-well potential, we give the asymptotic behavior of the solution ψ with a precise estimate of the error. In particular, the main result (Theorem 3) consists of proving that the solution of the Gross–Pitaevskii equation is approximated, with a rigorous control of the error, by means of the solution of an integrable two-dimensional dynamical system. As a result it follows (Theorem 4) that the beating motion between the two wells of a state initially made of the two lowest eigenstates disappears for increasing nonlinearity.

*Received by the editors October 2, 2002; accepted for publication (in revised form) May 9, 2003; published electronically January 6, 2004. This work was partially supported by the Italian MURST and INDAM-GNFM (project *Comportamenti Classici in Sistemi Quantistici*).
http://www.siam.org/journals/sima/35-5/41543.html
†Dipartimento di Matematica, Università di Modena e Reggio Emilia, Via Campi 213/B, I–41100 Modena, Italy (sacchetti@unimo.it).

A similar investigation was recently performed in [7], where the nonlinear perturbation is given by $\epsilon\langle\psi, g\psi\rangle g\psi$ and $g(x)$ is a given odd function, and in [14], where, in dimension $d = 1$ and $d = 3$, we consider the limit of large barrier between the two wells. In particular, in [14] I had to assume that the discrete spectrum of the Schrödinger operator H_0 consists of only two nondegenerate eigenvalues and that the restriction to the continuous eigenspace of the unitary evolution operator satisfies an a priori estimate uniformly with respect to the parameters of the model.

Finally, we mention other recent results concerning the study of the existence of stationary solutions for Gross–Pitaevskii equations with double-well potentials [2], [3] and, in the case of single-well-type potentials, the existence of solutions asymptotically given by solitary wave functions in the case when the discrete spectrum of the linear Schrödinger operator has only one nondegenerate eigenvalue [16], [21]. In the case of linear Hamiltonian H_0 with exactly two bound states Tsai and Yau [18], making use of some ideas by Soffer and Weinstein [17], proved that, in dimension $d = 3$ and under certain resonance conditions, if the initial data is near a nonlinear ground state, then the solution $\psi(t, x)$ asymptotically approaches to certain nonlinear ground state.

Our paper is organized as follows.

In section 2 we introduce the main notation and state the assumptions on the potential. Moreover, we collect some semiclassical results concerning the spectrum of the linear Schrödinger operator.

In section 3 we prove the global existence of the solution of the Gross–Pitaevskii equation, the existence of conservation laws, and an a priori estimate (Theorem 2). The global existence of the solution is proved for both repulsive and attractive nonlinear perturbation, where, in the second case, we have to assume that the strength of the nonlinear perturbation is small enough.

In section 4 we introduce the two-level approximation which, roughly speaking, consists of projecting the Gross–Pitaevskii equation onto the two-dimensional space spanned by the eigenvectors of the linear Schrödinger operator associated to the two lowest eigenvalues. For practical purposes, it is more convenient to choose, as a basis of such a two-dimensional space, the two *single-well* states. The dynamical system we obtain is exactly solvable.

In section 5 we give our main result (Theorem 3) proving the stability of the two-level approximation. Here, we make use of the comparison criterion between ordinary differential equations and an a priori estimate of the solution of the Gross–Pitaevskii equation. We emphasize that, in order to obtain such an estimate, assumption $d = 1$ on the dimension plays a crucial role.

In section 6 we give the full rigorous justification of the results by Vardi [19] proving the existence of a critical value for the nonlinearity parameter giving the destruction of the beating motion (Theorem 4).

2. Assumptions and preliminary results. Here, we consider the Cauchy problem

$$i\hbar\dot{\psi} = H_\epsilon\psi, \quad H_\epsilon = H_0 + W, \tag{2}$$
$$\psi(0, x) = \psi^0(x) \in L^2(\mathrm{R}), \quad \|\psi^0\| = 1,$$

where $\dot{\psi}$ denotes the derivative of ψ with respect to the time t, H_0 is the linear Schrödinger operator formally given by (here, x denotes the spatial variable in dimension 1)

$$H_0 = -\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + V, \tag{3}$$

V is a symmetric double-well potential, and

$$W = \epsilon|\psi|^2$$

is the nonlinear perturbation with strength ϵ.

In the following, for the sake of definiteness, we denote by C any positive constant independent of ϵ, $\hbar$, and t, we assume $\hbar$ small enough, that is, $\hbar \in (0, \hbar^\star]$ for some $\hbar^\star$, and we denote

$$\|\varphi\|_p = \|\varphi\|_{L^p} = \left\{ \int |\varphi(x)|^p dx \right\}^{1/p} \quad \text{and} \quad \|\varphi\| = \|\varphi\|_2.$$

Moreover, given $y = (y_1, \dots, y_m) \in \mathrm{R}^m$ for some $m \geq 1$, we denote

$$|y| = \max_{1 \leq j \leq m} |y_j|. \tag{4}$$

2.1. Assumptions on the potential. Here, we assume that the potential V is a regular symmetric function which admits two nondegenerate minima and it is bounded from below. More precisely, we have the following hypothesis.

HYPOTHESIS 1. *The potential $V(x)$ is a real-valued function such that*

(i) $V(-x) = V(x)$ $\forall x \in \mathrm{R}$*;*

(ii) $V \in C^2(\mathrm{R})$*;*

(iii) $V(x)$ *admits two nondegenerate minima at $x = \pm a$ for some $a > 0$ such that*

$$V(x) > V_{min} = V(\pm a) \quad \forall x \in \mathrm{R},\ x \neq \pm a; \tag{5}$$

in particular, for the sake of definiteness, we assume that

$$\frac{dV(\pm a)}{dx} = 0 \quad \textit{and} \quad \frac{d^2V(\pm a)}{dx^2} > 0;$$

(iv) *finally we assume that*

$$\liminf_{|x| \to \infty} V(x) = V_\infty > V_{min}.$$

It follows that the operator formally defined in (3) admits a self-adjoint realization (still denoted by H_0) on $L^2(\mathrm{R})$ (see, for instance, Theorem III.1.1 in [4]). Let $\sigma(H_0) = \sigma_d \cup \sigma_{ess}$ be the spectrum of the self-adjoint operator H_0, where σ_d denotes the discrete spectrum and σ_{ess} denotes the essential spectrum. From Hypothesis 1(iv) it follows that $\sigma_d \subset (V_{min}, V_\infty)$, $\sigma_{ess} = \emptyset$ if $V_\infty = +\infty$ (see Theorem XIII.67 in [13]) and that $\sigma_{ess} \subseteq [V_\infty, +\infty)$ if $V_\infty < \infty$ (see Theorem III.3.1 in [4]). Furthermore, the following two lemmas hold.

LEMMA 1. *Let σ_d be the discrete spectrum of H_0. Then, for any $\hbar \in (0, \hbar^\star]$, it follows that*

(i) *σ_d is not empty and, in particular, it contains two eigenvalues at least;*

(ii) *letting $\lambda_{1,2}$ be the lowest two eigenvalues of H_0, they are nondegenerate, in particular $\lambda_1 < \lambda_2$, and there exists $C > 0$, independent of $\hbar$, such that*

$$\inf_{\lambda \in \sigma(H_0) - \{\lambda_{1,2}\}} [\lambda - \lambda_2] \geq C\hbar.$$

Proof. The proof is an immediate consequence of the above assumptions and standard WKB arguments. □

LEMMA 2. *Let $\varphi_{1,2}$ be the normalized eigenvectors associated to $\lambda_{1,2}$. Then*

(i) *φ_j, $j = 1, 2$, can be chosen to be real-valued functions such that $\varphi_j(-x) = (-1)^{j-1}\varphi_j(x)$;*
(ii) *$\varphi_j \in H^1(\mathrm{R})$;*
(iii) *$\varphi_j \in L^p(\mathrm{R})$ for any $p \in [1, +\infty]$;*
(iv) *there exists a positive constant C such that*

$$\|\varphi_j\|_p \le C\hbar^{-\frac{p-2}{4p}} \quad \forall p \in [2, +\infty], \quad \forall \hbar \in (0, \hbar^\star]. \tag{6}$$

Proof. Property (i) immediately follows from assumption Hypothesis 1(i). Property (ii) follows from Lemma III.3.1 in [4]. Property (iii) follows from Theorem III.3.2 in [4]. Finally, property (iv) follows for $p = +\infty$ by means of standard WKB arguments. From this fact, from the normalization of the eigenvectors, and from the Hölder inequality, property (iv) follows for any $p \in [2, +\infty]$:

$$\|\varphi_j\|_p = \left[\|\varphi_j^2\varphi_j^{p-2}\|_1\right]^{1/p} \le \|\varphi_j\|_2^{2/p}\|\varphi_j\|_\infty^{(p-2)/p} = \|\varphi_j\|_\infty^{(p-2)/p}. \qquad \square$$

2.2. Splitting and single-well states. It is well known that the splitting between the two lowest eigenvalues vanishes as $\hbar$ goes to zero. In particular, we have the following lemma.

LEMMA 3. *Let*

$$\omega = \frac{\lambda_2 - \lambda_1}{2} \quad \textit{and} \quad \Omega = \frac{\lambda_2 + \lambda_1}{2}$$

and

$$\varphi_R = \frac{1}{\sqrt{2}}\left[\varphi_1 + \varphi_2\right] \quad \textit{and} \quad \varphi_L = \frac{1}{\sqrt{2}}\left[\varphi_1 - \varphi_2\right],$$

where $\varphi_{1,2}$ are the normalized eigenvectors associated to $\lambda_{1,2}$. Then there exist two positive constants C and Γ, independent of $\hbar$, such that

$$\|\varphi_R\varphi_L\|_\infty \le C\omega \tag{7}$$

and

$$\omega \le Ce^{-\Gamma/\hbar} \quad \forall \hbar \in (0, \hbar^\star]. \tag{8}$$

As a result it follows that

$$\lim_{\hbar \to 0} \omega = 0 \tag{9}$$

and

$$\lim_{\hbar \to 0} \frac{\Omega - V_{min}}{\hbar} = c \tag{10}$$

for some $c > 0$.

Proof. In order to prove this lemma we observe that V is a symmetric double-well potential with nonzero barrier between the wells. That is, let $\delta > 0$ be small enough and let us define the two sets

$$\left.\begin{aligned} B_R &= \left\{x \in \mathrm{R}^+ \ : \ V(x) \le V_{min} + \delta\right\} \\ B_L &= \left\{x \in \mathrm{R}^- \ : \ V(x) \le V_{min} + \delta\right\} \end{aligned}\right\}, \quad \text{i.e.,} \quad x \in B_R \Leftrightarrow -x \in B_L.$$

From condition (5) it follows also that

$$B_R = [b, c] \quad \text{and} \quad B_L = [-c, -b]$$

for some $c > a > b > 0$. The sets $B_{R,L}$ are usually called *wells*. Let

$$\Gamma_\delta = \int_{-b}^{b} \sqrt{\max[V(x) - (V_{min} + \delta), 0]} dx > 0$$

be the Agmon distance between the two wells. From these facts and from standard WKB arguments (see [8] and [9]) then (7)–(10) follow for some $\Gamma \in [\Gamma_0, \Gamma_\delta]$. □

Remark 1. By definition it follows that $\varphi_R(-x) = \varphi_L(x)$; moreover, from (7), it follows that these functions are localized on only one of the wells B_R and B_L; for example,

$$\int_{B_R} |\varphi_R(x)|^2 dx = 1 + O(e^{-C/\hbar})$$

for some $C > 0$. For such a reason we call them *single-well* (normalized) states.

Remark 2. We emphasize that, by assuming some regularity properties on the potential V, it is then possible to obtain the precise asymptotic behavior of the splitting as $\hbar$ goes to zero [9].

2.3. Assumptions on the parameters. We assume that the parameter ϵ is such that

$$\epsilon \to 0 \quad \text{as} \quad \hbar \to 0$$

and

$$\frac{c\epsilon}{\omega} \le C, \quad c = \|\varphi_R^2\| \quad \forall \hbar \in (0, \hbar^\star] \tag{11}$$

for some positive constant C. We recall also that the other parameter of the model, i.e., the splitting ω, satisfies the asymptotic estimate (8).

2.4. Assumption on the initial state. Let

$$\Pi_c = 1 - \langle \varphi_R, \cdot \rangle \varphi_R - \langle \varphi_L, \cdot \rangle \varphi_L \tag{12}$$

be the projection operator onto the eigenspace orthogonal to the two-dimensional eigenspace associated to the doublet $\{\lambda_{1,2}\}$. Letting ψ^0 be the initial wave function, we assume the following.

HYPOTHESIS 2. $\Pi_c \psi^0 = 0$.

3. Global existence of the solution and conservation laws. Here, we prove that the Cauchy problem (2) admits a solution for all time provided that Hypotheses 1–2 are satisfied and the strength ϵ of the nonlinear perturbation is small enough. Moreover, we prove a priori estimate of the solution ψ.

The following results hold.

THEOREM 1. *There exist $\hbar^\star > 0$ and $\epsilon_0 > 0$ such that for any $\hbar \in (0, \hbar^\star]$ and $\epsilon \in [-\epsilon_0, \epsilon_0]$ the Cauchy problem* (2) *admits a unique solution $\psi(t, x) \in H^1$ for any $t \in \mathrm{R}$. Moreover, the following conservation laws hold:*

$$\|\psi(t, \cdot)\| = \|\psi^0(\cdot)\| = 1 \tag{13}$$

and

$$E(\psi) = \frac{\hbar^2}{2m}\left\|\frac{\partial\psi}{\partial x}\right\|^2 + \langle V\psi,\psi\rangle + \frac{1}{2}\epsilon\|\psi^2\|^2 = E(\psi^0). \tag{14}$$

Proof. From Hypothesis 2 it follows that

$$\psi^0 = c_1\varphi_1 + c_2\varphi_2, \quad c_{1,2} = \langle\psi^0,\varphi_{1,2}\rangle.$$

From this fact and from Lemma 2, $\psi^0 \in H^1$. Therefore, existence of the global solution $\psi \in C(\mathrm{R}, H^1)$ and the conservation laws (13) and (14) follow from known results (see, for example, the papers quoted in [15] and [16]) for any $\epsilon > 0$ (repulsive nonlinear perturbation) and for any $\epsilon \in (-\epsilon_0, 0)$ for some $\epsilon_0 > 0$ (attractive nonlinear perturbation). □

Remark 3. There exists a positive constant C independent of $\hbar$ and ϵ such that

$$|E(\psi) - V_{min}| \le C(\omega + \hbar + \epsilon\hbar^{-1/2}) \quad \forall\hbar \in (0,\hbar^\star], \quad \forall\epsilon \in [-\epsilon_0,\epsilon_0]. \tag{15}$$

This estimate immediately follows from (14), from Hypothesis 2, and from Lemmas 1 and 2. Indeed, from Hypothesis 2 it follows that

$$E(\psi^0) = \langle H_0(c_1\varphi_1 + c_2\varphi_2), (c_1\varphi_1 + c_2\varphi_2)\rangle + \frac{1}{2}\epsilon\|\psi^0\|_4^4,$$

where $\|\psi^0\|_4 \le C\hbar^{-1/8}$ from (6) and where

$$\langle H_0(c_1\varphi_1 + c_2\varphi_2), (c_1\varphi_1 + c_2\varphi_2)\rangle = \lambda_1|c_1|^2 + \lambda_2|c_2|^2 = \Omega - \omega + 2\omega|c_2|^2.$$

From these facts and from (10), inequality (15) follows.

THEOREM 2. *Let $\epsilon_0(\hbar)$ be a function such that*

$$\lim_{\hbar\to 0} \epsilon_0(\hbar)/\hbar^2 = 0. \tag{16}$$

The solution ψ of (2) *satisfies the following uniform estimate: there exists a positive constant C independent of t, $\hbar$, and ϵ such that*

$$\|\psi\|_p \le C\left[\frac{|E(\psi^0) - V_{min}|}{\hbar^2}\right]^{\frac{p-2}{4p}} \quad \forall p \in [2,+\infty] \tag{17}$$

and

$$\left\|\frac{\partial\psi}{\partial x}\right\| \le C\left[\frac{|E(\psi^0) - V_{min}|}{\hbar^2}\right]^{\frac{1}{2}}$$

for all time and $\forall\hbar \in (0,\hbar^\star]$, $\forall\epsilon \in [-\epsilon_0(\hbar),\epsilon_0(\hbar)]$.

Proof. In order to prove the estimate (17) let

$$k = \frac{\hbar}{\sqrt{2m}}, \qquad \Lambda = \frac{E(\psi^0) - V_{min}}{k^2}.$$

Then the conservation laws (13) and (14) imply that

$$\left\|\frac{\partial\psi}{\partial x}\right\|^2 + \frac{1}{2}[\mathrm{sign}(\epsilon)]\rho^2\|\psi^2\|^2 \le \Lambda,$$

where

$$\rho = |\epsilon|^{1/2}/k \ll 1,$$

according to (16). In particular, if we set

$$\chi = \rho\psi,$$

then the above equation takes the form

$$\left\| \frac{\partial \chi}{\partial x} \right\|^2 + \frac{1}{2}[\mathrm{sign}(\epsilon)]\|\chi^2\|^2 \leq \Lambda\rho^2,$$

from which it follows that

$$\left\| \frac{\partial \chi}{\partial x} \right\|^2 \leq \rho^2|\Lambda| + \frac{1}{2}\|\chi^2\|^2 = \rho^2|\Lambda| + \frac{1}{2}\|\chi\|_4^4. \tag{18}$$

From the Gagliardo–Nirenberg inequality (see, for instance, [6] and [20], where the dimension is here equal to 1)

$$\|\chi\|_{2\sigma+2}^{2\sigma+2} \leq C \left\| \frac{\partial \chi}{\partial x} \right\|^{\sigma} \|\chi\|^{2+\sigma} \quad \forall \sigma \geq 0, \tag{19}$$

where we choose $\sigma = 1$, it follows that

$$\|\chi\|_4^4 \leq C \left\| \frac{\partial \chi}{\partial x} \right\| \|\chi\|^3 \leq C \left\| \frac{\partial \chi}{\partial x} \right\| \rho^3$$

since $\|\chi\| = \rho\|\psi\| = \rho$ and $\|\psi\| = 1$. By inserting this inequality in (18) it follows that $\|\frac{\partial \chi}{\partial x}\|$ satisfies

$$\left\| \frac{\partial \chi}{\partial x} \right\|^2 \leq \rho^2|\Lambda| + C\rho^3 \left\| \frac{\partial \chi}{\partial x} \right\| \tag{20}$$

for any $t \in \mathrm{R}$. From (20) it immediately follows that

$$\left\| \frac{\partial \chi}{\partial x} \right\| \leq \sqrt{|\Lambda|}\rho\,(1 + o(1)) \quad \text{as} \quad \rho \to 0.$$

Hence, $\|\frac{\partial \psi}{\partial x}\| \leq C\sqrt{|\Lambda|}$ and, from (19), we have that

$$\|\psi\|_p \leq C \left\| \frac{\partial \psi}{\partial x} \right\|^{\sigma/p} \leq C|\Lambda|^{(p-2)/4p},$$

where we choose now $\sigma = \frac{p-2}{2}$, i.e., $p = 2\sigma + 2$. □

Remark 4. Condition (16) is true in the semiclassical limit and under assumption (11).

Remark 5. From the fact $E(\psi_0) - V_{min} = O(\hbar)$, which follows from (8), (15), and (16), and from the bounds (17) and (11), it then follows that

$$\|\psi\|_p \leq C\hbar^{-\frac{p-2}{4p}} \quad \forall p \in [2, +\infty] \quad \text{and} \quad \left\| \frac{\partial \psi}{\partial x} \right\| \leq C\hbar^{-\frac{1}{2}} \tag{21}$$

for any $t \in \mathrm{R}$, $\hbar \in (0, \hbar^\star]$, and $\epsilon \in [-\epsilon_0(\hbar), \epsilon_0(\hbar)]$.

4. Two-level approximation. For our purposes it is more convenient to make the substitution $\psi \to e^{-i\Omega t/\hbar}\psi$; hence (2) takes the following form (where, with abuse of notation, we still denote the new function by ψ):

$$i\hbar\dot{\psi} = (H_0 - \Omega)\psi + \epsilon|\psi|^2\psi, \quad \psi(x,0) = \psi^0(x). \tag{22}$$

Let us write the solution of this equation in the form

$$\psi(t,x) = a_R(t)\varphi_R(x) + a_L(t)\varphi_L(x) + \psi_c(t,x), \tag{23}$$

where $a_R(t)$ and $a_L(t)$ are unknown complex-valued functions depending on time and $\psi_c = \Pi_c\psi$, Π_c, defined in (12), is the projection onto the space orthogonal to the two-dimensional space spanned by the two single-well states φ_R and φ_L; i.e.,

$$\langle\psi_c, \varphi_R\rangle = \langle\psi_c, \varphi_L\rangle = 0 \quad \forall t \in \mathrm{R}.$$

From the conservation law (13) it follows that

$$|a_R(t)|^2 + |a_L(t)|^2 + \|\psi_c(t,\cdot)\|^2 = 1 \quad \forall t \in \mathrm{R}. \tag{24}$$

By substituting ψ by (23) in (2) we obtain that a_R, a_L, and ψ_c must satisfy the system of differential equations

$$\begin{cases} i\hbar\dot{a}_R = -\omega a_L + \epsilon\langle\varphi_R, |\psi|^2\psi\rangle, \\ i\hbar\dot{a}_L = -\omega a_R + \epsilon\langle\varphi_L, |\psi|^2\psi\rangle, \\ i\hbar\dot{\psi}_c = (H_0 - \Omega)\psi_c + \epsilon\Pi_c|\psi|^2\psi. \end{cases} \tag{25}$$

By again substituting ψ by (23) in the first two equations of the above system, we obtain that these equations take the form

$$\begin{cases} i\hbar\dot{a}_R = -\omega a_L + \epsilon c|a_R|^2 a_R + \epsilon r_R, \\ i\hbar\dot{a}_L = -\omega a_R + \epsilon c|a_L|^2 a_L + \epsilon r_L, \end{cases} \tag{26}$$

where

$$c = \|\varphi_R^2\|^2 = \|\varphi_L^2\|^2 = O(\hbar^{-1}) \tag{27}$$

and where r_R and r_L are given by

$$\begin{aligned} r_R &= \langle\varphi_R, |\psi|^2\psi\rangle - |a_R^2|a_R\langle\varphi_R, |\varphi_R|^2\varphi_R\rangle \\ &= \langle\varphi_R, |\psi|^2\phi_L\rangle + a_R\langle|\varphi_R|^2, |\phi_L|^2 + a_R\varphi_R\bar{\phi}_L + \bar{a}_R\bar{\varphi}_R\phi_L\rangle, \\ r_L &= \langle\varphi_L, |\psi|^2\psi\rangle - |a_L^2|a_L\langle\varphi_L, |\varphi_L|^2\varphi_L\rangle \\ &= \langle\varphi_L, |\psi|^2\phi_R\rangle + a_L\langle|\varphi_L|^2, |\phi_R|^2 + a_L\varphi_L\bar{\phi}_R + \bar{a}_L\bar{\varphi}_L\phi_R\rangle, \end{aligned}$$

where

$$\phi_L = a_L\varphi_L + \psi_c \quad \text{and} \quad \phi_R = a_R\varphi_R + \psi_c.$$

We denote by *two-level approximation* the solutions b_R and b_L of the system of ordinary differential equations

$$\begin{cases} i\hbar\dot{b}_R = -\omega b_L + \epsilon c|b_R|^2 b_R, \\ i\hbar\dot{b}_L = -\omega b_R + \epsilon c|b_L|^2 b_L, \end{cases} \qquad b_{R,L}(0) = a_{R,L}(0), \tag{28}$$

obtained by neglecting the remainder terms r_R and r_L in (26). It is easy to see that the solution of this system satisfies the conservation law

$$|b_R(t)|^2 + |b_L(t)|^2 = |b_R(0)|^2 + |b_L(0)|^2 = |a_R(0)|^2 + |a_L(0)|^2 = 1, \tag{29}$$

and, moreover, it is also possible to explicitly compute (see [12] and Appendix B in [14]) the solution of (28) by means of elliptic functions cn and dn [1]. In particular, we obtain that the imbalance function, defined as

$$z(t) = |b_R(t)|^2 - |b_L(t)|^2, \tag{30}$$

is given by

$$z(t) = \begin{cases} A\mathrm{cn}\left[A\eta(\omega t/\hbar - \tau_0)/2k, k\right] & \text{if } k < 1, \\ A\mathrm{dn}\left[A\eta(\omega t/\hbar - \tau_0)/2, 1/k\right] & \text{if } k > 1, \end{cases}$$

where $\eta = \epsilon c/\omega$, τ_0 depends on the initial condition,

$$I = \sqrt{1 - z^2(0)}\cos[\theta(0)] - \eta z^2(0)/4,$$

$\theta = \arg(b_R) - \arg(b_L)$ is the relative phase,

$$A = \frac{2\sqrt{2}}{\eta}\left[\sqrt{\frac{1}{4}\eta^2 + 1 + I\eta} - \left(1 + \frac{1}{2}I\eta\right)\right]^{1/2},$$

and

$$k^2 = \frac{1}{2}\left[1 - \frac{1 + \frac{1}{2}I\eta}{\sqrt{\frac{1}{4}\eta^2 + 1 + I\eta}}\right]. \tag{31}$$

We emphasize that $z(t)$ periodically assumes positive and negative values if and only if $k < 1$.

5. Stability of the two-level approximation. Our main result consists of proving the stability of the two-level approximation when we restore the remainder terms r_R and r_L in (28).

We prove the following.

THEOREM 3. *Let $\psi_c = \Pi_c\psi$, $a_R(t) = \langle\psi, \varphi_R\rangle$, and $a_L(t) = \langle\psi, \varphi_L\rangle$, where ψ is the solution of* (22)*, and let $b_R(t)$ and $b_L(t)$ be the solution of the system of ordinary differential equations* (28)*. Let $\epsilon \in [-\epsilon_0(\hbar), \epsilon_0(\hbar)]$, where $\epsilon_0(\hbar)$ satisfies the condition* (16)*. Then, for any $\tau' > 0$, there exists a positive constant C independent of ϵ, $\hbar$, and t such that*

$$|b_{R,L}(t) - a_{R,L}(t)| \le Ce^{-C\hbar^{-1}} \quad \text{and} \quad \|\psi_c(\cdot, t)\| \le Ce^{-C\hbar^{-1}} \tag{32}$$

for any $\hbar \in (0, \hbar^\star]$ and for any $t \in [0, \hbar\tau'/\omega]$.

Proof. For the sake of simplicity, hereafter, we omit the parameters when doing so does not cause misunderstandings. In order to prove the theorem we introduce the *slow time* $\tau = \omega t/\hbar$ and let

$$\begin{cases} A_{R,L}(\tau) = a_{R,L}(t), \\ B_{R,L}(\tau) = b_{R,L}(t), \end{cases} \qquad R_{R,L}(\tau) = \frac{\epsilon}{\omega}r_{R,L}(t), \quad \text{and} \quad \eta = \frac{\epsilon c}{\omega}.$$

Then (26) and (28), respectively, take the form (here $'$ denotes the derivative with respect to τ)

$$\begin{cases} A_R' = iA_L - i\eta|A_R|^2A_R + R_R, \\ A_L' = iA_R - i\eta|A_L|^2A_L + R_L \end{cases} \tag{33}$$

and

$$\begin{cases} B_R' = iB_L - i\eta|B_R|^2B_R, \\ B_L' = iB_R - i\eta|B_L|^2B_L, \end{cases} \tag{34}$$

satisfying the same initial condition

$$B_{R,L}(0) = A_{R,L}(0) = a_{R,L}(0).$$

Due to (24) and (29), they are such that

$$|B_R(\tau)|^2 + |B_L(\tau)|^2 = 1, \quad |A_R(\tau)|^2 + |A_L(\tau)|^2 \le 1. \tag{35}$$

In a more concise way, with an obvious meaning of notation, we can write (33) and (34) as

$$A' = f(A) + R \quad \text{and} \quad B' = f(B), \quad A(0) = B(0) = a(0), \tag{36}$$

where $A, B \in S^2$ since (35), $S^2 = \{(z_1, z_2) \in \mathrm{C}^2 : |z_1|^2 + |z_2|^2 \le 1\}$.

LEMMA 4. *The function $f : S^2 \to C^2$ satisfies the Lipschitz condition*

$$|f(A) - f(B)| \le L|A - B|, \ L = 1 + 3\eta. \tag{37}$$

Proof. According to the notation (4) we have

$$|f(A) - f(B)| = \max\left[|f_R|, |f_L|\right],$$

where $|A| \le 1$ and $|B| \le 1$ since $A, B \in S^2$, and where

$$f_R = (A_L - B_L) - \eta(|A_R|^2A_R - |B_R|^2B_R),$$
$$f_L = (A_R - B_R) - \eta(|A_L|^2A_L - |B_L|^2B_L).$$

Then (37) immediately follows since

$$f_R = (A_L - B_L) - \eta\left[|B_R|^2(A_R - B_R) + A_R(|A_R| + |B_R|)(|A_R| - |B_R|)\right],$$

where $||A_R| - |B_R|| \le |A_R - B_R|$, and where the other term f_L will be treated the same way. ∎

LEMMA 5. *Let*

$$\beta = \max[c\epsilon, \omega],$$

where c is defined in (27). *Let $\psi_c = \Pi_c\psi$, where ψ is the solution of* (22); *it satisfies the uniform estimate*

$$\|\psi_c\| \le C\beta\hbar^{-3/2}\left[\exp[C\hbar^{-1/2}(\epsilon t/\hbar)] + 1\right] \quad \forall t \in \mathrm{R} \tag{38}$$

for some positive constant C independent of $\hbar$, ϵ, and t.

Proof. As a first step we consider the following raw estimates:

$$\|\psi_c\|_p \leq C\hbar^{-\frac{p-2}{4p}} \quad \forall p \in [2,+\infty], \quad \forall t \in \mathrm{R} \tag{39}$$

and

$$|r_{R,L}| \leq C\hbar^{-1/2} \quad \forall t \in \mathrm{R}.$$

Indeed, (39) immediately follows from the Minkowski inequality and from (21):

$$C\hbar^{-\frac{p-2}{4p}} \geq \|\psi\|_p \geq -\left(|a_R(t)|\|\varphi_R\|_p + |a_L(t)|\|\varphi_L\|_p\right) + \|\psi_c\|_p,$$

where $|a_{R,L}(t)| \leq 1$, and where $\varphi_{R,L}$ satisfy the bound (6). In the same way, from Lemma 2 and Theorem 2, it follows that

$$\begin{aligned} |r_R| &\leq C\|\varphi_R\psi^2\| \cdot \|\psi\| + \|\varphi_R\|_4^4 \\ &\leq C\|\varphi_R\|_\infty \|\psi\|_4^2\|\psi\| + C\|\varphi_R\|_4^4 \\ &\leq C\hbar^{-1/2}, \end{aligned}$$

and similarly for $|r_L|$.

Now, in order to prove the estimate (38) we make use of the third equation of (25), from which it follows that

$$\psi_c(\cdot,t) = -i\frac{\epsilon}{\hbar}\int_0^t e^{-i(H_0-\Omega)(t-s)/\hbar}\Pi_c|\psi(\cdot,s)|^2\psi(\cdot,s)ds$$

since $\psi_c^0 = \Pi_c\psi^0 = 0$ from Hypothesis 2.

Let $\psi = \varphi + \psi_c$, where $\varphi = a_R\varphi_R + a_L\varphi_L$. Then

$$|\psi|^2\psi = \varphi_I + \psi_c\varphi_{II} + \bar{\psi}_c\varphi_{III}, \quad \begin{cases} \varphi_I = |\varphi|^2\varphi, \\ \varphi_{II} = 2|\varphi|^2 + 2\bar{\psi}_c\varphi + |\psi_c|^2 + \bar{\varphi}\psi_c, \\ \varphi_{III} = \varphi^2. \end{cases}$$

Therefore, we can write

$$\psi_c = -i\frac{\epsilon}{\hbar}\left[I + II + III\right],$$

where

$$I = \int_0^t e^{-i(H_0-\Omega)(t-s)/\hbar}\Pi_c\varphi_I ds,$$

$$II = \int_0^t e^{-i(H_0-\Omega)(t-s)/\hbar}\Pi_c\psi_c\varphi_{II} ds,$$

$$III = \int_0^t e^{-i(H_0-\Omega)(t-s)/\hbar}\Pi_c\bar{\psi}_c\varphi_{III} ds.$$

For the first term we have, by integrating by parts, that

$$\begin{aligned} I = &\left[-i\hbar e^{-i(H_0-\Omega)(t-s)/\hbar}[H_0-\Omega]^{-1}\Pi_c|\varphi|^2\varphi\right]_0^t \\ &+ i\hbar\int_0^t e^{-i(H_0-\Omega)(t-s)/\hbar}[H_0-\Omega]^{-1}\Pi_c\frac{\partial|\varphi|^2\varphi}{\partial s}ds. \end{aligned}$$

Let us emphasize that from Lemma 1 it follows that the following operators, from L^2 into L^2, are bounded:

$$\left\| e^{-i(H_0-\Omega)(t-s)/\hbar} \right\| = 1, \quad \|\hbar[H_0-\Omega]^{-1}\Pi_c\| \le C.$$

Also, from Lemma 2 and (24), (26), and (27), we have the following uniform estimate for any $t \in \mathrm{R}$:

$$\begin{aligned}\|\dot\varphi\|_p &\le (|\dot a_r| + |\dot a_L|)\left(\|\varphi_R\|_p + \|\varphi_L\|_p\right) \le C\hbar^{-1}\max[c\epsilon, \omega, \epsilon\hbar^{-\frac{1}{2}}]\hbar^{-\frac{p-2}{4p}} \\ &\le C\hbar^{-1}\beta\hbar^{-\frac{p-2}{4p}}.\end{aligned}$$

Then we have that

$$\begin{aligned}\|I\| &\le C \max_{s\in[0,t]}\left\{\|\varphi^3(s,\cdot)\| + t\|\dot\varphi(s,\cdot)\varphi^2(s,\cdot)\|\right\} \\ &\le C \max_{s\in[0,t]}\left\{\|\varphi(s,\cdot)\|_6^3 + t\|\dot\varphi(s,\cdot)\| \cdot \|\varphi(s,\cdot)\|_\infty^2\right\} \\ &\le C\left\{\hbar^{-1/2} + t\hbar^{-1}\beta\hbar^{-1/2}\right\}.\end{aligned}$$

For the other two terms we have that

$$\|II\| \le \int_0^t \|\psi_c\| \cdot \|\varphi_{II}\|_\infty ds \le C\hbar^{-1/2}\int_0^t \|\psi_c\| ds$$

since $\|\varphi_{II}\|_\infty \le C\hbar^{-1/2}$, and similarly

$$\|III\| \le \int_0^t \|\psi_c\| \cdot \|\varphi_{III}\|_\infty ds \le C\hbar^{-1/2}\int_0^t \|\psi_c\| ds.$$

Indeed, from Lemma 2 and (39) it follows that

$$\|\varphi_{II}\|_\infty \le C\left\{\|\varphi\|_\infty^2 + \|\psi_c\|_\infty\|\varphi\|_\infty + \|\psi_c\|_\infty^2\right\} \le C\hbar^{-1/2}$$

and

$$\|\varphi_{III}\|_\infty \le \|\varphi\|_\infty^2 \le C\hbar^{-1/2}.$$

Collecting all these results and denoting

$$g(t) = \|\psi_c(\cdot,t)\|$$

we have that $g(t)$ is a positive real-valued function satisfying the estimate

$$\begin{aligned}g(t) &\le C\frac{\epsilon}{\hbar}\left\{\hbar^{-1/2}\int_0^t g(s)ds + \hbar^{-1/2}\left(1 + t\hbar^{-1}\beta\right)\right\} \\ &\le a\int_0^t g(s)ds + a + abt, \quad a = C\frac{\epsilon}{\hbar^{3/2}}, \quad b = \frac{\beta}{\hbar}.\end{aligned}$$

From this estimate, since $\psi_c(0) = 0$, and from Gronwall's lemma (see [10], page 19) it follows that

$$\begin{aligned}g(t) &\le a + abt + a\int_0^t e^{a(t-s)}(a + abs)ds = -b + ae^{at} + be^{at} \\ &\le \frac{C\beta}{\hbar^{3/2}}\left[e^{C\epsilon t\hbar^{-3/2}} + 1\right],\end{aligned}$$

proving the result. □

From the inequality (8) and from assumption (11) it follows that for any fixed $\tau' > 0$ there exists $C > 0$ satisfying the second inequality in (32).

LEMMA 6. *For any fixed $\tau' > 0$ the remainder terms r_R and r_L satisfy the uniform estimate*

$$\max\left[|r_R|, |r_L|\right] \leq C\beta\hbar^{-2}e^{C\hbar^{-1/2}} \quad \forall t \in [0, \tau'\hbar/\omega]$$

for some positive constant C independent of $\hbar$, ϵ, and t.

Proof. Let us consider only the term $|r_R|$; the other term $|r_L|$ could be treated the same way. By definition, and since $\max[|a_R|, |a_L|] \leq 1$, it follows that

$$|r_R| \leq + \left|\langle \varphi_R\bar{\varphi}_L, |\psi|^2\rangle\right| \tag{40}$$

$$+ \left|\langle \varphi_R|\psi|^2, \psi_c\rangle\right| \tag{41}$$

$$+ \left|\langle |\varphi_R|^2, |\phi_L|^2 + a_R\varphi_R\bar{\phi}_L + \bar{a}_R\bar{\varphi}_R\phi_L\rangle\right|, \tag{42}$$

and we estimate separately each term.

From Lemma 3, equation (13), and the Hölder inequality, it follows that the term (40) satisfies the estimate

$$\left|\langle \varphi_R\varphi_L, |\psi|^2\rangle\right| \leq \|\varphi_R\bar{\varphi}_L\|_\infty \cdot \|\psi^2\|_1 \leq C\omega.$$

From Lemma 5 and the Hölder inequality, it follows that the term (41) satisfies the estimates

$$\left|\langle \varphi_R|\psi|^2, \psi_c\rangle\right| \leq \|\varphi_R\|_\infty \cdot \|\psi^2\| \cdot \|\psi_c\| \leq C\beta\hbar^{-2}e^{C\hbar^{-1/2}}$$

and that the term (42) satisfies the estimate

$$\begin{aligned}&\left|\langle |\varphi_R|^2, |\phi_L|^2 + a_R\varphi_R\bar{\phi}_L + \bar{a}_R\bar{\varphi}_R\phi_L\rangle\right| \\ &\quad \leq C\left[\|\varphi_R\varphi_L\|_\infty + \|\varphi_R^2\|_\infty\|\psi_c\|^2 + \|\varphi_R\varphi_L\|_\infty\|\psi_c\|\right] \leq C\omega.\end{aligned}$$

Collecting all these estimates, we obtain the proof of the lemma. □

The proof of the theorem is almost complete. Indeed, equations (36) can be rewritten in the integral form

$$A(\tau) = A(0) + \int_0^\tau f[A(s)]ds + \int_0^\tau Rds$$

and

$$B(\tau) = B(0) + \int_0^\tau f[B(s)]ds,$$

from which, and from Lemmas 4 and 5, it follows that for any $\tau \in [0, \tau']$,

$$\begin{aligned}|A(\tau) - B(\tau)| &\leq \int_0^\tau |f[A(s)] - f[B(s)]|\, ds + \int_0^\tau |R|ds \\ &\leq a\int_0^\tau |A(s) - B(s)|\, ds + b\tau, \quad a = L, \quad b = C\frac{\epsilon\beta\hbar^{-2}e^{C\hbar^{-1/2}}}{\omega}.\end{aligned}$$

From this inequality and by means of Gronwall's lemma we finally obtain that

$$|A(\tau) - B(\tau)| \le b\tau + ab\int_0^\tau e^{a(\tau-s)}sds = \frac{b}{a}\left[e^{a\tau} - 1\right]$$
$$\le \frac{C}{L}\frac{\epsilon\beta\hbar^{-2}e^{C\hbar^{-1/2}}}{\omega},$$

proving (32) since

$$\frac{\omega+\epsilon}{C'\omega} \le L = 1 + 3\eta \le C'\frac{\omega+\epsilon}{\omega}$$

for some $C' > 0$, which implies that $\frac{\beta}{L\omega} \le C$ for some $C > 0$. □

Remark 6. Since $\omega = O(e^{-\Gamma/\hbar})$ the above theorem implies that for any $\alpha < 1$ and for any $\tau' > 0$, there exists C such that

$$|b_{R,L}(t) - a_{R,L}(t)| \le C\omega^\alpha \quad \text{and} \quad \|\psi_c(\cdot,t)\| \le C\omega^\alpha \quad \forall t \in [0, \hbar\tau'/\omega].$$

6. Destruction of the beating motion for large nonlinearity.

6.1. The unperturbed case $\epsilon = 0$. Under Hypothesis 2 it follows that the solution of the unperturbed equation

$$i\hbar\dot{\psi} = H_0\psi, \quad \psi(0,x) = \psi^0(x)$$

is simply given by

$$\psi(t,x) = e^{-i\Omega t/\hbar}\left[\frac{c_1+c_2}{\sqrt{2}}\cos(\omega t/\hbar) + i\frac{c_2-c_1}{\sqrt{2}}\sin(\omega t/\hbar)\right]\varphi_R(x)$$
$$+ e^{-i\Omega t/\hbar}\left[\frac{c_1-c_2}{\sqrt{2}}\cos(\omega t/\hbar) - i\frac{c_1+c_2}{\sqrt{2}}\sin(\omega t/\hbar),\right]\varphi_L(x),$$

where

$$c_{1,2} = \langle\varphi_{1,2}, \psi^0\rangle, \quad |c_1|^2 + |c_2|^2 = 1.$$

Hence, $\psi(t,x)$ is, up to the phase factor $e^{-i(\Omega-\omega)t/\hbar}$, a periodic function with period $T = \pi\hbar/\omega$.

In particular, if ψ initially coincides with a single-well state, e.g., $\psi^0 = \varphi_R$, then

$$\psi(t,x) = e^{-i(\Omega-\omega)t/\hbar}\left[e^{-i\omega t/\hbar}\cos(\omega t/\hbar)\varphi_R(x) - ie^{-i\omega t/\hbar}\sin(\omega t/\hbar)\varphi_L(x)\right]$$

and the state $\psi(t,x)$ performs a beating motion. That is, the state, initially localized on the well B_R, is localized on the other well B_L after half a period and, after a whole

period, it returns to the initial well, and so on. In particular, let us consider the motion of the *center of mass* defined here as

$$\langle X\rangle^t = \langle X\psi, \psi\rangle = \int_{\mathrm{R}} X(x)|\psi(t,x)|^2 dx,$$

where $X \in C(R) \cap L^2(\mathrm{R})$ is a given bounded function such that $X(-x) = -X(x)$. We have that

$$\langle X\rangle^t = X_0 \left[\cos^2(\omega t/\hbar) - \sin^2(\omega t/\hbar)\right],$$

where

$$X_0 = \langle \varphi_R, X\varphi_R\rangle = \int_{\mathrm{R}} X(x)|\varphi_R(x)|^2 dx.$$

Hence, $\langle X\rangle^t$ is a periodic function which periodically assumes positive and negative values; i.e., we have the well-known beating motion for the double-well problem.

6.2. The perturbed case $\epsilon \neq 0$. In such a case it follows that the *center of mass* is given by

$$\langle X\rangle^t = X_0[|a_R(t)|^2 - |a_L(t)|^2] + r,$$

where X_0 has been previously defined and the remainder term r satisfies the uniform estimate

$$\begin{aligned} |r| &= 2\,|\Re\,[a_R\bar{a}_L\langle X\varphi_R, \varphi_L\rangle + \langle X\psi, \psi_c\rangle]| \\ &\leq 2\,[\|\varphi_R\varphi_L\|_\infty + \|X\|_\infty\|\psi\|\|\psi_c\|] \\ &\leq Ce^{-C\hbar^{-1}} \quad \forall t \in [0, \hbar\tau'/\omega]. \end{aligned}$$

If we denote by $z(t)$ the imbalance function defined in (30), then, in the semiclassical limit, it follows that

$$|a_R(t)|^2 - |a_L(t)|^2 \sim z(t) \quad \forall t \in [0, \hbar\tau'/\omega];$$

hence

$$\langle X\rangle^t \sim X_0 z(t) \quad \forall t \in [0, \hbar\tau'/\omega].$$

Then we have the following.

THEOREM 4. *Let Hypotheses* 1 *and* 2 *be satisfied. Let* k^2 *be defined as in* (31)*, depending on the initial wave function* ψ^0*. Let* $\tau' > 0$ *be fixed, and let* $\langle X\rangle^t$*, up to a remainder term, be a periodic function for any* $t \in [0, \hbar\tau'/\omega]$*. In particular, if*

(i) $k^2 < 1$*, then* $\langle X\rangle^t$ *periodically assumes positive and negative values (i.e., the beating motion still persists);*

(ii) $k^2 > 1$*, then* $\langle X\rangle^t$ *has a definite sign (i.e., the beating motion is forbidden).*

Remark 7. Let us close by emphasizing that when the wave function is initially prepared on just one well, e.g., $\psi^0 = \varphi_R$, then

$$I = -\frac{1}{4}\eta \quad \text{and} \quad k^2 = \frac{1}{16}\eta^2.$$

Therefore, from the theorem above it follows that for $|\eta|$ larger than the critical value 4 the beating motion is forbidden (see Figure 1). In such a way, we put on a fully rigorous basis the results obtained by [19] in the two-level approximation.

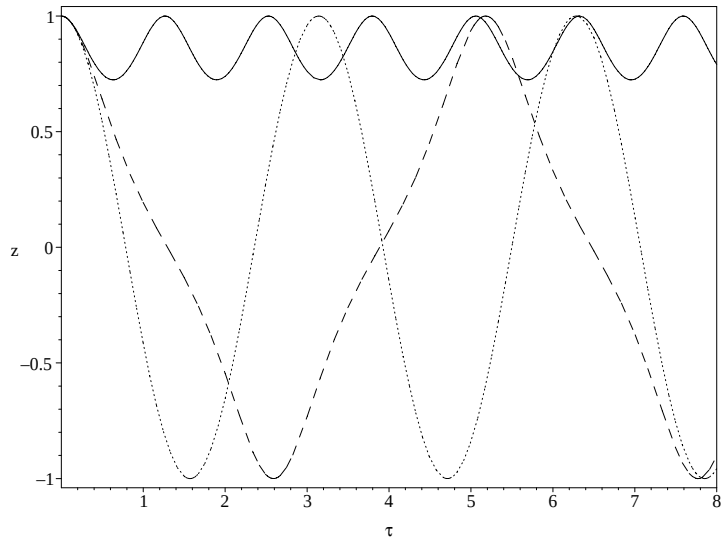


FIG. 1. *Absence of the beating motion of the* center of mass *for nonlinearity larger than a critical value. Here, we plot the* imbalance function $z(\tau)$ *for different values of the nonlinearity parameter* η*, where* $\tau = \omega t/\hbar$ *denotes the* slow time. *For* $\eta = 0$ *(point line) and* $\eta = 3.8$ *(broken line) we still have a beating motion; in contrast, for* η *larger than the critical value* 4*, e.g.,* $\eta = 6.5$ *(full line), the beating motion is forbidden.*

Acknowledgment. I thank Prof. Vincenzo Grecchi and Prof. André Martinez for helpful discussions and remarks.

REFERENCES

[1] M. ABRAMOWITZ AND I.A. STEGUN, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Wiley, New York, 1972.

[2] W.H. ASCHBACHER, J. FROEHLICH, G.M. GRAF, K. SCHNEE, AND M. TROYER, *Symmetry breaking regime in the nonlinear Hartree equation*, J. Math. Phys., 43 (2002), pp. 3879–3891.

[3] R. D'AGOSTA, B.A. MALOMED, AND C. PRESILLA, *Stationary solutions of the Gross-Pitaevskii equation with linear counterpart*, Phys. Lett. A, 275 (2002), pp. 424–434.

[4] F.A. BEREZIN AND M.A. SHUBIN, *The Schrödinger Equation*, Kluwer Academic, Dordrecht, The Netherlands, 1991.

[5] F. DALFOVO, S. GIORGINI, L.P. PITAVSKII, AND S. STRINGARI, *Theory of Bose-Einstein condensation in trapped gases*, Rev. Modern Phys., 71 (1999), pp. 463–512.

[6] G. FIBICH, B. ILAN, AND G. PAPANICOLAOU, *Self-focusing with fourth-order dispersion*, SIAM J. Appl. Math., 62 (2002), pp. 1437–1462.

[7] V. GRECCHI, A. MARTINEZ, AND A. SACCHETTI, *Destruction of the beating effect for a nonlinear Schrödinger equation*, Comm. Math. Phys., 227 (2002), pp. 191–209.

[8] E.M. HARREL, *Double wells*, Comm. Math. Phys., 75 (1980), pp. 239–261.

[9] B. HELFFER AND J. SJÖSTRAND, *Multiple wells in the semi-classical limit* I, Comm. Partial Differential Equations, 9 (1984), pp. 337–408.

[10] E. HILLE, *Lectures on Ordinary Differential Equations*, Addison-Wesley, London, 1969.

[11] E.H. LIEB, R. SEIRINGER, AND J. YNGVASON, *A rigorous derivation of the Gross-Pitaevskii energy functional for a two-dimensional Bose gas*, Comm. Math. Phys., 224 (2001), pp. 17–31.

[12] S. RAGHAVAN, A. SMERZI, S. FANTONI, AND S.R. SHENOY, *Coherent oscillations between two weakly coupled Bose-Einstein condensates: Josephson effects, π oscillations, and macroscopic quantum self-trapping*, Phys. Rev. A, 59 (1999), pp. 620–633.

[13] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics:* IV. *Analysis of Operators*, Academic Press, New York, 1972.

[14] A. Sacchetti, *Nonlinear Time-Dependent Schrödinger Equations: The Gross-Pitaevskii Equation with Double-Well Potential*, mp-arc 02-208, preprint, 2002.
[15] C. Sulem and P.L. Sulem, *The Nonlinear Schrödinger Equation: Self-focusing and Wave Collapse*, Springer-Verlag, New York, 1999.
[16] A. Soffer and M.I. Weinstein, *Multichannel nonlinear scattering for nonintegrable equations*, Comm. Math. Phys., 133 (1990), pp. 119–146.
[17] A. Soffer and M.I. Weinstein, *Resonances, radiation damping and instability in Hamiltonian nonlinear wave equations*, Invent. Math., 136 (1999), pp. 9–74.
[18] T.P. Tsai and H.T. Yau, *Asymptotic dynamics of nonlinear Schrödinger equations: Resonance-dominated and dispersion-dominated solutions*, Comm. Pure Appl. Math., 55 (2002), pp. 153–216.
[19] A. Vardi, *On the role of intermolecular interactions in establishing chiral stability*, J. Chem. Phys., 112 (2000), pp. 8743–8746.
[20] M.I. Weinstein, *Nonlinear Schrödinger equations and sharp interpolation estimates*, Comm. Math. Phys., 87 (1983), pp. 567–576.
[21] R. Weder, *Center manifold for nonintegrable nonlinear Schrödinger equations on the line*, Comm. Math. Phys., 215 (2000), pp. 343–356.

SIAM J. MATH. ANAL.
Vol. 35, No. 5, pp. 1177–1202

A QUASI-DUAL FUNCTION METHOD FOR EXTRACTING EDGE STRESS INTENSITY FUNCTIONS*

MARTIN COSTABEL†, MONIQUE DAUGE†, AND ZOHAR YOSIBASH‡

Abstract. We present a method for the computation of the coefficients of singularities along the edges of a polyhedron for second-order elliptic boundary value problems. The class of problems considered includes problems of stress concentration along edges or crack fronts in general linear three-dimensional elasticity. Our method uses an incomplete construction of three-dimensional dual singular functions, based on explicitly known dual singular functions of two-dimensional problems tensorized by test functions along the edge and combined with complementary terms improving their orthogonality properties with respect to the edge singularities. Our method is aimed at the numerical computation of the stress intensity functions. It is suitable for a postprocessing procedure in the finite element approximation of the solution of the boundary value problem.

Key words. edge singularities, dual singularities, stress intensity factors

AMS subject classifications. 35J25, 35B65

DOI. 10.1137/S0036141002404863

1. Introduction.

1.1. The problem. The solutions of elliptic boundary problems, for example, those arising from linear elasticity, when posed and solved in nonsmooth domains like polygons and polyhedra, have nonsmooth parts. It is well known how to describe these singularities in terms of special *singular functions* depending on the geometry and the differential operators on one hand, and of unknown *coefficients* depending on the given right-hand sides (for example, volume forces and surface tractions or displacements) on the other hand.

Concerning the *singular functions*, they are extensively covered in the literature. In many cases, like corners in two dimensions or edges in three dimensions, they can be written analytically (see, for example, [18, 3, 29]) or semianalytically [12]. In other cases, like polyhedral corners, there exist well-known numerical methods for their computation (see, for example, [1, 35, 33, 36]).

Concerning the *coefficients*, there are two cases to distinguish: *corners* and *edges*.

1. In the case of a *corner* in two or three dimensions, i.e., the vertex of a cone, the space of singular functions up to a given regularity is finite-dimensional. Therefore only finitely many numbers have to be computed, and there exist several well-established methods to do this. Let us mention some of them:

In the "singular function method," also known as the Fix method in the finite element literature, singular basis functions are added to the space of trial functions, so that their coefficients are computed immediately as a part of the numerical solution of the boundary value problem (see [4, 6, 8, 17, 28, 32]).

In the "dual singular function method," one uses the fact that the coefficients depend linearly on the solution and therefore also on the right-hand side; see [21, 23]

*Received by the editors March 28, 2002; accepted for publication (in revised form) June 6, 2003; published electronically January 6, 2004.
http://www.siam.org/journals/sima/35-5/40486.html
†UMR CNRS 6625–IRMAR, Université de Rennes I, Campus de Beaulieu, 35042 Rennex Cedex, France (costabel@univ-rennes1.fr, dauge@univ-rennes1.fr).
‡Pearlstone Center for Aeronautical Engineering Studies, Mechanical Engineering Department, Ben Gurion University of the Negev, Beer-Sheva, Israel (zohary@bgu.ac.il).

where this was first developed. There exist several different ways to express these linear functionals that extract the coefficients. One can use functionals acting on the solution of the boundary value problem, and these can then have a simple explicit form and can be localized. Or one can write them as functionals acting directly on the right-hand side. These are the dual singular functions, properly speaking, and they are solutions of a boundary value problem themselves (see [5, 15, 16, 7, 2, 34]).

2. In the case of an *edge* in three dimensions, the space of singular functions is infinite-dimensional. Theoretical formulas for the extraction of coefficients then involve an infinite number of dual singular functions in general; see [22, 26]. The coefficients can be understood as functions defined on the edge, and their computation now requires approximation of function spaces on the edge. There exist some papers describing versions of the singular function method in this case. In [13], the case of a half-space crack in three-dimensional elasticity is considered. An algorithm is proposed and analyzed consisting of boundary elements on the crack surface combined with singular elements that are parametrized by one-dimensional finite elements on the crack front. This method and the corresponding error analysis are described for smooth curved cracks in three dimensions in [31]. In [19], the simple case of a circular edge is treated with Fourier expansion, error estimates are given, and results of numerical computations are shown.

Every linear functional acting on the edge coefficient functions now gives rise to a dual singular function. Such linear functionals can be the point evaluation at each point of the edge or, more regularly, moments, i.e., scalar products with some polynomial basis functions. Computing a finite number of such point values or moments, one obtains an approximation of the coefficient function. Such a procedure has been studied in [20] for the simple case of the Laplace equation at a flat crack. In [30] the coefficients are given by convolution integrals which contain the dual singular functions, and examples for the Lamé system are provided.

With the exception of the computations in the case of the simple geometries and operators of [19] and [20], the formulas and theoretical algorithms for the extraction of *edge coefficients* mentioned above have not lead to numerical implementations or serious computational results. A first step towards an algorithm suitable for implementation in an engineering stress analysis code is described in [36], where point values of edge coefficients are computed in the case of the Laplace equation near a straight edge. Very special orthogonality conditions of the Laplace edge singular functions are used to construct extraction formulas that are essentially two-dimensional.

Whereas this idea cannot be extended directly to more general geometrical and physical situations like Lamé equations in a polyhedral domain, our paper is an extension of [36] to such situations in the practical sense of suitability for implementation in engineering codes.

1.2. Outline. In the present paper we construct an algorithm for the approximate computation of moments of the edge coefficient functions. The algorithm has a twofold purpose: It is sufficiently general to be applicable to real-life three-dimensional boundary value problems and their singularities near polyhedral edges, and it is simple enough to be implemented in the framework of professional finite element codes. In a forthcoming paper we will show practical applications in the computation of stress concentration coefficients in three-dimensional anisotropic elasticity.

Our paper is organized as follows:

After a more detailed description of the idea of our algorithm in this first section, we recall in section 2 the structure of edge singularities for second-order linear Dirichlet

boundary value problems in three dimensions. We describe how the leading term in each singular function is obtained from a two-dimensional problem in a sector and can be computed from the principal Mellin symbol of the partial differential operator. For a complete description of the singular function one has to construct higher order "shadow terms," for which we also give formulas involving Mellin symbols of the operator.

In section 3, the structure of dual singular functions is described first in two dimensions and then for the case of the three-dimensional edge. The dual singular functions have an asymptotic expansion in terms that have tensor product form in cylindrical coordinates and are homogeneous with respect to the distance to the edge. This form allows us to prove a certain approximate duality between finite partial sums of these asymptotic expansions. These sums can be constructed explicitly from the Mellin symbols of the operator, and the duality holds approximately on cylindrical domains in the sense that the error is of the order of an arbitrarily high power of the radius of the cylinder.

In section 4, we construct the extraction algorithm for moments of the coefficients of the edge singularities. The algorithm requires the integration of the solution of the boundary value problem against a smooth function on a cylindrical surface of distance R to the edge, and it is exact modulo a given arbitrarily high power of R.

In section 5, we discuss generalizations to more general domains and boundary conditions, and the special case of a crack.

In section 6, we compare our algorithm with possible alternatives based on other formulas for the extraction of coefficients.

1.3. The main framework. Any three-dimensional elliptic boundary value problem posed on a polyhedron defines infinite-dimensional singularity spaces corresponding to each of the *edges*. Each singularity along an edge E is characterized

- by an *exponent* α, which is a complex number depending only on the geometry and the operator, and which determines the level of nonsmoothness of the singularity, and
- by a *coefficient* a_α, which is a function along the edge E.

Of great interest are the coefficients a_α when $\operatorname{Re}\alpha$ is less than 1, corresponding to non-H^2 solutions. In many situations, $\operatorname{Re}\alpha < 1$ when the opening at the edge is nonconvex. For example, α can be equal to $\frac{1}{2}$ in elasticity problems in the presence of cracks. Sometimes in such a situation the coefficients are called *stress intensity factors*. Herein we propose a method for the computation of these coefficients, which can be applied to any edge (including a crack front) of any polyhedron.

For the exposition of the method we use a model domain Ω where only one edge E is of interest (in particular, E will be the only possible nonconvex edge). Nevertheless this method applies, almost without alteration, to any polyhedron; see section 5.

As model domain, we take the tensor product $\Omega = G \times I$, where I is an interval, let us say $[-1, 1]$, and G is a plane bounded sector of opening $\omega \in (0, 2\pi]$ and radius 1 (the case of a crack, $\omega = 2\pi$, is included). See Figure 1. The variables are (x, y) in G and z in I, and we denote the coordinates (x, y, z) by $\mathbf{x}$. Let (r, θ) be the polar coordinates centered at the vertex of G so that $G = \{(x, y) \in \mathbb{R}^2 \mid r \in (0, 1),\ \theta \in (0, \omega)\}$. The domain Ω has an *edge* E which is the set $\{(x, y, z) \in \mathbb{R}^3 \mid r = 0,\ z \in I\}$.

The operator L is a homogeneous second-order partial differential $N \times N$ system with constant real coefficients, which means that

$$L = \sum_{j=1}^{3} \sum_{i=1}^{3} L_{ij} \partial_i \partial_j \quad \text{with} \quad \partial_1 = \frac{\partial}{\partial x},\ \partial_2 = \frac{\partial}{\partial y},\ \partial_3 = \frac{\partial}{\partial z},$$

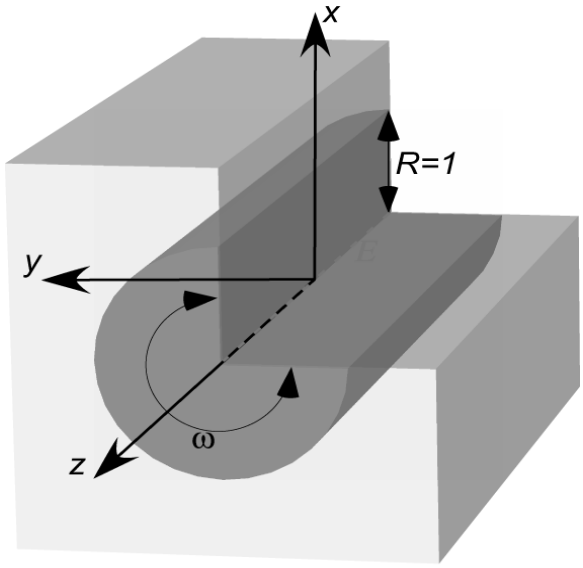


FIG. 1. *The domain of interest* Ω.

with coefficient matrices L_{ij} in $\mathbb{R}^{N\times N}$. We moreover assume that the matrices L_{ij} are symmetric. Therefore L is formally self-adjoint.

We assume moreover that L is associated with an elliptic bilinear form B, i.e., that for any u and v in $H^2(\Omega)^N$ and any subdomain $\Omega' \subset \Omega$ there holds

$$\begin{aligned}\int_{\Omega'} Lu\cdot v\ \mathrm{d}\mathbf{x} &= B(u,v)+\int_{\Gamma'} T_{\Gamma'}u\cdot v\ \mathrm{d}\sigma\\ &= \int_{\Omega'} u\cdot Lv\ \mathrm{d}\mathbf{x}+\int_{\Gamma'}\left(T_{\Gamma'}u\cdot v-u\cdot T_{\Gamma'}v\right)\mathrm{d}\sigma,\end{aligned}\tag{1.1}$$

where $T_{\Gamma'}$ is the Neumann trace operator associated with L via B on the boundary Γ' of Ω'. Our aim is the determination of the edge structure of any solution u of the problem

$$u\in H^1_0(\Omega)\quad \forall v\in H^1_0(\Omega)\quad B(u,v)=\int_\Omega f\cdot v\ \mathrm{d}\mathbf{x},\tag{1.2}$$

where f is a smooth vector function in $\mathcal{C}^\infty(\overline{\Omega})^N$. Away from the end points of the edge, the solution u can be expanded in edge singularities $S[\alpha\,;a_\alpha]$ associated with the exponents α and the coefficients a_α. These singularities $S[\alpha\,;a_\alpha]$ are the sums of terms in tensor product form $\partial_z^j a_\alpha(z)\,\Phi_j[\alpha](x,y)$, where only the generating coefficients a_α depend on the right-hand side f of problem (1.2).

1.4. The extraction method. In this paper, we construct for each exponent α a set of *quasi-dual* singular functions $K^m[\alpha\,;b]$, where m is a natural integer, which is the *order* of the quasi-dual function, and b a test coefficient. We then extract not the pointwise values of a_α, but its scalar product versus b on E with the help of the following *antisymmetric* internal boundary integrals $J[R]$ over the surface

$$\Gamma_R := \left\{\mathbf{x}\in\mathbb{R}^3 \mid r=R,\ \theta\in(0,\omega),\ z\in I\right\},$$

depending on the radius R:

$$J[R](u,v) := \int_{\Gamma_R} \left(T_{\Gamma_R} u \cdot \overline{v} - u \cdot T_{\Gamma_R} \overline{v}\right) \mathrm{d}\sigma. \tag{1.3}$$

Roughly, and with certain limitations (see Theorem 4.3 and its extensions in section 5), we find that for the lowest values of $\operatorname{Re}\alpha$, there holds

$$J[R](u, K^m[\alpha\,;b]) = \int_I a_\alpha(z)\,\bar{b}(z)\,\mathrm{d}z + \mathcal{O}\left(R^{m+1}\right) \quad \text{as} \quad R \to 0, \tag{1.4}$$

which allows a precise determination of $\int_I a_\alpha\,\bar{b}$ by extrapolation in R and a reconstruction of a_α by the choice of a suitable set of test coefficients b.

One of the fundamental tools for the proof of (1.4) consists of algebraic relations based on integration by parts in the domains $\Omega_{\varepsilon,R}$, where for any ε and R with $0 < \varepsilon < R$ we denote by $G_{\varepsilon,R}$ the annulus

$$G_{\varepsilon,R} := \{(x,y) \in \mathbb{R}^2 \mid r \in (\varepsilon, R),\ \theta \in (0,\omega)\}$$

and by $\Omega_{\varepsilon,R}$ the tensor domain $G_{\varepsilon,R} \times I$. We note that

$$\partial\Omega_{\varepsilon,R} = \Gamma_\varepsilon \cup \Gamma_R \cup (G_{\varepsilon,R} \times \partial I).$$

Finally we also denote by G_∞ the infinite sector of opening ω and by Ω_∞ the infinite wedge $G_\infty \times I$.

2. Edge singularities. Edge singularities are investigated in several works. Let us quote Maz'ya and Plamenevskii [24], Maz'ya and Rossmann [27], Dauge [14], and Costabel and Dauge [9]. Here, as a model problem, we concentrate on the simplest case of a homogeneous operator with constant coefficients.

The structure and the expansion of edge singularities rely on the splitting of the operator L into three parts,

$$L = M_0(\partial_x, \partial_y) + M_1(\partial_x, \partial_y)\,\partial_z + M_2\,\partial_z^2,$$

where M_0 is an $N \times N$ matrix of second-order partial differential operators in (x,y), M_1 is an $N \times N$ matrix of first-order partial differential operators in (x,y), and M_2 is a scalar $N \times N$ matrix.

We can check that for any smooth function $a(z)$ in I and any sequence $(\Phi_j)_{j\geq 0}$ of functions of (x,y) satisfying the relations

$$\begin{cases} M_0\Phi_0 = 0, \\ M_0\Phi_1 + M_1\Phi_0 = 0, & \text{in } G_\infty, \\ M_0\Phi_j + M_1\Phi_{j-1} + M_2\Phi_{j-2} = 0, \qquad j \geq 2, \end{cases} \tag{2.1}$$

the series

$$u \sim \sum_{j\geq 0} \partial_z^j a(z)\,\Phi_j(x,y)$$

formally satisfies the equation $Lu \sim 0$ in Ω_∞. If, moreover, all derivatives of a are zero in ± 1 and if the Φ_j satisfy the Dirichlet conditions on ∂G_∞, then $u \sim 0$ on $\partial\Omega_\infty$. In order to provide a more precise meaning, we need a description of solutions of the system of equations (2.1).

2.1. Two-dimensional leading singularities. The first terms Φ_0 are the solutions of the Dirichlet problem in the infinite sector:

$$\left\{\begin{array}{rcll} M_0\Phi_0 &=& 0 & \text{in} \quad G_\infty, \\ \Phi_0 &=& 0 & \text{on} \quad \partial G_\infty. \end{array}\right. \tag{2.2}$$

From the general theory we know that the solutions of problem (2.2) are generated by functions having the particular form in polar coordinates (r,θ)

$$\Phi_0 = r^\alpha \varphi_0(\theta), \quad \alpha \in \mathbb{C}. \tag{2.3}$$

Since it is homogeneous of degree 2, the system M_0 can be written in polar coordinates in the form

$$M_0(\partial_x, \partial_y) = r^{-2}\mathcal{M}_0(\theta; r\partial_r, \partial_\theta).$$

With the ansatz (2.3), the system (2.2) becomes

$$\left\{\begin{array}{rcll} \mathcal{M}_0(\theta;\alpha,\partial_\theta)\varphi_0 &=& 0 & \text{in} \quad (0,\omega), \\ \varphi_0 &=& 0 & \text{on} \quad 0 \quad \text{and} \quad \omega. \end{array}\right. \tag{2.4}$$

The operator $\varphi \mapsto \mathcal{M}_0(\theta;\alpha,\partial_\theta)\varphi$ acting from $H_0^1(0,\omega)$ into $H^{-1}(0,\omega)$ is the *Mellin symbol* of M_0, and we denote it by $\mathfrak{M}_0(\alpha)$.

The system (2.4) has nonzero solutions, i.e., $\mathfrak{M}_0(\alpha)$ is not invertible, only for a discrete subset $\mathfrak{A} = \mathfrak{A}(M_0)$ of $\mathbb{C}$. We call the numbers $\alpha \in \mathfrak{A}$ the *edge exponents*.

The ellipticity of L implies the ellipticity of M_0, and as a consequence, any strip $\operatorname{Re}\alpha \in (\xi_1,\xi_2)$ contains at most a finite number of elements of $\mathfrak{A}$. As the coefficients of M_0 are real, if α belongs to $\mathfrak{A}$, then $\bar\alpha$ also belongs to $\mathfrak{A}$. Moreover we have the general property that

$$\mathfrak{M}_0(\alpha)^* = \mathfrak{M}_0^*(-\bar\alpha),$$

where $\mathfrak{M}_0(\alpha)^*$ is the adjoint of $\mathfrak{M}_0(\alpha)$ and $\mathfrak{M}_0^*$ denotes the Mellin symbol of the adjoint M_0^* of M_0. Now M_0 is formally self-adjoint: $M_0^* = M_0$, and there holds

$$\mathfrak{M}_0(\alpha)^* = \mathfrak{M}_0(-\bar\alpha).$$

By the Fredholm alternative, this implies that if α belongs to $\mathfrak{A}$, then $-\bar\alpha$ also belongs to $\mathfrak{A}$.

The operator valued function $\alpha \mapsto \mathfrak{M}_0(\alpha)^{-1}$ is meromorphic on $\mathbb{C}$. If

$$\forall \alpha \in \mathfrak{A}, \quad \alpha \text{ is a pole of degree 1 of } \mathfrak{M}_0^{-1}, \tag{$\mathfrak{H}_1$}$$

then any solution of (2.2) is a linear combination of solutions of type (2.3) with $\alpha \in \mathfrak{A}$ and φ_0 a nonzero solution of (2.4). For simplicity we assume hypothesis ($\mathfrak{H}_1$) and will explain in what follows the implications if it does not hold.

2.2. Further two-dimensional generators for singularities. The second equation of system (2.1) with Dirichlet conditions reduces to finding Φ_1 such that

$$\left\{\begin{array}{rcll} M_0\Phi_1 &=& -M_1\Phi_0 & \text{in} \quad G_\infty, \\ \Phi_1 &=& 0 & \text{on} \quad \partial G_\infty, \end{array}\right. \tag{2.5}$$

where $\Phi_0 = r^\alpha \varphi_0(\theta)$ as determined in the previous subsection. Since it is homogeneous of degree 1, the system M_1 can be written in polar coordinates in the form

$$M_1(\partial_x, \partial_y) = r^{-1} \mathcal{M}_1(\theta; r\partial_r, \partial_\theta).$$

Therefore, $M_1\Phi_0 = r^{\alpha-1}\mathcal{M}_1(\theta; \alpha, \partial_\theta)\varphi_0$ and an ansatz like (2.3) for the solution of problem (2.5) is

$$\Phi_1 = r^{\alpha+1}\varphi_1(\theta), \tag{2.6}$$

with φ_1 solution of the Dirichlet problem

$$\begin{cases} \mathcal{M}_0(\theta; \alpha+1, \partial_\theta)\varphi_1 = -\mathcal{M}_1(\theta; \alpha, \partial_\theta)\varphi_0 & \text{in } (0, \omega), \\ \varphi_1 = 0 & \text{on } 0 \text{ and } \omega; \end{cases} \tag{2.7}$$

in other words, φ_1 solves $\mathfrak{M}_0(\alpha+1)\varphi_1 = -\mathcal{M}_1(\alpha)\varphi_0$. Therefore, if $\alpha+1$ does not belong to $\mathfrak{A}$, the previous problem has a unique solution. This is why we assume hypothesis $(\mathfrak{H}_2)$:

$$\forall \alpha \in \mathfrak{A}, \quad \forall j \in \mathbb{N},\ j \geq 1, \quad \alpha + j \notin \mathfrak{A}. \tag{$\mathfrak{H}_2$}$$

If $(\mathfrak{H}_2)$ holds, then for each solution $\Phi_0 = r^\alpha\varphi_0$ of problem (2.4), we obtain by induction a unique sequence $(\Phi_j)_{j\geq 0}$ solution of (2.1) with Dirichlet conditions in the form

$$\Phi_j = r^{\alpha+j}\varphi_j(\theta),$$

where φ_j solves

$$\mathfrak{M}_0(\alpha+j)\varphi_j = -\mathcal{M}_1(\alpha+j-1)\varphi_{j-1} - M_2\varphi_{j-2}\,. \tag{2.8}$$

We recall that M_2, being a scalar matrix, has the same expression in Cartesian coordinates as in polar coordinates (viz. $M_2 = \mathcal{M}_2$).

2.3. Three-dimensional singularities. Assuming hypotheses $(\mathfrak{H}_1)$ and $(\mathfrak{H}_2)$, for any $\alpha \in \mathfrak{A}$ with $\operatorname{Re}\alpha > 0$, let p_α denote the dimension of the kernel of $\mathfrak{M}_0(\alpha)$ and let $\Phi_0[\alpha, p]$, for $p = 1, \ldots, p_\alpha$, be a basis of $\ker \mathfrak{M}_0(\alpha)$. Moreover, for any $j \geq 1$, let $\Phi_j[\alpha, p]$ be the solution of (2.7) or (2.8) (also called "*shadow singularities*") generated by $\Phi_0[\alpha, p]$.

For any integer $n \geq 0$ we call "*singularity at the order n*" any expression of the form

$$S^n[\alpha, p\,;\, a] := \sum_{j=0}^{n} \partial_z^j a(z)\, \Phi_j[\alpha, p](x, y), \tag{2.9}$$

where a belongs to $\mathcal{C}^{n+2}(\overline{I})$.

By construction, there holds

$$LS^n[\alpha, p\,;\, a] = \partial_z^{n+1} a\,(M_1\Phi_n + M_2\Phi_{n-1}) + \partial_z^{n+2} a\, M_2\Phi_n, \tag{2.10}$$

whence we have the following lemma.

LEMMA 2.1. *For any $\alpha \in \mathfrak{A}$, $\operatorname{Re}\alpha > 0$, and $a \in \mathcal{C}^{n+2}(\overline{I})$ we have*

$$LS^n[\alpha, p\,;\, a] = \mathcal{O}\left(r^{\operatorname{Re}\alpha+n-1}\right); \tag{2.11}$$

i.e., $r^{-\operatorname{Re}\alpha-n+1} LS^n[\alpha, p\,;\, a]$ is bounded in $\overline{\Omega}$. Moreover $S^n[\alpha, p\,;\, a] = 0$ on $\partial G_\infty \times I$.

3. Dual singular functions. We first recall and reformulate well-known facts about the dual singular functions for two-dimensional problems (cf. Maz'ya and Plamenevskii [21, 23, 25], Babuška and Miller [2], Dauge et al. [15, 16]) and then extend these notions in the framework of our edge problem, so that we obtain what we call "*quasi-dual singular functions*" (compare with the extraction functions in [1] by Andersson, Falk, and Babuška) as opposed to exact dual singular functions; cf. Maz'ya and Plamenevskii [22], Maz'ya and Rossmann [26] (pointwise duality), and Lenczner [20] (Sobolev duality).

3.1. Two-dimensional dual singular functions. The two-dimensional operator is the homogeneous second-order operator M_0 with real coefficients. We develop its symbol $\mathcal{M}_0(\alpha)$ in powers of α (of degree 2):

$$\mathcal{M}_0(\theta;\alpha,\partial_\theta) = \mathcal{N}_0(\theta;\partial_\theta) + \alpha\mathcal{N}_1(\theta;\partial_\theta) + \alpha^2\mathcal{N}_2(\theta). \tag{3.1}$$

Since M_0 is self-adjoint, we can deduce that

$$\mathcal{N}_0 \text{ and } \mathcal{N}_2 \text{ are self-adjoint} \quad \text{and} \quad \mathcal{N}_1 \text{ is anti-self-adjoint.} \tag{3.2}$$

LEMMA 3.1. *Let α, β be in $\mathfrak{A}$ and let φ, ψ be in the kernels of $\mathfrak{M}_0(\alpha)$, $\mathfrak{M}_0(\beta)$, respectively. Then there holds the identity*

$$(\alpha+\bar\beta)\int_0^\omega \big(\mathcal{N}_1 + (\alpha-\bar\beta)\mathcal{N}_2\big)\varphi\cdot\overline{\psi}\,d\theta = 0. \tag{3.3}$$

Proof. We start with the duality relation:

$$0 \;=\; \int_0^\omega \varphi\cdot\overline{\mathfrak{M}_0(\beta)\psi} = \int_0^\omega \mathfrak{M}_0(\beta)^*\varphi\cdot\overline{\psi} \;=\; \int_0^\omega \mathfrak{M}_0(-\bar\beta)\varphi\cdot\overline{\psi}.$$

Then we use the identity

$$\mathfrak{M}_0(-\bar\beta) = \mathfrak{M}_0(\alpha) - (\bar\beta+\alpha)\mathcal{N}_1 + (\bar\beta^2-\alpha^2)\mathcal{N}_2.$$

From $\mathfrak{M}_0(\alpha)\varphi = 0$, we obtain

$$\begin{aligned} 0 \;=\; \int_0^\omega \mathfrak{M}_0(-\bar\beta)\varphi\cdot\overline{\psi} &= \int_0^\omega \big(-(\bar\beta+\alpha)\mathcal{N}_1 + (\bar\beta^2-\alpha^2)\mathcal{N}_2\big)\varphi\cdot\overline{\psi} \\ &= -(\alpha+\bar\beta)\int_0^\omega \big(\mathcal{N}_1+(\alpha-\bar\beta)\mathcal{N}_2\big)\varphi\cdot\overline{\psi}. \qquad \square \end{aligned}$$

LEMMA 3.2. *Let α, β, φ, and ψ be as in Lemma* 3.1.
(i) *If $-\bar\beta \neq \alpha$, then*

$$\int_0^\omega \big(\mathcal{N}_1+(\alpha-\bar\beta)\mathcal{N}_2\big)\varphi\cdot\overline{\psi} = 0. \tag{3.4}$$

(ii) *If $-\bar\beta = \alpha$, then the left-hand side of* (3.4) *becomes*

$$\int_0^\omega \big(\mathcal{N}_1+2\alpha\mathcal{N}_2\big)\varphi\cdot\overline{\psi} = \int_0^\omega \left(\frac{d}{d\alpha}\mathfrak{M}_0(\alpha)\right)\varphi\cdot\overline{\psi}, \tag{3.5}$$

and, if we moreover assume hypothesis $(\mathfrak{H}_1)$*, then for any basis* $\big(\varphi[\alpha,p]\big)_p$ *of* $\ker\mathfrak{M}_0(\alpha)$ *there exists a unique dual basis* $\big(\psi[\alpha,p]\big)_p$ *of* $\ker\mathfrak{M}_0(-\bar\alpha)$ *such that*

$$\int_0^\omega \big(\mathcal{N}_1+2\alpha\mathcal{N}_2\big)\varphi[\alpha,p]\cdot\overline{\psi}[\alpha,q]=\delta_{p,q}. \tag{3.6}$$

Proof. (i) is a straightforward consequence of Lemma 3.1.

(ii) Identity (3.5) is clear. Concerning (3.6), we first note that since $\mathfrak{M}_0(\alpha)^*=\mathfrak{M}_0(-\bar\alpha)$, the dimension of the kernel of $\mathfrak{M}_0(\alpha)$ is equal to the codimension of the closure of the range of $\mathfrak{M}_0(-\bar\alpha)$. On the other hand, as for any $\alpha'\in\mathbb{C}\setminus\mathfrak{A}$, $\mathfrak{M}_0(\alpha')$ is invertible, and since $\mathfrak{M}_0(\alpha)-\mathfrak{M}_0(\alpha')$ is a compact operator, $\mathfrak{M}_0(\alpha)$ is a Fredholm operator of index 0. As a consequence,

$$\dim\ker\mathfrak{M}_0(\alpha)=\dim\ker\mathfrak{M}_0(-\bar\alpha).$$

In order to obtain (3.6) it suffices now to prove that if $\varphi\in\ker\mathfrak{M}_0(\alpha)$ satisfies

$$\forall\psi\in\ker\mathfrak{M}_0(-\bar\alpha),\quad \int_0^\omega\big(\mathcal{N}_1+2\alpha\mathcal{N}_2\big)\varphi\cdot\overline{\psi}=0,$$

then $\varphi=0$. If this does not hold, thanks to (3.5) there exists $\varphi\in\ker\mathfrak{M}_0(\alpha)$ such that

$$\forall\psi\in\ker\mathfrak{M}_0(-\bar\alpha),\quad \int_0^\omega\left(\frac{d}{d\alpha}\,\mathfrak{M}_0(\alpha)\right)\varphi\cdot\overline{\psi}=0.$$

By the Fredholm alternative, there exists φ' such that

$$\mathfrak{M}_0(\alpha)\varphi'+\frac{d}{d\alpha}\,\mathfrak{M}_0(\alpha)\varphi=0.$$

As a consequence the function

$$\alpha'\quad\longmapsto\quad(\alpha'-\alpha)^{-2}\,\mathfrak{M}_0(\alpha')\big(\varphi+(\alpha'-\alpha)\varphi'\big)$$

has an analytic extension in α. This contradicts hypothesis $(\mathfrak{H}_1)$ according to which $\mathfrak{M}_0^{-1}$ has a pole of order 1 in α. □

We end this subsection with a relation between the expression in the left-hand sides of (3.4) and (3.6) and a trace obtained by integration by parts.

Considering the Green formula (1.1) in the domain $\Omega_{\varepsilon,R}$ for functions u and v, which are zero on the two faces $\theta=0$ and $\theta=\omega$ of Ω, we have contributions on the parts Γ_R and Γ_ε of the boundary of $\Omega_{\varepsilon,R}$, where $r=R$ and $r=\varepsilon$, respectively. We denote by $T(r)$ the Neumann trace operator on Γ_r. It has the form

$$T(r)=T(r,\theta;\partial_r,\partial_\theta,\partial_z)=r^{-1}T_0(\theta;r\partial_r,\partial_\theta)+T_1(\theta)\,\partial_z\,. \tag{3.7}$$

We also have contributions of the lateral sides $G_{\varepsilon,R}\times\partial I$. Denoting by $T_{\partial I}$ the Neumann trace on these sides, we have the Green formula

$$\begin{aligned}\int_{\Omega_{\varepsilon,R}}Lu\cdot v-u\cdot Lv\,\mathrm{d}\mathbf{x}\;=\;&\int_I\int_0^\omega T(R)u\cdot v-u\cdot T(R)v\;R\,\mathrm{d}\theta\,\mathrm{d}z\\&-\int_I\int_0^\omega T(\varepsilon)u\cdot v-u\cdot T(\varepsilon)v\;\varepsilon\,\mathrm{d}\theta\,\mathrm{d}z\\&+\int_{G_{\varepsilon,R}\times\partial I}T_{\partial I}\,u\cdot v-u\cdot T_{\partial I}\,v\,\mathrm{d}\sigma.\end{aligned} \tag{3.8}$$

Applying the above identity to functions u and v independent of z (and zero on the two sides $\theta = 0$ and $\theta = \omega$), we note that the contributions on the two sides $G_{\varepsilon,R} \times \{\pm 1\}$ cancel out because the two Neumann operators $T_{\pm 1}$ which compose $T_{\partial I}$ are opposite to each other. Thus we obtain

$$\begin{aligned}\int_{G_{\varepsilon,R}} M_0 u \cdot v - u \cdot M_0 v \,\mathrm{d}x\,\mathrm{d}y \;&=\; \int_0^\omega T_0(R)u \cdot v - u \cdot T_0(R)v \,\mathrm{d}\theta \\ &\quad - \int_0^\omega T_0(\varepsilon)u \cdot v - u \cdot T_0(\varepsilon)v \,\mathrm{d}\theta,\end{aligned} \tag{3.9}$$

where $T_0(R)$ denotes $T_0(\theta; R\partial_r, \partial_\theta)$.

LEMMA 3.3. *Let α and β be complex numbers and let φ and ψ belong to $H_0^1(0,\omega)^N$. Set $\Phi := r^\alpha \varphi(\theta)$ and $\Psi := r^{-\bar\beta}\psi(\theta)$. For any $R > 0$ there holds*

$$\int_0^\omega T_0(R)\Phi \cdot \overline{\Psi} - \Phi \cdot T_0(R)\overline{\Psi} \,\mathrm{d}\theta = R^{\alpha-\beta} \int_0^\omega \big(\mathcal{N}_1 + (\alpha+\beta)\mathcal{N}_2\big)\varphi \cdot \overline{\psi} \,\mathrm{d}\theta\,. \tag{3.10}$$

Proof. Formula (3.9) and the splitting (3.1) of $\mathcal{M}_0 = r^2 M_0$ yield for any $\varepsilon < R$

$$\begin{aligned}\int_{G_{\varepsilon,R}} \Big(\big(\mathcal{N}_0 + r\partial_r\mathcal{N}_1 + (r\partial_r)^2\mathcal{N}_2\big)\Phi \cdot \overline{\Psi} \;-\; \Phi \cdot \big(\mathcal{N}_0 + r\partial_r\mathcal{N}_1 + (r\partial_r)^2\mathcal{N}_2\big)\overline{\Psi}\Big)\tfrac{1}{r}\,\mathrm{d}r\,\mathrm{d}\theta \\ = \int_0^\omega T_0(R)\Phi \cdot \overline{\Psi} - \Phi \cdot T_0(R)\overline{\Psi} \,\mathrm{d}\theta \\ - \int_0^\omega T_0(\varepsilon)\Phi \cdot \overline{\Psi} - \Phi \cdot T_0(\varepsilon)\overline{\Psi} \,\mathrm{d}\theta.\end{aligned}$$

Since $\mathcal{N}_0$ is self-adjoint, integration by parts gives

$$\begin{aligned}\int_\varepsilon^R \Big(\big(\mathcal{N}_0 + r\partial_r\mathcal{N}_1 + (r\partial_r)^2\mathcal{N}_2\big)\Phi \cdot \overline{\Psi} \;-\; \Phi \cdot \big(\mathcal{N}_0 + r\partial_r\mathcal{N}_1 + (r\partial_r)^2\mathcal{N}_2\big)\overline{\Psi}\Big)\tfrac{1}{r}\,\mathrm{d}r \\ = \big[\mathcal{N}_1\Phi \cdot \overline{\Psi} + (r\partial_r)\mathcal{N}_2\Phi \cdot \overline{\Psi} - \Phi \cdot (r\partial_r)\mathcal{N}_2\overline{\Psi}\big]_\varepsilon^R.\end{aligned}$$

We have

$$\mathcal{N}_1\Phi \cdot \overline{\Psi} + (r\partial_r)\mathcal{N}_2\Phi \cdot \overline{\Psi} - \mathcal{N}_2\Phi \cdot (r\partial_r)\overline{\Psi} = r^{\alpha-\beta}\big(\mathcal{N}_1\varphi \cdot \overline{\psi} + \alpha\mathcal{N}_2\varphi \cdot \overline{\psi} + \varphi \cdot \beta\mathcal{N}_2\overline{\psi}\big),$$

and as $\mathcal{N}_2$ is self-adjoint (cf. (3.2)), we finally obtain

$$\begin{aligned}\big(\mathcal{N}_1\varphi \cdot \overline{\psi} + (\alpha+\beta)\mathcal{N}_2\varphi \cdot \overline{\psi}\big)\big(R^{\alpha-\beta} - \varepsilon^{\alpha-\beta}\big) \;&=\; \int_0^\omega T_0(R)\Phi \cdot \overline{\Psi} - \Phi \cdot T_0(R)\overline{\Psi} \,\mathrm{d}\theta \\ &\quad - \int_0^\omega T_0(\varepsilon)\Phi \cdot \overline{\Psi} - \Phi \cdot T_0(\varepsilon)\overline{\Psi} \,\mathrm{d}\theta.\end{aligned}$$

Now the right-hand side of the above equality also has the form $c(\alpha,\beta)(R^{\alpha-\beta} - \varepsilon^{\alpha-\beta})$, and we deduce (3.10) for any $\alpha \neq \beta$. Since, for fixed β, φ, ψ, and R, both members of (3.10) depend continuously on α, we deduce (3.10) for $\alpha = \beta$ by continuity. ∎

3.2. Three-dimensional dual singular functions. We assume hypotheses $(\mathfrak{H}_1)$ and $(\mathfrak{H}_2)$, and for any $\alpha \in \mathfrak{A}$, $\operatorname{Re}\alpha > 0$, we choose a basis $\varphi[\alpha,p]$, $p = 1,\dots,p_\alpha$, of $\ker \mathfrak{M}_0(\alpha)$. Then we denote by $\psi[\alpha,p]$, $p = 1,\dots,p_\alpha$, the corresponding dual basis according to Lemma 3.2. We recall that we have denoted $r^\alpha\varphi[\alpha,p]$ by $\Phi_0[\alpha,p]$ and that associated singularities at the order n are defined in (2.9).

Following the same lines, we set

$$\Psi_0[\alpha,p] := r^{-\bar\alpha}\psi[\alpha,p],$$

and for any integer $n \geq 0$, we define the "*quasi-dual singular function at the order n*" by

$$K^n[\alpha,p\,;\,b] := \sum_{j=0}^{n} \partial_z^j b(z)\,\Psi_j[\alpha,p](x,y), \tag{3.11}$$

where b belongs to $\mathcal{C}^{n+2}(\overline{I})$ and the sequence $(\Psi_j)_{j\geq 0}$ is defined by induction as solution of (2.1) in the form

$$\Psi_j = r^{-\bar\alpha+j}\psi_j(\theta),$$

where ψ_j solves

$$\mathfrak{M}_0(-\bar\alpha+j)\psi_j = -\mathcal{M}_1(-\bar\alpha+j-1)\psi_{j-1} - M_2\psi_{j-2}. \tag{3.12}$$

Of course, $K^n[\alpha,p\,;\,b]$ is but $S^n[-\bar\alpha,p\,;\,b]$ (generated by Ψ_0). Therefore by (2.11) there holds, for any $\alpha \in \mathfrak{A}$, $\operatorname{Re}\alpha > 0$, and $b \in \mathcal{C}^{n+2}(\overline{I})$,

$$LK^n[\alpha,p\,;\,b] = \mathcal{O}\left(r^{-\operatorname{Re}\alpha+n-1}\right). \tag{3.13}$$

In the next proposition we state that the singularities $S^n[\alpha,p\,;\,a]$ and the quasi-dual singular functions $K^n[\beta,q\,;\,b]$ are in duality with each other (modulo a remainder) if linked by the following antisymmetric sesquilinear form:

$$J[R](u,v) := \int_{\Gamma_R} \left(Tu\cdot\overline{v} - u\cdot T\overline{v}\right)\mathrm{d}\sigma = \int_I\int_0^\omega \left(Tu\cdot\overline{v} - u\cdot T\overline{v}\right)\big|_{r=R}\, R\,\mathrm{d}\theta\,\mathrm{d}z, \tag{3.14}$$

where $T = T(R)$ is the radial Neumann trace operator (3.7).

Proposition 3.4. *Let $\alpha,\beta \in \mathfrak{A}$ with $\operatorname{Re}\alpha,\ \operatorname{Re}\beta > 0$. We assume that hypotheses $(\mathfrak{H}_1)$ and $(\mathfrak{H}_2)$ hold. For an integer $n \geq 0$, let the coefficients a and b be in $\mathcal{C}^{n+2}(\overline{I})$. We assume moreover that $\partial_z^j b = 0$ for $j = 0,\dots,n-1$ on ∂I. Then for any $R > 0$ there holds*

$$J[R]\Big(S^n[\alpha,p\,;\,a]\,,K^n[\beta,q\,;\,b]\Big) = \delta_{\alpha,\beta}\,\delta_{p,q}\int_I a(z)\,\bar b(z)\,\mathrm{d}z + \mathcal{O}\left(R^{\operatorname{Re}\alpha-\operatorname{Re}\beta+n+1}\right). \tag{3.15}$$

Proof. We use the Green formula (3.8) on $G_{\varepsilon,R}$ for

$$u = S^n[\alpha,p\,;\,a] \quad\text{and}\quad v = \overline{K^n[\alpha,q\,;\,b]}.$$

Since $u = \mathcal{O}\left(r^{\operatorname{Re}\alpha}\right)$ and $v = \mathcal{O}\left(r^{-\operatorname{Re}\beta}\right)$, (2.11) and (3.13) imply

$$\int_{\Omega_{\varepsilon,R}} Lu\cdot v - u\cdot Lv\,\mathrm{d}\mathbf{x} = \mathcal{O}\left(\int_\varepsilon^R r^{\operatorname{Re}\alpha-\operatorname{Re}\beta+n-1}\,r\mathrm{d}r\right).$$

With formula (2.10), we even obtain the more precise expression

$$\int_{\Omega_{\varepsilon,R}} Lu\cdot v - u\cdot Lv\,\mathrm{d}\mathbf{x} = \sum_{k=n-1}^{2n}\gamma_k\int_\varepsilon^R r^{\alpha-\beta+k}\,r\mathrm{d}r\,,$$

with coefficients γ_k independent of R and ε. As a consequence of hypothesis $(\mathfrak{H}_2)$ we know that $\alpha-\beta+k$ is different from -1 for $k=n,\dots,2n+1$. Thus

$$\int_{\Omega_{\varepsilon,R}} Lu\cdot v-u\cdot Lv\,\mathrm{d}\mathbf{x}=\sum_{k=n+1}^{2n+2}\lambda_k\big(R^{\alpha-\beta+k}-\varepsilon^{\alpha-\beta+k}\big),$$

with coefficients λ_k independent of R and ε. For the boundary integral $J[r](u,v)$ (3.14), we omit the mention of (u,v). Thus the Green formula (3.8) gives

$$J[R]-J[\varepsilon]+\int_{G_{\varepsilon,R}\times\partial I} T_{\partial I}u\cdot v-u\cdot T_{\partial I}v\;r\mathrm{d}r\,\mathrm{d}\theta=\sum_{k=n+1}^{2n+2}\lambda_k\big(R^{\alpha-\beta+k}-\varepsilon^{\alpha-\beta+k}\big).$$

As $T_{\partial I}$ is of the form $r^{-1}T_{\partial I,0}(\theta;r\partial_r,\partial_\theta)+T_{\partial I,1}(\theta)\,\partial_z$ (cf. (3.7)), and as the ends ∂I are zeros of order n of b, we are left with

$$T_{\partial I}u\cdot v-u\cdot T_{\partial I}v=T_{\partial I}u\cdot\partial_z^n b\,\Psi_n-u\cdot\partial_z^n b\,\big(r^{-1}T_{\partial I,0}\Psi_n+T_{\partial I,1}\Psi_{n-1}\big)-u\cdot\partial_z^{n+1}b\,T_{\partial I,1}\Psi_n.$$

Integrating on $G_{\varepsilon,R}\times\partial I$ and using the structure of Ψ_j, we obtain as before

$$T_{\partial I}u\cdot v-u\cdot T_{\partial I}v\;=\;\sum_{k=n+1}^{2n+2}\lambda'_k\big(R^{\alpha-\beta+k}-\varepsilon^{\alpha-\beta+k}\big).$$

From the last three equalities we obtain

$$J[R]-J[\varepsilon]=\sum_{k=n+1}^{2n+2}\lambda''_k\big(R^{\alpha-\beta+k}-\varepsilon^{\alpha-\beta+k}\big). \tag{3.16}$$

It remains to expand $J[r]$ in homogeneous parts: we have

$$J[r]=\sum_{k=0}^{2n+1}J_k r^{\alpha-\beta+k} \tag{3.17}$$

with (cf. (3.7))

$$\begin{aligned}J_k=&\sum_{j+\ell=k}\int_I\int_0^\omega \partial_z^j a\,\partial_z^\ell\bar b\,\Big(T_0(\theta;\alpha+j,\partial_\theta)\varphi_j\cdot\overline{\psi}_\ell-\varphi_j\cdot T_0(\theta;-\beta+\ell,\partial_\theta)\overline{\psi}_\ell\Big)\,\mathrm{d}\theta\mathrm{d}z\\&+\sum_{j+\ell=k-1}\int_I\int_0^\omega\Big(\partial_z^{j+1}a\,\partial_z^\ell\bar b\,T_1(\theta)\varphi_j\cdot\overline{\psi}_\ell-\partial_z^j a\,\partial_z^{\ell+1}\bar b\,\varphi_j\cdot T_1(\theta)\overline{\psi}_\ell\Big)\mathrm{d}\theta\mathrm{d}z.\end{aligned} \tag{3.18}$$

Combining (3.16) with (3.17) we obtain

$$\sum_{k=0}^{2n+1}J_k(R^{\alpha-\beta+k}-\varepsilon^{\alpha-\beta+k})=\sum_{k=n+1}^{2n+2}\lambda''_k\big(R^{\alpha-\beta+k}-\varepsilon^{\alpha-\beta+k}\big).$$

By identification of terms, we immediately deduce that

$$\forall k\le n,\quad J_k(R^{\alpha-\beta+k}-\varepsilon^{\alpha-\beta+k})=0.$$

Therefore

$$\forall k \leq n \ \text{ such that } \ \alpha - \beta + k \neq 0, \quad J_k = 0.$$

By hypothesis $(\mathfrak{H}_2)$, the number $\alpha - \beta + k$ can be 0 only if $k = 0$. Therefore

$$\forall k,\ 1 \leq k \leq n, \quad J_k = 0 \quad \text{and} \quad \forall \alpha, \beta \in \mathfrak{A},\ \alpha \neq \beta, \quad J_0 = 0.$$

In order to obtain (3.15), it remains to study J_0 when $\alpha = \beta$. Formula (3.18) yields, for J_0,

$$J_0 = \int_I \int_0^\omega a\,\bar{b}\left(T_0(\theta;\alpha,\partial_\theta)\varphi_0[\alpha,p]\cdot\overline{\psi}_0[\alpha,q] - \varphi_j[\alpha,p]\cdot T_0(\theta;-\beta,\partial_\theta)\overline{\psi}_0[\alpha,q]\right)\mathrm{d}\theta\mathrm{d}z.$$

Applying Lemma 3.3 for $\alpha = \beta$ we have

$$J_0 = \Big(\int_I a\,\bar{b}\,\mathrm{d}z\Big)\Big(\int_0^\omega (\mathcal{N}_1 + 2\alpha\mathcal{N}_2)\varphi[\alpha,p]\cdot\overline{\psi}[\alpha,q]\,\mathrm{d}\theta\Big),$$

and with the orthogonality relation (3.6) we deduce that

$$J_0 = \delta_{p,q}\int_I a\,\bar{b}\,\mathrm{d}z. \qquad \square$$

Note that in formula (3.18), we can integrate by parts in z without any boundary contribution for $k \leq n$, because $\partial_z^j b = 0$ for $j = 0,\dots,n-1$ on ∂I. Therefore

$$J_k = \Big(\int_I a\,\partial_z^k\bar{b}\,\mathrm{d}z\Big)\,H_k[\alpha,p\,;\,\beta,q], \tag{3.19}$$

where

$$\begin{aligned} H_k[\alpha,p\,;\,\beta,q] = &\sum_{j+\ell=k}\int_0^\omega (-1)^j\Big(T_0(\theta;\alpha+j,\partial_\theta)\varphi_j\cdot\overline{\psi}_\ell - \varphi_j\cdot T_0(\theta;-\beta+\ell,\partial_\theta)\overline{\psi}_\ell\Big)\,\mathrm{d}\theta \\ &- \sum_{j+\ell=k-1}\int_0^\omega (-1)^j\Big(T_1(\theta)\varphi_j\cdot\overline{\psi}_\ell + \varphi_j\cdot T_1(\theta)\overline{\psi}_\ell\Big)\mathrm{d}\theta. \end{aligned} \tag{3.20}$$

As a consequence of the proof of Proposition 3.4 we have

$$\forall \alpha,\ \beta \in \mathfrak{A}, \quad \forall p,\ q, \quad \forall k \in \mathbb{N}, \qquad H_k[\alpha,p\,;\,\beta,q] = \delta_{k,0}\,\delta_{\alpha,\beta}\,\delta_{p,q}. \tag{3.21}$$

Later on, we will use formula (3.21), and not Proposition 3.4, to extract the singularity coefficients of a true solution of problem (1.2).

4. Extraction of singularity coefficients. In this section, we first describe asymptotic expansions of the solution u of problem (1.2). The right-hand side f is $\mathcal{C}^\infty(\overline{\Omega})$, and we suppose in a preliminary step that $f \equiv 0$ in a neighborhood of the edge E. The expansions of u show edge singularity coefficients $a_{\alpha,p}$ along the edge E. We propose a method based on the duality formula (3.15) to determine these coefficients.

4.1. Expansion of the solution along the edge. The edge expansions are valid only away from the sides $G \times \partial I$. This is the reason why we introduce for any $\delta \in (0,1)$ the subinterval

$$I_\delta = (-1+\delta, 1-\delta)$$

and consider the subdomains $G \times I_\delta$. We need the introduction of weighted spaces to describe the remainders in the expansions. For $\xi \in \mathbb{R}$, let

$$\mathcal{V}_\eta(G \times I_\delta) := \big\{ v \in \mathcal{C}^\infty(G \times I_\delta) \mid \forall \mathbf{m} \in \mathbb{N}^3,\ r^{-\eta+|\mathbf{m}|} \partial_{\mathbf{x}}^{\mathbf{m}} v \in \mathrm{L}^\infty(G \times I_\delta) \big\}.$$

Then the following theorem holds; cf. [27].

THEOREM 4.1. *Let $\delta \in (0,1)$ and $\eta > 0$ be given. Then for any $\alpha \in \mathfrak{A}$ such that $\operatorname{Re}\alpha \in (0,\eta)$ and for any $p \in \{1,\dots,p_\alpha\}$, there exists a unique coefficient $a_{\alpha,p} \in \mathcal{C}^\infty(I_\delta)$ such that*

$$u - \sum_{\alpha,\ 0<\operatorname{Re}\alpha<\eta} \sum_{p} S^n\big[\alpha, p\,;\, a_{\alpha,p}\big] \in \mathcal{V}_\eta(G \times I_\delta), \tag{4.1}$$

where $n = n(\alpha)$ is the smallest integer such that $\operatorname{Re}\alpha + n > \eta$.

Letting δ tend to 0, this clearly defines unique coefficients $a_{\alpha,p}$ in $\mathcal{C}^\infty(I)$ such that for any δ (4.1) holds with $a_{\alpha,p}\big|_{I_\delta}$. But this *does not imply that* (4.1) holds in Ω, because in general the remainders on $G \times I_\delta$ depend on δ and their norms blow up as $\delta \to 0$. This is due to the presence of corner singularities at the corners $\mathbf{c}^\pm := (0,0,\pm 1)$. We have to analyze these corner singularities in order to obtain uniform estimates in Ω.

4.2. Corner exponents. We describe the situation in a neighborhood of the corner $\mathbf{c}^+$ and particularize the notation by the superscript $^+$. A similar situation holds for the other corner $\mathbf{c}^-$. Let K^+ be the infinite cone coinciding with Ω in a neighborhood of $\mathbf{c}^+$. Let $\mathbb{S}^+$ denote the sphere of radius 1 centered at $\mathbf{c}^+$, ρ^+ the distance to $\mathbf{c}^+$, and ϑ^+ the coordinates on $\mathbb{S}^+$. Thus (ρ^+, ϑ^+) are spherical coordinates centered at $\mathbf{c}^+$. Finally, let S^+ denote the intersection $\mathbb{S}^+ \cap K^+$. The operator L can be written in these spherical coordinates as

$$L = (\rho^+)^{-2} \mathcal{L}^+(\vartheta^+; \rho^+ \partial_{\rho^+}, \partial_{\vartheta^+}),$$

which defines the Mellin symbol $\gamma \mapsto \mathfrak{L}^+(\gamma)$ of L at $\mathbf{c}^+$, where $\mathfrak{L}^+(\gamma)$ is the operator $\phi \mapsto \mathcal{L}^+(\vartheta^+; \gamma, \partial_{\vartheta^+})\phi$ acting from $H_0^1(S^+)$ into $H^{-1}(S^+)$. We denote by $\mathfrak{G}^+$ the set of $\gamma \in \mathbb{C}$ such that $\mathfrak{L}^+(\gamma)$ is not invertible. We call these γ the *corner exponents*. We introduce the analogue of hypothesis $(\mathfrak{H}_1)$ for $\mathfrak{L}^+$:

$$(\mathfrak{H}_3) \qquad \forall \gamma \in \mathfrak{G}^+, \quad \gamma \text{ is a pole of degree 1 of } (\mathfrak{L}^+)^{-1}.$$

For each $\gamma \in \mathfrak{G}^+$, we denote by $\phi[\gamma, q]$, $q = 1, \dots, q_\gamma$, a basis of $\ker \mathfrak{L}^+(\gamma)$.

We need a new family of weighted spaces: Let us introduce r^+ on S^+ as the distance to the corner $(r = 0, z = 0)$ of S^+ corresponding to the edge E and extend it by homogeneity: $r^+(\mathbf{x}) = r^+\big(\vartheta^+(\mathbf{x})\big)$. Note that we have the equivalence

$$r^+(\mathbf{x}) \simeq r(\mathbf{x})/\rho^+(\mathbf{x}). \tag{4.2}$$

In the same way we define $\tilde{r}^+$ on S^+ as the distance to the two other corners of S^+, $(r = 1, \theta = 0, z = 1)$ and $(r = 1, \theta = \omega, z = 1)$, and extend $\tilde{r}^+$ by homogeneity. We

define for $\xi > -\frac{1}{2}$ and $\eta > 0$

$$\mathcal{V}_{\xi,\,\eta}(\Omega^+) := \big\{v \in \mathcal{C}^\infty(\Omega^+) \mid \\ \forall \mathbf{m} \in \mathbb{N}^3,\ (\rho^+)^{-\xi+|\mathbf{m}|}\,(r^+)^{-\eta+|\mathbf{m}|}\,(\tilde{r}^+)^{|\mathbf{m}|}\partial_{\mathbf{x}}^{\mathbf{m}} v \in \mathrm{L}^\infty(\Omega^+)\big\},$$

with $\Omega^+ = G \times (0,1)$. There holds the corner expansion for any fixed $\xi > -\frac{1}{2}$:

$$u - \sum_{\gamma,\ -1/2<\mathrm{Re}\,\gamma<\xi}\ \sum_{q} c_{\gamma,q}\,(\rho^+)^\gamma \phi[\gamma,q](\vartheta^+) \in \mathcal{V}_{\xi,\,0}(\Omega^+), \tag{4.3}$$

where the coefficients $c_{\gamma,q}$ are complex numbers. Note that the remainder in (4.3) is *flat* with respect to the "distance" ρ^+ to the corner $\mathbf{c}^+$ and *not* with respect to the edge E. Thus, the expansions (4.1) and (4.3) give complementary and seemingly independent information about the structure of u.

In fact, we will use this result only to obtain the optimal corner regularity of u without splitting u into regular and singular parts at this corner. We define the set of exponents $\mathfrak{G}^-$ attached to the corner c^- in a similar way as $\mathfrak{G}^+$. We define $\xi_1^\pm$ as

$$\xi_1^\pm = \min\big\{\,\mathrm{Re}\,\gamma \mid \gamma \in \mathfrak{G}^\pm \ \text{ and } \ \mathrm{Re}\,\gamma > -\tfrac{1}{2}\big\}. \tag{4.4}$$

The choice $\xi = \xi_1^+$ is the best possible so that the corner expansion in (4.3) is empty. There holds

$$u \in \mathcal{V}_{\xi_1^+,\,0}(\Omega^+) \quad \text{and} \quad u \in \mathcal{V}_{\xi_1^-,\,0}(\Omega^-). \tag{4.5}$$

4.3. Edge expansion up to the corner. Relying on [14, Chap. 17] we can expand u along the edge E while taking its corner regularity into account. Near $\mathbf{c}^+$ the edge coefficients will themselves belong to weighted spaces of the type $\mathcal{V}_\xi(0,1)$ on the half-edge $\{z \in (0,1)\}$ (here ρ^+ coincides with $1-z$),

$$\mathcal{V}_\xi(0,1) := \big\{a \in \mathcal{C}^\infty(0,1) \mid \ \forall m \in \mathbb{N},\ \ (\rho^+)^{-\xi+m}\,\partial_{\rho^+}^m a \in \mathrm{L}^\infty(0,1)\big\},$$

and near c^- the coefficients will belong to a space $\mathcal{V}_\xi(-1,0)$ where the weight function is $\rho^-(z) = 1+z$ instead of ρ^+.

THEOREM 4.2. *Let $\eta > 0$ be given. Then for any $\alpha \in \mathfrak{A}$ such that $\mathrm{Re}\,\alpha \in (0,\eta)$ and any $p = 1,\dots,p_\alpha$, the coefficient $a_{\alpha,p}$ appearing in the splitting* (4.1) *belongs to $\mathcal{V}_{\xi_1^+-\mathrm{Re}\,\alpha}(0,1)$ and there holds*

$$u - \sum_{\alpha,\ 0<\mathrm{Re}\,\alpha<\eta}\ \sum_{p} \chi(r^+)\,S^n\big[\alpha,p\,;\,a_{\alpha,p}\big] =: u^+_{\mathrm{reg},\eta} \in \mathcal{V}_{\xi_1^+,\,\eta}(\Omega^+), \tag{4.6}$$

where χ is a smooth cut-off function which is 1 in a neighborhood of 0, $r^+ = r^+(\mathbf{x})$ is defined in (4.2)*, and $n = n(\alpha)$ is the smallest integer such that $\mathrm{Re}\,\alpha + n > \eta$. Similarly, $a_{\alpha,p}\big|_{(-1,0)}$ belongs to $\mathcal{V}_{\xi_1^--\mathrm{Re}\,\alpha}(-1,0)$ and there holds*

$$u - \sum_{\alpha,\ 0<\mathrm{Re}\,\alpha<\eta}\ \sum_{p} \chi(r^-)\,S^n\big[\alpha,p\,;\,a_{\alpha,p}\big] =: u^-_{\mathrm{reg},\eta} \in \mathcal{V}_{\xi_1^-,\,\eta}(\Omega^-). \tag{4.7}$$

4.4. Extraction of edge coefficients. Our main goal is the determination and the computation of the edge coefficients $a_{\alpha,p}$—at least those corresponding to the smallest values of $\operatorname{Re}\alpha$. These coefficients are defined via the expansion (4.1), and a sharp estimate of both the coefficients and the remainder is given in Theorem 4.2. The method for extracting them is based on the use of the antisymmetric bilinear form $J[R](u,v)$ defined in (3.14), where v is chosen as $K^n[\beta,p\,;\,b]$ for a certain range of $\beta\in\mathfrak{A}$ and of test edge coefficients b. The choice of the order n will determine the order of the error, which is a positive power of R. We introduce a last technical hypothesis

$$(\mathfrak{H}_4)\qquad \forall\alpha\in\mathfrak{A},\ \operatorname{Re}\alpha\geq 0,\quad \xi_1^+-\operatorname{Re}\alpha\notin\mathbb{N},\quad \xi_1^--\operatorname{Re}\alpha\notin\mathbb{N}.$$

The main result of our work is the following.

THEOREM 4.3. *Let u be the solution of problem* (1.2) *with a smooth right-hand side f, which is zero in a neighborhood of the edge E. We assume the hypotheses* $(\mathfrak{H}_1)$–$(\mathfrak{H}_4)$. *The function u admits the edge expansion* (4.1) *for all $\delta>0$. Let $\beta\in\mathfrak{A}$ with $\operatorname{Re}\beta>0$. We fix an integer $n\geq 0$ such that*

$$n\geq \operatorname{Re}\beta-\xi_1-1\quad\text{with}\quad \xi_1=\min\{\xi_1^+,\xi_1^-\}\,, \tag{4.8}$$

where we recall that ξ_1^+ defined in (4.4) *is attached to the corner $\mathbf{c}^+$ and ξ_1^- is its analogue for the corner $\mathbf{c}^-$. Let m be an integer $m\geq n$ and finally let $b\in\mathcal{C}^m(\overline{I})$ be such that $\partial_z^j b(\pm 1)=0$ for all $j=0,\dots,n-1$. Then there holds*

$$J[R]\big(u,K^m[\beta,p\,;\,b]\big)=\int_I a_{\beta,p}(z)\,\bar b(z)\,\mathrm{d}z\ +\mathcal{O}\left(R^{\min\{n+\xi_1\,,\,m+\eta_1\}-\operatorname{Re}\beta+1}\right) \tag{4.9}$$

as $R\to 0$, where

$$\eta_1=\min\big\{\operatorname{Re}\alpha\mid\alpha\in\mathfrak{A}\ \text{ and }\ \operatorname{Re}\alpha>0\big\}. \tag{4.10}$$

Before starting the proof, we give a corollary of identity (3.21). For this, we first introduce the decomposition of the bilinear form $J[R]$ according to the splitting (3.7) of the radial traction T:

$$J^0[R](u,v):=\int_{\Gamma_R}\big(T_0u\cdot\overline{v}-u\cdot T_0\overline{v}\big)\,R^{-1}\mathrm{d}\sigma=\int_I\int_0^\omega\big(T_0u\cdot\overline{v}-u\cdot T_0\overline{v}\big)\big|_{r=R}\,\mathrm{d}\theta\,\mathrm{d}z$$

and

$$J^1[R](u,v):=\int_{\Gamma_R}\big(T_1u\cdot\overline{v}-u\cdot T_1\overline{v}\big)\,\mathrm{d}\sigma=\int_I\int_0^\omega\big(T_1u\cdot\overline{v}-u\cdot T_1\overline{v}\big)\big|_{r=R}\,R\,\mathrm{d}\theta\,\mathrm{d}z.$$

LEMMA 4.4. *Let $\alpha,\ \beta\in\mathfrak{A}$. Let $m\in\mathbb{N}$ and integers $0\leq n\leq m$, $0\leq k\leq m$. Let $b\in\mathcal{C}^m(\overline{I})$ such that $\partial_z^j b(\pm 1)=0$ for all $j=0,\dots,n-1$. Let $a\in\mathcal{V}_\xi(-1,1)$. If $\xi+n-k+1>0$, then*

$$\sum_{j+\ell=k}J^0[R]\big(\partial_z^j a\,\Phi_j[\alpha,q]\,,\partial_z^\ell b\,\Psi_\ell[\beta,p]\big)+\sum_{j+\ell=k-1}J^1[R]\big(\partial_z^j a\,\Phi_j[\alpha,q]\,,\partial_z^\ell b\,\Psi_\ell[\beta,p]\big)=\delta_{k,0}\,\delta_{\alpha,\beta}\,\delta_{p,q}\int_I a(z)\,\bar b(z)\,\mathrm{d}z.$$

This lemma is merely a consequence of identity (3.21). Indeed, the assumptions about a and b ensure that (i) all integrals in z are convergent, and (ii) integrations by

parts in z (to have all derivatives on b) do not produce any boundary contribution. Therefore we can separate the integrals over I and $(0,\omega)$ as in (3.19). The integrals over $(0,\omega)$ are zero (or 1) thanks to (3.21), which correspondingly yields the lemma.

Proof of Theorem 4.3. Relying on the decompositions (4.6)–(4.7) of u, we split the integral $J[R]\big(u, K^m[\beta,p\,;\,b]\big)$ into several pieces, $I_0 + I_1^+ + I_2^+ + I_1^- + I_2^- + I_3$, and estimate each of them.

- We first assume that $m > n + \xi_1 - \eta_1$.

(A) We define I_0 as

$$I_0 = \sum_{\substack{\alpha,\,q,\,k\\ \xi_1 - \operatorname{Re}\alpha + n - k + 1 > 0}} \left(\sum_{j+\ell=k} J^0[R]\big(\partial_z^j a_{\alpha,q}\,\Phi_j[\alpha,q]\,,\partial_z^\ell b\,\Psi_\ell[\beta,p]\big) + \sum_{j+\ell=k-1} J^1[R]\big(\partial_z^j a_{\alpha,q}\,\Phi_j[\alpha,q]\,,\partial_z^\ell b\,\Psi_\ell[\beta,p]\big)\right),$$

where the coefficients $a_{\alpha,q}$ are those of expansion (4.6). The assumptions of Lemma 4.4 are fulfilled because of the following:

(a) The inequality $\xi_1 - \operatorname{Re}\alpha + n - k + 1 > 0$ implies that $k < \xi_1 - \operatorname{Re}\alpha + n + 1$, which is $\le n + \xi_1 - \eta_1$; since we have assumed that $m > n + \xi_1 - \eta_1$, then $k \le m$.

(b) By Theorem 4.2, $a_{\alpha,q}$ belongs to the weighted space $\mathcal{V}_{\xi_1^+ - \operatorname{Re}\alpha}(0,1)$ in the part of the edge which belongs to Ω^+, and similarly in Ω^-; therefore the inequality $\xi_1 - \operatorname{Re}\alpha + n - k + 1 > 0$ is the assumption $\eta + n - k + 1 > 0$ of Lemma 4.4.

Moreover, the assumption $n \ge \operatorname{Re}\beta - \xi_1 - 1$ implies that the triple $(\alpha = \beta, q = p, k = 0)$ belongs to the sum defining I_0. Therefore

$$I_0 = \int_I a_{\beta,p}(z)\,\bar{b}(z)\,\mathrm{d}z.$$

(B) We define I_1^+ as

$$I_1^+ = \sum_{\substack{\alpha,\,q,\,k\\ \xi_1^+ - \operatorname{Re}\alpha + n - k + 1 > 0}} \left(\sum_{j+\ell=k} J^0[R]\big((\chi(r^+)-1)\partial_z^j a_{\alpha,q}\,\Phi_j[\alpha,q]\,,\partial_z^\ell b\,\Psi_\ell[\beta,p]\big) + \sum_{j+\ell=k-1} J^1[R]\big((\chi(r^+)-1)\partial_z^j a_{\alpha,q}\,\Phi_j[\alpha,q]\,,\partial_z^\ell b\,\Psi_\ell[\beta,p]\big)\right).$$

Let us define z^+ as $1 - z$. The domain of integration of the terms in I_1^+ is $\Gamma_R \cap \operatorname{supp}(\chi(r^+) - 1)$ and is contained in a set of the form

$$\big\{\mathbf{x} \in \mathbb{R}^3 \mid r = R,\ \theta \in (0,\omega),\ z^+ \in (0, cR)\big\},$$

where c is a positive constant.

Each term in I_1^+ can be estimated by a product of three terms:

(i) an integral in z^+ over $(0,cR)$ of a function depending on z^+ but not on R or θ;

(ii) an integral in θ over $(0,\omega)$ of a function depending on θ but not on R or z^+;

(iii) a power of R corresponding to the restriction on Γ_R of a power of r.

We observe the following:

(i) The integral over $(0, cR)$ is $\int_0^{cR} (z^+)^{\xi_+^1 - \operatorname{Re}\alpha + n - k}\, \mathrm{d}z^+$, which is $\mathcal{O}\big(R^{\xi_+^1 - \operatorname{Re}\alpha + n - k + 1}\big)$ since $\xi_+^1 - \operatorname{Re}\alpha + n - k + 1 > 0$.

(ii) The integral over $(0, \omega)$ does not depend on R.

(iii) The power of R is $R^{\operatorname{Re}\alpha - \operatorname{Re}\beta + k}$.

Therefore

$$I_1^+ = \mathcal{O}\left(R^{\xi_+^1 + n - \operatorname{Re}\beta + 1}\right).$$

The corresponding part I_1^- in the neighborhood of $\mathbf{c}^-$ has a similar bound.

(C) We define I_2^+ as

$$I_2^+ = \sum_{\substack{\alpha,\, q,\, j,\, \ell \\ \xi_1^+ - \operatorname{Re}\alpha + n - j - \ell + 1 < 0 \\ \ell \le m,\ \operatorname{Re}\alpha + j < n + \xi_1^+ + 1}} \left(\sum_{j+\ell=k} J_+^0[R]\big(\chi(r^+)\, \partial_z^j a_{\alpha,q}\, \Phi_j[\alpha, q]\ , \partial_z^\ell b\, \Psi_\ell[\beta, p]\big) \right.$$
$$\left. + \sum_{j+\ell=k-1} J_+^1[R]\big(\chi(r^+)\, \partial_z^j a_{\alpha,q}\, \Phi_j[\alpha, q]\ , \partial_z^\ell b\, \Psi_\ell[\beta, p]\big) \right),$$

where J_+^0 and J_+^1 are the contributions over Ω^+ of J^0 and J^1.

As for I_1^+, each term of I_2^+ can be estimated by the product of three terms (i)–(iii). The only difference is that the integral (i) in z^+ is over $(cR, 1)$ instead of $(0, cR)$ and is equal to $\int_{cR}^1 (z^+)^{\xi_+^1 - \operatorname{Re}\alpha + n - k}\, \mathrm{d}z^+$, which is still $\mathcal{O}\big(R^{\xi_+^1 - \operatorname{Re}\alpha + n - k + 1}\big)$ since $\xi_+^1 - \operatorname{Re}\alpha + n - k + 1$ is < 0. The power (iii) of R is the same; thus we obtain, as above, that

$$I_2^+ = \mathcal{O}\left(R^{\xi_+^1 + n - \operatorname{Re}\beta + 1}\right).$$

(D) We set $\eta := n + \xi_1^+ + 1$. We check that

$$I_0 + I_1^+ + I_1^- + I_2^+ + I_2^- = \sum_{\substack{\alpha,\, q,\, j \\ \operatorname{Re}\alpha + j < \eta}} J[R]\big(\chi(r^+)\, \partial_z^j a_{\alpha,q}\, \Phi_j[\alpha, q]\ , K^m[\beta, p\,;\, b]\big).$$

But according to Theorem 4.2

$$u^+_{\mathrm{reg},\eta} := u - \sum_{\substack{\alpha,\, q,\, j \\ \operatorname{Re}\alpha + j < \eta}} \chi(r^+)\, \partial_z^j a_{\alpha,q}\, \Phi_j[\alpha, q] \in \mathcal{V}_{\xi_1^+,\,\eta}(\Omega^+),$$

and similarly for the other corner. Therefore it remains to estimate

$$I_3 := J[R]\big(u^+_{\mathrm{reg},\eta}\ , K^m[\beta, p\,;\, b]\big)$$

and, more precisely, each contribution $J[R]\big(u^+_{\mathrm{reg},\eta}, \partial_z^\ell b\, \Psi_\ell\big)$ for $\ell = 0, \dots, m$. Since $u^+_{\mathrm{reg},\eta}$ belongs to $\mathcal{V}_{\xi_1^+,\,\eta}(\Omega^+)$,

$$u^+_{\mathrm{reg},\eta} = \mathcal{O}\left((\rho^+)^{\xi_1^+} (r^+)^\eta\right) = \mathcal{O}\left((\rho^+)^{\xi_1^+ - \eta}\, r^\eta\right) \quad \text{and} \quad \nabla u^+_{\mathrm{reg},\eta} = \mathcal{O}\left((\rho^+)^{\xi_1^+ - \eta}\, r^{\eta - 1}\right).$$

For the bounding of $J[R]\big(u^+_{\mathrm{reg},\eta}\,, \partial_z^\ell b\, \Psi_\ell\big)$, we split the integral over Γ_R into (a) the contribution on $z^+ \in (0, R)$, and (b) the contribution on $z^+ \in (R, 1)$, and we estimate each piece by a product of three terms as we did before.

(a) When $z^+ \in (0,R)$, the distance ρ^+ is equivalent to R on Γ_R. Therefore the weight over $u^+_{\mathrm{reg},\eta}$ is equivalent to $R^{\xi_1^+}$ in that region. Part (i) is the integral $\int_0^R (z^+)^{n-\ell}\,\mathrm{d}z^+ = \mathcal{O}\big(R^{n-\ell+1}\big)$, and the power (iii) of R is $R^{\xi_1^+ - \mathrm{Re}\,\beta+\ell}$. Their product is $R^{n+\xi_1^+-\mathrm{Re}\,\beta+1}$.

(b) When $z^+ \in (R,1)$, the distance ρ^+ is equivalent to z^+ on Γ_R. Therefore the weight over $u^+_{\mathrm{reg},\eta}$ is equivalent to $(z^+)^{\xi_1^+-\eta}\,r^\eta$ in that region. Part (i) is the integral $\int_R^1 (z^+)^{\xi_1^+-\eta+n-\ell}\,\mathrm{d}z^+ = \mathcal{O}\big(R^{\xi_1^+-\eta+n-\ell+1}\big)$ (since $\xi_1^+ - \eta + n - \ell + 1 < 0$), and the power (iii) of R is $R^{\eta-\mathrm{Re}\,\beta+\ell}$. The product of both is $R^{\xi_1^+ + n + 1 - \mathrm{Re}\,\beta}$.

Gathering all the previous results of parts (A)–(D), we obtain formula (4.9) in the case $m > n + \xi_1 - \eta_1$.

• When $m < n+\xi_1-\eta_1$, we follow the same lines with the corresponding changes: For I_0 we reduce the sum by the extra condition that $k \le m$, and the same for $I_1^\pm$. The conclusions are still the same. For I_2^+ the sum is augmented by the set of (α, q, j, ℓ) such that $\xi_1^+ - \mathrm{Re}\,\alpha + n - j - \ell + 1 > 0$ and $j + \ell > m$. The new terms do not satisfy the same estimates as the old ones since the corresponding contribution (i) in z^+ is now $\mathcal{O}(1)$. As the power (iii) of R is still $R^{\mathrm{Re}\,\alpha - \mathrm{Re}\,\beta + j + \ell}$, we obtain that

$$I_2^+ = \min\Big\{\mathcal{O}\big(R^{\xi_1^1+n-\mathrm{Re}\,\beta+1}\big), \mathcal{O}\left(R^{\mathrm{Re}\,\alpha-\mathrm{Re}\,\beta+j+\ell}\right)\Big\},$$

where the min is taken over (α, j, ℓ) such that $\xi_1^+ - \mathrm{Re}\,\alpha + n - j - \ell + 1 > 0$ and $j + \ell > m$. The minimum of $\mathrm{Re}\,\alpha + j + \ell$ is attained for $\alpha = \beta_1$ and $j + \ell = m + 1$, whence

$$I_2^+ = \mathcal{O}\left(R^{\eta_1 - \mathrm{Re}\,\beta + m + 1}\right).$$

We have proved formula (4.9) in the case $m < n + \xi_1 - \eta_1$. □

Remark 4.5. (i) Formula (4.9) is, of course, still valid if hypotheses $(\mathfrak{H}_1)$–$(\mathfrak{H}_4)$ are only assumed to hold for the exponents which are used in the proof, namely, $\mathrm{Re}\,\beta < \eta = n + \xi_1 + 1$ for $(\mathfrak{H}_1), (\mathfrak{H}_2)$, and $(\mathfrak{H}_4)$ and $\mathrm{Re}\,\gamma = \xi_1^+$ for $(\mathfrak{H}_3)$.

(ii) If we discard hypotheses $(\mathfrak{H}_3)$ and $(\mathfrak{H}_4)$, we can still prove a formula like (4.9), up to the possible multiplication of the remainder by $|\log^M R|$ for some integer M.

(iii) We still obtain formula (4.9) if we relax the assumption on the right-hand side so that f is no longer supposed to be zero in the neighborhood of the edge, but only flat up to a specified order, in relation to what is needed in the proof of (4.9): it suffices that f belongs to the weighted spaces $\mathcal{V}_{\xi_1-2\,,\,\eta-2}(\Omega^+)$ and $\mathcal{V}_{\xi_1-2\,,\,\eta-2}(\Omega^-)$, with ξ_1 defined in (4.8) and $\eta = n + \xi_1 + 1$. Then the edge expansion up to the corner (4.6) still holds with such a right-hand side, which makes part (D) of the proof of (4.9) still valid.

Remark 4.6. The assumptions about the test edge coefficients b can be slightly relaxed.

(i) Instead of the boundary conditions $\partial_z^j b(\pm 1) = 0$ for any $j = 0, \dots, n-1$, we may assume that $(1-z)^{-n+j}(z+1)^{-n+j}\partial_z^j b \in \mathrm{L}^\infty(I)$ for $j \le m$, and the statement of Theorem 4.3 can be extended to *noninteger* n.

(ii) We may assume that b is only $\mathcal{C}^{m-1}(\overline{I})$ globally and piecewise $\mathcal{C}^m$ on a finite partition of I.

5. A wider range of applications for quasi-dual methods. We extend the results of Theorem 4.3 to any edge of a general polyhedron and discuss the case of cracks (where $\omega = 2\pi$). We also evaluate the limitation of the convergence rate in R when the right-hand side is not flat along the edge.

5.1. The domain. By a slight modification we can adapt our method to the determination of edge singularities along any edge of a three-dimensional polyhedron, that is, a domain Ω with plane faces and, therefore, straight edges.

Let E be an edge of Ω. E is an open segment whose end points $\mathbf{c}^+$ and $\mathbf{c}^-$ are corners of Ω. We choose cylindrical coordinates (r,θ,z) adapted to Ω around E:

$$E = \big\{\mathbf{x} \sim (r,\theta,z) \mid \ r = 0,\ z \in (-\tfrac{h}{2}, \tfrac{h}{2})\big\},$$

where h is the length of E. There exists a conical neighborhood[1] Θ of E such that

$$\Omega \cap \Theta = \big\{\mathbf{x} \sim (r,\theta,z) \mid \ r = (0,1),\ \omega \in (0,\omega),\ z \in (-\tfrac{h}{2}, \tfrac{h}{2})\big\} \cap \Theta,$$

where ω is the opening of Ω along the edge E.

We still define, for any $R < 1$, the internal cylinder Γ_R as

$$\Gamma_R = \big\{\mathbf{x} \sim (r,\theta,z) \mid \ r = R,\ \omega \in (0,\omega),\ z \in (-\tfrac{h}{2}, \tfrac{h}{2})\big\}.$$

But it may happen that even for small R, Γ_R is not included in Ω. Then we define the reduced internal cylinder $\breve{\Gamma}_R$ as

$$\breve{\Gamma}_R = \big\{\mathbf{x} \sim (r,\theta,z) \mid \ r = R,\ \omega \in (0,\omega),\ z \in (-\tfrac{h}{2} + kR, \tfrac{h}{2} - kR)\big\},$$

where $k > 0$ defines the conical neighborhood Θ. In other words, for any $R \le R_0$, $\breve{\Gamma}_R = \Gamma_R \cap \Theta$.

On the same model as (3.14), we define

$$\breve{J}[R](u,v) := \int_{\breve{\Gamma}_R} \big(Tu\cdot\overline{v} - u\cdot T\overline{v}\big)\,\mathrm{d}\sigma = \int_{-\frac{h}{2}+kR}^{\frac{h}{2}-kR} \int_0^\omega \big(Tu\cdot\overline{v} - u\cdot T\overline{v}\big)\big|_{r=R}\, R\,\mathrm{d}\theta\,\mathrm{d}z. \tag{5.1}$$

Then defining the sets $\mathfrak{G}^\pm$ of corner exponents at $\mathbf{c}^\pm$ as before, but now on the polyhedral cones $K^\pm$ coinciding with Ω in neighborhoods of $\mathbf{c}^\pm$, and defining $\xi_1^\pm$ in the same way, we have expansions[2] (4.6)–(4.7), and there holds, with the same assumptions as in Theorem 4.3,

$$\breve{J}[R]\big(u, K^m[\beta,p\,;b]\big) = \int_{-\frac{h}{2}}^{\frac{h}{2}} a_{\beta,p}(z)\,\bar{b}(z)\,\mathrm{d}z \ + \mathcal{O}\left(R^{\min\{n+\xi_1\,,\,m+\eta_1\}-\mathrm{Re}\,\beta+1}\right). \tag{5.2}$$

The proof follows exactly the same steps as the proof of (4.9). The parts I_0, $I_1^\pm$, and $I_2^\pm$ are still defined by integrals over Γ_R. We modify only part (D), noting that, thanks to the condition on the support of χ, the expansion (4.6) now gives

$$\breve{J}[R](u, K^m[\beta,p\,;b]) = \breve{J}[R](u^+_{\mathrm{reg},\eta}, K^m[\beta,p\,;b]) + I_0 + I_1^+ + I_1^- + I_2^+ + I_2^-.$$

The conclusion follows by the same arguments as before.

[1] In cylindrical coordinates, Θ has the form

$$\Theta = \{\mathbf{x} \sim (r,\theta,z) \mid \ r = (0,R_0),\ \omega \in (0,\omega),\ z \in (-\tfrac{h}{2} + kr, \tfrac{h}{2} - kr)\},$$

with a $k > 0$ and $R_0 > 0$.

[2] With the cut-off function χ chosen so that in the cylinder $r \le R_0$, the support of $\mathbf{x} \mapsto \chi(r^\pm)$ is contained in the conical neighborhood Θ. The subdomains Ω^+ and Ω^- correspond to the regions $z > 0$ and $z < 0$, respectively.

5.2. In the presence of cracks. We now consider the case where the opening ω is equal to 2π. This means that the model domain Ω is the cylinder of radius 1 with an internal boundary formed by the plane rectangle

$$\Sigma = \{\mathbf{x} \in \mathbb{R}^3 \mid x \in (0,1),\ y = 0,\ z \in I\}.$$

This case is in principle included in our analysis. But the special situation of the singularity exponents prevents hypothesis $(\mathfrak{H}_2)$ from being satisfied: By the result of [10], the set $\mathfrak{A}$ of singular exponents is included in the set of half-integers and, moreover,

$$\forall \bar{j} \in \mathbb{N}, \quad \dim\ker \mathfrak{M}_0(\tfrac{1}{2} + \bar{j}) = N, \tag{5.3}$$

where we recall that N is the size of the system L. But our method can still be applied in this case! We are going to explain why.

The first place where we use $(\mathfrak{H}_2)$ is for the definition of the shadow singularities $\Phi_j[\alpha, p]$. The general theory gives that $\Phi_j[\alpha, p]$ can be found in the form of a finite sum of the form $r^{\alpha+j} \sum \log^q r\, \varphi_{j,q}(\theta)$. But in this situation of cracks, it is proved in [11] that the logarithmic terms are absent. But still, the solution of (2.8), though existing, is not unique. This circumstance will help in the second place where we use $(\mathfrak{H}_2)$.

We used $(\mathfrak{H}_2)$ to prove (3.21), in particular that $H_k[\alpha, p\,; \beta, q] = 0$ for all α and β in $\mathfrak{A}$ when $k \neq 0$.

LEMMA 5.1. *For all $\bar{j} \in \mathbb{N}$, $p = 1, \dots, N$, and $j \geq 1$ let the singularities $\Phi_0[\frac{1}{2} + \bar{j}, p]$ and their shadows $\Phi_j[\frac{1}{2} + \bar{j}, p]$ be fixed. The dual singularities $\Psi_0[\frac{1}{2} + \bar{\ell}, q]$ are still determined according to Lemma* 3.2*, and there exists a choice of the shadows $\Psi_\ell[\frac{1}{2} + \bar{\ell}, q]$ such that there holds (cf.* (3.20) *and* (3.21)*)*

$$\forall \bar{j},\ \bar{\ell} \in \mathbb{N}, \quad \forall p, q \leq N, \quad \forall k > 0, \quad H_k[\tfrac{1}{2} + \bar{j}, p\,; \tfrac{1}{2} + \bar{\ell}, q] = 0. \tag{5.4}$$

Proof. By the proof of Proposition 3.4, we know that for any choice of the $\Psi_\ell[\beta, q]$, the identity $H_k[\alpha, p\,; \beta, q] = 0$ holds as soon as $\alpha - \beta + k \neq 0$, i.e., in our case, when $\frac{1}{2} + \bar{j} - \frac{1}{2} - \bar{\ell} + k \neq 0$. Thus it remains to prove (5.4) when $\bar{j} - \bar{\ell} + k = 0$.

Let $\bar{\ell}$ and q be fixed. The proof uses induction over k. For $k = 1$, $\bar{j} = \bar{\ell} - 1$. Let us fix a particular solution $\breve{\psi}_1[\frac{1}{2} + \bar{\ell}, q]$ of (3.12). Any solution of (3.12) is the sum of $\breve{\psi}_1[\frac{1}{2} + \bar{\ell}, q]$ and of an element of $\ker \mathfrak{M}_0(-\frac{1}{2} - \bar{\ell} + 1) = \ker \mathfrak{M}_0(-\frac{1}{2} - \bar{j})$. A basis of this kernel is the set of $\psi_0[\frac{1}{2} + \bar{j}, p']$, $p' = 1, \dots, N$. Therefore $H_1[\frac{1}{2} + \bar{j}, p\,; \frac{1}{2} + \bar{\ell}, q]$ is the sum of a fixed contribution and of a linear combination of the contributions of the $\psi_0[\frac{1}{2} + \bar{j}, p']$, i.e., of $H_0[\frac{1}{2} + \bar{j}, p\,; \frac{1}{2} + \bar{j}, p']$. By Lemma 3.2, we can determine elements of the kernel $\ker \mathfrak{M}_0(-\frac{1}{2} - \bar{j})$ so that $H_1[\frac{1}{2} + \bar{j}, p\,; \frac{1}{2} + \bar{\ell}, q] = 0$ for all $p = 1, \dots, N$.

For a general k, we assume that the $\Psi_\ell[\frac{1}{2} + \bar{\ell}, q]$ are determined for $\ell < k$ and have to prove (5.4) for $\bar{j} = \bar{\ell} - k$. We isolate the contribution $j = 0$, $\ell = k$ in H_k, and the proof is similar to the case $k = 1$. ∎

5.3. The right-hand side. Let us consider now a standard smooth right-hand side $f \in \mathcal{C}^\infty(\overline{\Omega})$. Then f belongs to the weighted spaces $\mathcal{V}_{0,0}(\Omega^+)$ and $\mathcal{V}_{0,0}(\Omega^-)$. With

$$\xi_0^+ = \min\{\xi_1^+, 2\} \quad \text{and} \quad \xi_0^- = \min\{\xi_1^-, 2\}, \tag{5.5}$$

there holds, for $\eta = 2$,

$$f \in \mathcal{V}_{\xi_0^+ - 2\,,\,\eta - 2}(\Omega^+) \quad \text{and} \quad f \in \mathcal{V}_{\xi_0^- - 2\,,\,\eta - 2}(\Omega^-). \tag{5.6}$$

Thus a general smooth interior right-hand side alters the asymptotics of the solution only in the region of exponents $\operatorname{Re}\alpha \geq 2$ and $\operatorname{Re}\gamma \geq 2$. The corresponding parts in the asymptotics of u (either polynomial or singular) are no longer orthogonal in the sense of the bilinear form $J[R]$ versus the standard singularities associated with a zero (or flat) right-hand side.

In connection with Remark 4.5(iii), we see that in order to take (5.6) into account, we first have to replace ξ_1 by $\xi_0 := \min\{\xi_0^+, \xi_0^-\}$ in the statement of Theorem 4.3 and investigate the consequences on the estimates of the limitation $\eta = 2$.

We assume that $m > n + \xi_0 - \eta_1$. We make changes in the general proof of Theorem 4.3 in the same spirit as at the end of this proof: For I_0 we reduce the sum by the extra condition that $\operatorname{Re}\beta + k < 2$, and the same for $I_1^{\pm}$. Thus we require that $\operatorname{Re}\alpha < 2$ so that the triple $(\beta = \alpha, q = p, k = 0)$ belongs to the sum defining I_0. The conclusions are still the same.

For I_2^+ the sum is augmented by the set of (β, q, j, ℓ) such that $\xi_1^+ - \operatorname{Re}\beta + n - j - \ell + 1 > 0$ and $\operatorname{Re}\beta + j + \ell \geq 2$. The new terms do not satisfy the same estimates as the old ones since the corresponding contribution (i) in z^+ is now $\mathcal{O}(1)$. As the power (iii) of R is still $R^{\operatorname{Re}\beta - \operatorname{Re}\alpha + j + \ell}$, we obtain

$$I_2^+ = \min\Big\{\mathcal{O}\big(R^{\xi_+^1 + n - \operatorname{Re}\alpha + 1}\big), \mathcal{O}\left(R^{\operatorname{Re}\beta - \operatorname{Re}\alpha + j + \ell}\right)\Big\},$$

where the min is taken over (β, j, ℓ) such that $\xi_1^+ - \operatorname{Re}\beta + n - j - \ell + 1 > 0$ and $\operatorname{Re}\beta + j + \ell \geq 2$.

We have also to consider I_3 anew with the constraint that $\eta = 2$. Part (a) of the estimate is the same, but concerning part (b), we now have to deal with the possibility that $\xi_0^+ - \eta + n + 1 = \xi_0^+ - 2 + n + 1$ may be ≥ 0. In this case, the contribution (i) is $\mathcal{O}(1)$ and the contribution (iii) is $R^{\eta - \operatorname{Re}\alpha} = R^{2 - \operatorname{Re}\alpha}$.

Let $Q[R](u, K^m[\alpha, p; b])$ be the remainder $J[R](u, K^m[\alpha, p; b]) - \int_I a_{\alpha,p}(z)\,\bar{b}(z)\,\mathrm{d}z$.

THEOREM 5.2. *Let u be the solution of problem (1.2) with a smooth right-hand side $f \in \mathcal{C}^\infty(\overline{\Omega})$. We assume the hypotheses $(\mathfrak{H}_1)$–$(\mathfrak{H}_4)$. Let $\alpha \in \mathfrak{A}$ with $\operatorname{Re}\alpha \in (0, 2)$. We fix an integer $n \geq 0$ such that*

$$n \geq \operatorname{Re}\alpha - \xi_0 - 1. \tag{5.7}$$

Let m be an integer $m \geq n$ and let $b \in \mathcal{C}^m(\overline{I})$ be such that $\partial_z^j b(\pm 1) = 0$ for all $j = 0, \ldots, n-1$. Then there holds

$$Q[R]\big(u, K^m[\alpha, p\,;\, b]\big) \;=\; \mathcal{O}\left(R^{\min\{1\,,\,n+\xi_1\,,\,m+\eta_1\} - \operatorname{Re}\alpha + 1}\right). \tag{5.8}$$

Remark 5.3. If f is zero on the edge E, then f belongs to $\mathcal{V}_{1,1}(\Omega^{\pm})$ and the above statement can be improved by replacing everywhere 2 by 3, including in the definition (5.5) of $\xi_0^{\pm}$, and we obtain the following estimate for the remainder:

$$Q[R]\big(u, K^m[\alpha, p\,;\, b]\big) \;=\; \mathcal{O}\left(R^{\min\{2\,,\,n+\xi_1\,,\,m+\eta_1\} - \operatorname{Re}\alpha + 1}\right). \tag{5.9}$$

5.4. Other boundary conditions. In a way similar to that described in detail for Dirichlet boundary conditions, we can treat other self-adjoint boundary conditions such as Neumann conditions or mixed conditions in several forms, i.e., Dirichlet on certain faces and Neumann on the others, or of mixed type for systems, where, for example, in elasticity some components of the displacement are prescribed to 0 and the complementing components of the traction are also prescribed.

We may also consider transmission conditions based on a coercive bilinear form B with piecewise constant coefficients.

Once the correct Mellin symbols $\mathfrak{M}_0$ and $\mathfrak{L}^{\pm}$ are defined, we consider their respective spectra $\mathfrak{A}$ and $\mathfrak{G}^{\pm}$ and everything works in the same way, mutatis mutandis. But we have to emphasize that the sets of exponents $\mathfrak{A}$ and $\mathfrak{G}^{\pm}$ may systematically contain (small) integers. For example, if we consider a Neumann problem, 0 always belongs to $\mathfrak{A}$ and $\mathfrak{G}^{\pm}$, which implies that $\alpha_1 = 0$ (and, in general, $\xi_1 = 0$), though this zero exponent corresponds to a "singular function" Φ_0 which is constant.

Also the consideration of nonzero boundary data in the neighborhood of the edge would introduce more perturbation in the orthogonality relations between the asymptotics of the solution and the standard singularities associated with a zero right-hand side.

6. Other methods and formulas: A comparison. Inspired by [26] and [20] we can provide other families of formulas for the determination of the edge coefficients. We present them and then compare them with each other. All of them are valid in the extended framework of polyhedral domains as in section 5.1.

6.1. Pointwise dual formulas. Adapting [26] we find the formula, valid for any solution u of (1.2) with smooth $Lu = f$, sufficiently flat near the edge E: For each fixed $z_0 \in I$,

$$a_{\alpha,p}(z_0) = \int_\Omega Lu \cdot \overline{K}_{z_0}[\alpha, p]\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z. \tag{6.1}$$

The three-dimensional dual function $(x, y, z) \mapsto K_{z_0}[\alpha, p](x, y, z)$ is defined as

$$K_{z_0}[\alpha, p] := \Psi^{3\mathrm{D}}_{z_0}[\alpha, p] - X_{z_0}[\alpha, p],$$

where the following hold:

1. $\Psi^{3\mathrm{D}}_{z_0}[\alpha, p]$ is a dual three-dimensional "*corner*" singularity at $(0, 0, z_0)$ considered the vertex of a cone: With ρ_0 the distance to the point $(0, 0, z_0)$, and ϑ_0 the corresponding spherical coordinates, $\Psi^{3\mathrm{D}}_{z_0}[\alpha, p]$ has the form

$$\Psi^{3\mathrm{D}}_{z_0}[\alpha, p](\rho_0, \vartheta_0) = \rho_0^{-1-\bar{\alpha}}\, \psi[\alpha, p](\vartheta_0)$$

and satisfies on the infinite wedge W_I coinciding with Ω in the conical neighborhood Θ

$$\begin{cases} L\Psi^{3\mathrm{D}}_{z_0}[\alpha, p] = 0 & \text{in} \quad W_I, \\ \Psi^{3\mathrm{D}}_{z_0}[\alpha, p] = 0 & \text{on} \quad \partial W_I. \end{cases}$$

It does not belong to H^1 in any neighborhood of z_0 due to its strong singularity in $\rho_0^{-1-\bar{\alpha}}$. The spherical pattern ψ depends only on the wedge W_I and the operator L, but not on the particular point z_0 since we have supposed that the operator has constant coefficients.

2. $X_{z_0}[\alpha, p]$ is the correction in $H^1(G)$, solution of

$$\begin{cases} LX_{z_0}[\alpha, p] = 0 & \text{in} \quad \Omega, \\ X_{z_0}[\alpha, p] = \Psi^{3\mathrm{D}}_{z_0}[\alpha, p]\big|_{\partial\Omega} & \text{on} \quad \partial\Omega. \end{cases} \tag{6.2}$$

Note that X_{z_0} strongly depends on z_0, because the trace of $\Psi^{3\mathrm{D}}_{z_0}[\alpha, p]$ on $\partial\Omega$ depends on z_0.

6.2. Global dual formulas. In the same spirit as formulas (6.1)–(6.2), we can also obtain exact formulas for moments of the coefficients: For test functions $b \in \mathcal{C}_0^\infty(I)$ (or, more generally, b as in Theorem 4.3 with n large enough)

$$\int_{-1}^{1} a_{\alpha,p}(z)\, b(z)\, \mathrm{d}z = \int_\Omega Lu \cdot \overline{K}_b[\alpha,p]\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z. \tag{6.3}$$

Here $K_b[\alpha,p] := K^m[\alpha,p;b] - X_b[\alpha,p]$, where $K^m[\alpha,p;b]$ is defined in (3.11) with $m > \operatorname{Re}\alpha - 1$ (i.e., so that $LK^m[\alpha,p;b]$ belongs to $H^{-1}(\Omega)$; see (3.13)) and $X_b[\alpha,p]$ is the correction in $H^1(G)$, solution of

$$\left\{ \begin{array}{rll} LX_b[\alpha,p] = LK^m[\alpha,p;b] & \text{in} & \Omega, \\ X_b[\alpha,p] = K^m[\alpha,p;b]\big|_{\partial\Omega} & \text{on} & \partial\Omega. \end{array} \right. \tag{6.4}$$

Compare with [20], where the case $L = \Delta$ with $m = 0$ is considered.

An alternative to (6.3) in the spirit of [15] is the following mixed formula:

$$\int_{-1}^{1} a_{\alpha,p}(z)\, b(z)\, \mathrm{d}z = \int_\Omega Lu \cdot \chi\overline{K}^m[\alpha,p;b] - u \cdot L(\chi\overline{K}^m[\alpha,p;b])\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z. \tag{6.5}$$

Here the cut-off χ can be taken as in the expansions (4.6)–(4.7), i.e., $\chi(\mathbf{x}) = \chi(r^+)$ in Ω^+ and $\chi(\mathbf{x}) = \chi(r^-)$ in Ω^-. Simpler cut-off can be used if Ω contains a cylinder of the form $\{\mathbf{x},\ r < r_0,\ 0 < \theta < \omega,\ z \in I\}$: then $\chi = \chi(r)$ with $\chi(r) \equiv 1$ for $r < r_0/2$ and $\equiv 0$ for $r \geq r_0$.

6.3. Comparison. Formula (6.1) yields exact pointwise values for the edge coefficient, provided the right-hand side is smooth enough to ensure the continuity of the coefficient and flat enough to cancel any Taylor part of degree $\leq \operatorname{Re}\alpha$ in the solution u. This formula makes use of the right-hand side only and does not need the computation of u. But its main drawback is its own computation. The determination of the dual spherical pattern $\psi[\alpha,p]$ is seldom explicit and difficult in general: In addition to the Laplace operator, this is done only for the Lamé system under Neumann boundary conditions for a crack situation ($\omega = 2\pi$); see [30]. Moreover the solution of the three-dimensional problem (6.2) is necessary for each value of z_0 where we want to have the value of the coefficient $a_{\alpha,p}$. Finally, the application of formula (6.1) requires the computation of a volume integral.

Formula (6.3) yields exact evaluation of the moment of the coefficient against the test function. It has the following advantages over (6.1): the continuity of the coefficients is no longer necessary; the basic function $K^m[\alpha,p;b]$ is easier to determine (one-dimensional problems on $(0,\omega)$) and less singular than $\Psi_{z_0}^{3\mathrm{D}}$. But it is still necessary to solve as many three-dimensional problems (6.4) as values of test functions b.

Formula (6.5) is closer to the idea of the quasi-dual formulas, since it is no longer necessary to solve three-dimensional problems for the determination of the dual functionals, but it does require the knowledge of the solution u. Still (6.5) is a volume integral, and the determination of the cut-off terms $\chi\overline{K}^m[\alpha,p;b]$ and $L(\chi\overline{K}^m[\alpha,p;b])$ is not obvious.

The quasi-dual formulas (4.9) and (5.2) need the determination of the same basic functions $K^m[\alpha,p;b]$ and the computation of the solution u itself, but no other three-dimensional solution. It requires only one (or a few) surface integrals, away from the edge where the functions $K^m[\alpha,p;b]$ are the most singular. Each determination of

$J[R]\big(u, K^m[\beta, p\,;\, b]\big)$ does not provide the exact value of the moment of $a_{\alpha,p}$ against b but its value modulo a (known) power of R, which allows a Richardson extrapolation of the limit from the computation of $J[R]\big(u, K^m[\beta, p\,;\, b]\big)$ for 3 values of R.

The works [34] in two dimensions and [36] in three dimensions also introduce an extraction method based on integration over a circular arc of radius R, followed by Richardson extrapolation in R. They are successfully implemented in an engineering stress analysis code. In a certain sense, they are precursory to our present method, with the following important distinction: In these two references the antisymmetric duality pairing $J[R]$ is replaced by a simple scalar product involving only the angular part of the singular functions. This possibility exists only for the Laplace operator due to its natural separation of variables (see [36]) and for the Lamé equations in two dimensions (see [34]). In order to reach a wide generality, we are led to deal with the universal duality pairing $J[R]$. On the other hand, the extraction done in [36] yields *pointwise* values of the coefficients. Extracting *moments* is more suitable to the regularity properties of the edge coefficients near corners and to the approximation by finite elements.

REFERENCES

[1] B. Andersson, U. Falk, and I. Babuška, *Reliable determination of edge and vertex stress intensity factors in three-dimensional elastomechanics*, in Proceedings of the 17th Congress of the International Council of the Aeronautical Sciences, Stockholm, 1990, ICAS-90-4.9.2, AIAA, Washington, D.C., 1990, pp. 1730–1746.

[2] I. Babuška and A. Miller, *The post-processing approach in the finite element method. Part* II: *The calculation of the stress intensity factors*, Internat. J. Numer. Methods Engrg., 20 (1984), pp. 1111–1129.

[3] A. Beagles and A.-M. Sändig, *Singularities of rotationally symmetric solutions of boundary value problems for the Lamé equations*, ZAMM Z. Angew. Math. Mech., 71 (1991), pp. 423–431.

[4] H. Blum, *On the approximation of linear elliptic systems on polygonal domains*, in Singularities and Constructive Methods for Their Treatment (Oberwolfach, 1983), Lecture Notes in Math. 1121, Springer-Verlag, Berlin, pp. 28–37.

[5] H. Blum, *Numerical treatment of corner and crack singularities*, in Finite Element and Boundary Element Techniques from Mathematical and Engineering Point of View, CISM Courses and Lectures 301, Springer-Verlag, Vienna, 1988, pp. 171–212.

[6] H. Blum and M. Dobrowolski, *On finite element methods for elliptic equations on domains with corners*, Computing, 28 (1982), pp. 53–63.

[7] M. Bourlard, M. Dauge, M.-S. Lubuma, and S. Nicaise, *Coefficients of the singularities for elliptic boundary value problems on domains with conical points.* III. *Finite element methods on polygonal domains*, SIAM J. Numer. Anal., 29 (1992), pp. 136–155.

[8] M. Bourlard, M. Dauge, and S. Nicaise, *Error estimates on the coefficients obtained by the singular function method*, Numer. Funct. Anal. Optim., 10 (1989), pp. 1077–1113.

[9] M. Costabel and M. Dauge, *General edge asymptotics of solutions of second order elliptic boundary value problems.* I *and* II, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 109–184.

[10] M. Costabel and M. Dauge, *Crack singularities for general elliptic systems*, Math. Nachr., 235 (2002), pp. 29–49.

[11] M. Costabel, M. Dauge, and R. Duduchava, *Asymptotics without logarithmic terms for crack problems*, Comm. Partial Differential Equations, 28 (2003), pp. 869–926.

[12] M. Costabel, M. Dauge, and Y. Lafranche, *Fast semi-analytic computation of elastic edge singularities*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 2111–2134.

[13] M. Costabel and E. P. Stephan, *An improved boundary element Galerkin method for three-dimensional crack problems*, Integral Equations Operator Theory, 10 (1987), pp. 467–504.

[14] M. Dauge, *Elliptic Boundary Value Problems in Corner Domains. Smoothness and Asymptotics of Solutions*, Lecture Notes in Math. 1341, Springer-Verlag, Berlin, 1988.

[15] M. Dauge, S. Nicaise, M. Bourlard, and J. M.-S. Lubuma, *Coefficients des singularités pour des problèmes aux limites elliptiques sur un domaine à points coniques.* I. *Résultats*

généraux pour le problème de Dirichlet, RAIRO Modél. Math. Anal. Numér., 24 (1990), pp. 27–52.

[16] M. Dauge, S. Nicaise, M. Bourlard, and J. M.-S. Lubuma, *Coefficients des singularités pour des problèmes aux limites elliptiques sur un domaine à points coniques.* II. *Quelques opérateurs particuliers*, RAIRO Modél. Math. Anal. Numér., 24 (1990), pp. 343–367.

[17] P. Destuynder and M. Djaoua, *Estimation de l'erreur sur le coefficient de la singularité de la solution d'un problème elliptique sur un ouvert avec coin*, RAIRO Anal. Numér., 14 (1980), pp. 239–248.

[18] P. Grisvard, *Singularités en élasticité*, Arch. Ration. Mech. Anal., 107 (2) (1989), pp. 157–180.

[19] B. Heinrich, S. Nicaise, and B. Weber, *Elliptic interface problems in axisymmetric domains.* II. *The Fourier-finite-element approximation of non-tensorial singularities*, Adv. Math. Sci. Appl., 10 (2000), pp. 571–600.

[20] M. Lenczner, *Méthode de calcul du coefficient de singularité pour la solution du problème de Laplace dans un domaine diédral*, RAIRO Modél. Math. Anal. Numér., 27 (1993), pp. 395–420.

[21] V. Maz'ya and B. Plamenevskii, *On the coefficients in the asymptotics of solutions of elliptic boundary-value problems near conical points*, Soviet Math. Dokl., 15 (1974), pp. 1570–1575.

[22] V. Maz'ya and B. Plamenevskii, *On the coefficients in the asymptotics of solutions of elliptic boundary value problems near the edge*, Soviet Math. Dokl., 17 (1976), pp. 970–974.

[23] V. G. Maz'ya and B. A. Plamenevskii, *Coefficients in the asymptotics of the solutions of an elliptic boundary value problem in a cone*, J. Soviet Math., 9 (1978), pp. 750–764.

[24] V. G. Maz'ya and B. A. Plamenevskii, *L^p estimates of solutions of elliptic boundary value problems in a domain with edges*, Trans. Moscow Math. Soc., 1 (1980), pp. 49–97.

[25] V. G. Maz'ya and B. A. Plamenevskii, *On the coefficients in the asymptotic of solutions of the elliptic boundary problem in domains with conical points*, Amer. Math. Soc. Transl. (2), 123 (1984), pp. 57–88.

[26] V. G. Maz'ya and J. Rossmann, *Über die Lösbarkeit und die Asymptotik der Lösungen elliptischer Randwertaufgaben in Gebieten mit Kanten.* III, preprint P-MATH-31/84, Akad. Wiss. DDR, Inst. Math., Berlin, 1984.

[27] V. G. Maz'ya and J. Rossmann, *Über die Asymptotik der Lösungen elliptischer Randwertaufgaben in der Umgebung von Kanten*, Math. Nachr., 138 (1988), pp. 27–53.

[28] M.-A. Moussaoui, *Sur l'approximation des solutions du problème de Dirichlet dans un ouvert avec coins*, in Singularities and Constructive Methods for Their Treatment (Oberwolfach, 1983), Springer-Verlag, Berlin, pp. 199–206.

[29] A. Rössle, *Corner singularities and regularity of weak solutions for the two-dimensional Lamé equations on domains with angular corners*, J. Elasticity, 60 (2000), pp. 57–75.

[30] J. Rossmann and A.-M. Sändig, *Formulas for the coefficients in the asymptotics of solutions of boundary value problems for second order systems near edges*, ZAMM Z. Angew. Math. Mech., 76 (1996), pp. 181–184.

[31] E. P. Stephan and M. Costabel, *A boundary element method for three-dimensional crack problems*, in Innovative Numerical Methods in Engineering (Atlanta, GA, 1986), Computational Mechanics, Southampton, 1986, pp. 351–360.

[32] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[33] B. A. Szabó and Z. Yosibash, *Numerical analysis of singularities in two dimensions.* I. *Computation of eigenpairs*, Internat. J. Numer. Methods Engrg., 38 (1995), pp. 2055–2082.

[34] B. A. Szabó and Z. Yosibash, *Numerical analysis of singularities in two dimensions.* II. *Computation of generalized flux/stress intensity factors*, Internat. J. Numer. Methods Engrg., 39 (1996), pp. 409–434.

[35] Z. Yosibash, *Computing singular solutions of elliptic boundary value problems in polyhedral domains using the p-FEM*, Appl. Numer. Math. 33 (2000), pp. 71–93.

[36] Z. Yosibash, R. Actis, and B. Szabó, *Extracting edge flux intensity functions for the Laplacian*, Internat. J. Numer. Methods Engrg., 53 (2002), pp. 225–242.

SIAM J. MATH. ANAL.
Vol. 35, No. 5, pp. 1203–1212

THE BRAMBLE–HILBERT LEMMA FOR CONVEX DOMAINS*

S. DEKEL† AND D. LEVIATAN‡

Abstract. The Bramble–Hilbert lemma is a fundamental result on multivariate polynomial approximation. It is frequently applied in the analysis of finite elements methods (FEM) used for numerical solutions of PDEs. However, this classical estimate depends on the geometry of the domain and may "blow up" for simple examples such as a sequence of triangles of equivalent diameter that become thinner and thinner. Thus, in FEM applications one usually requires that the mesh has "quasi-uniform" geometry. This assumption is perhaps too restrictive when one tries to obtain estimates of nonlinear approximation methods that use piecewise polynomials.

Our main result that improves upon this point is the following. Let $\Omega \subset \mathbb{R}^n$ be a bounded convex domain and let $g \in W_p^m(\Omega)$, $m \in \mathbb{N}$, $1 \le p \le \infty$, where $W_p^m(\Omega)$ is the Sobolev space. Then there exists a polynomial P of total degree $m-1$ for which

$$|g - P|_{k,p} \le C(n,m)\big(\operatorname{diam}\Omega\big)^{m-k}|g|_{m,p}, \quad k = 0, 1, \dots, m,$$

where $|\cdot|_{k,p} := \sum_{|\alpha|=k} \|D^\alpha \cdot\|_{L_p(\Omega)}$ is the Sobolev seminorm of order k. As a consequence we get that for $f \in L_p(\Omega)$,

$$E_{m-1}(f,\Omega)_p \approx K_m\Big(f, \big(\operatorname{diam}\Omega\big)^m\Big)_p,$$

where $E_{m-1}(f,\Omega)_p := \inf_{P\in\Pi_{m-1}} \|f-P\|_{L_p(\Omega)}$ is the error of polynomial approximation of degree $m-1$ and $K_m(\ ,\ )_p$ is the K-functional associated with the pair $(L_p(\Omega), W_p^m(\Omega))$, and where the constants of equivalence depend only on m and n.

For the case of convex domains (elements) this extends a recent result for $p = 2$, and for $m = 1$ and $2 < p \le \infty$. This also improves previous results where the constant in the estimate further depends on the geometry of the domain, or where there is a constraint $p > n(\ge 2)$.

Key words. Bramble–Hilbert lemma, multivariate nonlinear approximation, finite element methods

AMS subject classifications. 41A10, 41A25, 41A63, 65M60

DOI. 10.1137/S0036141002417589

1. Introduction.

We begin by recalling classical smoothness measures over multivariate domains. Here and throughout the paper we assume that the domain $\Omega \subset \mathbb{R}^n$ is compact with a nonempty interior. A first notion of smoothness uses the *Sobolev spaces* $W_p^m(\Omega)$. These are spaces of functions $g \in L_p(\Omega)$ which have all their distributional derivatives of order up to m, $D^\alpha g := \frac{\partial^k g}{\partial x_1^{\alpha_1}\cdots\partial x_n^{\alpha_n}}$, $\alpha = (\alpha_1, \dots, \alpha_n)$, $\alpha \in \mathbb{Z}_+^n$, $|\alpha| := \sum_{i=1}^n \alpha_i = k$, $0 \le k \le m$, in $L_p(\Omega)$. The seminorm of $W_p^m(\Omega)$ is given by $|g|_{m,p} := \sum_{|\alpha|=m} \|D^\alpha g\|_{L_p(\Omega)} < \infty$ and may be regarded as a measure of the smoothness of order m of a function, provided the function is in the appropriate Sobolev space. The *K-functional* of order m of $f \in L_p(\Omega)$ (see, e.g., [De], [BeSh]) is defined by

$$K_m(f,t)_p := K\big(f,t,L_p(\Omega),W_p^m(\Omega)\big) := \inf_{g\in W_p^m(\Omega)} \{\|f-g\|_p + t|g|_{m,p}\}. \tag{1.1}$$

*Received by the editors November 12, 2002; accepted for publication (in revised form) August 8, 2003; published electronically January 6, 2004.
http://www.siam.org/journals/sima/35-5/41758.html
†RealTimeImage, 6 Hamasger St., Or-Yehuda 60408, Israel (shai.dekel@turboimage.com).
‡School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel (leviatan@math.tau.ac.il).

Since we assume Ω to be compact we may denote

$$K_m(f,\Omega)_p := K_m(f,d^m)_p, \tag{1.2}$$

where $d := \operatorname{diam}\Omega$.

For $f \in L_p(\Omega)$, $1 \le p \le \infty$, $h \in \mathbb{R}^n$, and $m \in \mathbb{N}$, we recall the mth order difference operator $\Delta_h^m(f,\cdot) : \Omega \to \mathbb{R}$

$$\Delta_h^m(f,x) := \Delta_h^m(f,\Omega,x) := \begin{cases} \sum_{k=0}^m (-1)^{m-k}\binom{m}{k} f(x+kh), & [x,x+mh] \subset \Omega, \\ 0 & \text{otherwise,} \end{cases}$$

where $[x,y]$ denotes the line segment connecting any two points $x, y \in \mathbb{R}^n$. The *modulus of smoothness* (see, e.g., [De], [BeSh]) is defined by

$$\omega_m(f,t)_p := \sup_{|h|\le t} \|\Delta_h^m(f,\Omega,\cdot)\|_{L_p(\Omega)}, \qquad t > 0, \tag{1.3}$$

where for $h \in \mathbb{R}^n$, $|h|$ denotes the norm of h. We also denote

$$\omega_m(f,\Omega)_p := \sup_{h\in\mathbb{R}^n} \|\Delta_h^m(f,\Omega,\cdot)\|_{L_p(\Omega)}. \tag{1.4}$$

It is known that the above two notions of smoothness, (1.1) and (1.3), are sometimes equivalent (see section 5.4 in [BeSh] for the case $\Omega = \mathbb{R}^n$ and [JS] for the case of multivariate Lipschitz domains). That is, there exist $C_1, C_2 > 0$, such that for any $t > 0$

$$C_1 K_m(f,t^m)_p \le \omega_m(f,t)_p \le C_2 K_m(f,t^m)_p. \tag{1.5}$$

However, while it is easy to show that C_2 in (1.5) depends only on m (see [BeSh, (5.4.33)]), the constant C_1 may further depend on the geometry of Ω.

Let $\Pi_{m-1} := \Pi_{m-1}(\mathbb{R}^n)$ denote the multivariate polynomials of total degree $m-1$ (order m) in n variables. Given a "nontrivial" multivariate domain, our goal is to estimate the degree of approximation of a function $f \in L_p(\Omega)$, $1 \le p \le \infty$,

$$E_{m-1}(f,\Omega)_p := \inf_{P\in\Pi_{m-1}} \|f - P\|_{L_p(\Omega)},$$

using one of the above notions of smoothness. One of the classical results in this direction is the *Bramble–Hilbert lemma* [BrHi]. To introduce it we require the following definitions.

A domain Ω is *star-shaped with respect to a ball* $B \subseteq \Omega$ if for each point $x \in \Omega$, the closed convex-hull of $\{x\} \cup B$ is contained in Ω. Let $\rho_{\max} = \max\{\rho : \Omega \text{ is star-shaped with respect to a ball } B \subseteq \Omega \text{ of radius } \rho\}$. The *chunkiness parameter* of Ω is defined by

$$\gamma := \frac{d}{\rho_{\max}} \qquad (d = \operatorname{diam}\Omega). \tag{1.6}$$

This leads to the following formulation of the Bramble–Hilbert lemma (a weaker formulation estimates, instead, sublinear functionals; see Corollary 1.5).

Bramble–Hilbert Lemma. *Let Ω be star-shaped with respect to some ball B and let $g \in W_p^m(\Omega)$, $1 \le p \le \infty$, $m \in \mathbb{N}$. Then there exists a polynomial $P \in \Pi_{m-1}$ for which*

$$|g - P|_{k,p} \le C(n,m,\gamma) d^{m-k} |g|_{m,p}, \qquad k = 0, 1, \dots, m. \tag{1.7}$$

See Chapter 4 in [BrSc] for a proof of this result and [H] for a slightly stronger version of (1.7). Obviously the main drawback of (1.7) is that the constant depends on the chunkiness parameter (1.6) which "blows up," for example, in the case of a sequence of triangles of equivalent diameter that become thinner and thinner. This problem is usually resolved in the finite elements methods (FEM) literature by assuming that the mesh is *quasi-uniform*, i.e., that the collection of domains (elements) used to discretize the given problem has a uniformly bounded chunkiness parameter.

Perhaps another limitation of (1.7) is that it is too restrictive to be applied in estimates in nonlinear approximation by piecewise polynomials. For instance, let $f \in L_p([0,1]^2)$ and define $S_N^m(\mathbb{R}^2)$ to be the collection

$$\sum_{k=1}^{N} \mathbf{1}_{\Delta_k} P_k,$$

where Δ_k are triangles with disjoint interiors and $P_k \in \Pi_{m-1}(\mathbb{R}^2)$, and we wish to estimate (see [KP], [DLS])

$$\sigma_{N,m}(f)_p := \inf_{\varphi \in S_N^m} \|f - \varphi\|_{L_p([0,1]^2)}.$$

Thus, there have been quite a few attempts at removing the dependence of the constants on the geometry of Ω, and of estimating them. Perhaps the most significant result has recently been obtained by Verfürth [V], in the case of convex domains and $p = 2$. Using the notation $H^m := W_2^m$, Verfürth has proved the following proposition.

PROPOSITION (see [V]). *Let Ω be a convex domain and let $g \in H^m(\Omega)$, $m \in \mathbb{N}$. Then there exists a polynomial $P \in \Pi_{m-1}$ for which*

$$|g - P|_{H^k} \le C(n,m) d^{m-k} |g|_{H^m}, \qquad k = 0, 1, \dots, m-1. \tag{1.8}$$

Also if $m = 1$, and if $g \in W_p^1$, $2 < p \le \infty$, then

$$\|g - P\|_{L_p(\Omega)} \le C(n,p) d |g|_{W_p^1}. \tag{1.9}$$

Verfürth gives concrete estimates of the above constants and has some further results for star-shaped domains as well.

Earlier, Dechevski, and Quak [DQ] improved the Bramble–Hilbert lemma in some cases. Their result applies to the larger class of domains that are star-shaped with respect to a point. A domain Ω is *star-shaped with respect to a point $x_0 \in \Omega$* if for any point $x \in \Omega$ the line segment $[x_0, x]$ is contained in Ω. The following is a modified version of their result.

PROPOSITION (see [DQ]). *Let Ω be a Lipschitz domain, which is star-shaped with respect to a point $x_0 \in \Omega$. Then for $m \in \mathbb{N}$ and $2 \le n < p \le \infty$, there exists a polynomial $P \in \Pi_{m-1}$ for which*

$$|g - P|_{k,p} \le C(n,m,p) d^{m-k} |g|_{m,p}, \qquad k = 0, 1, \dots, m. \tag{1.10}$$

Although the constant in (1.10) does not depend on geometrical parameters such as (1.6), the above proposition assumes the constraint $n < p$ that does not cover one of the most common cases in applications of the FEM, namely, $n = p = 2$.

Our approach differs from previous work in one crucial detail. For convex domains we can construct an approximating polynomial that is more adaptive to the shape

of the domain. Thus, instead of constructing a polynomial using either some center point $x_0 \in \Omega$ or some maximal but relatively small ball $B \subset \Omega$, our construction uses John's "maximal" ellipsoid (see Proposition 3.2) combined with a simple affine transformation argument. Our main result is

THEOREM 1.1. *Let $\Omega \subset \mathbb{R}^n$ be convex, and let $g \in W_p^m(\Omega)$, $m \in \mathbb{N}$, $1 \le p \le \infty$. Then there exists a polynomial $P \in \Pi_{m-1}$ for which*

$$|g - P|_{k,p} \le C(n,m) d^{m-k} |g|_{m,p}, \quad k = 0, 1, \dots, m. \tag{1.11}$$

We emphasize that our proof of Theorem 1.1 is constructive and we are going to specify the polynomial P which yields (1.11). In fact we show that one may take $P(x) := Q^m(g(A\cdot)(A^{-1}x)$, where Q^m is the averaged Taylor polynomial over the ball $B(0,1) \subset \mathbb{R}^n$, and A is an affine transformation related to Ω (see definitions and details in sections 2 and 3).

A direct consequence of Theorem 1.1 is the following.

COROLLARY 1.2. *For all convex domains $\Omega \subset \mathbb{R}^n$ and functions $f \in L_p(\Omega)$, $1 \le p \le \infty$,*

$$E_{m-1}(f,\Omega)_p \approx K_m(f,\Omega)_p,$$

where $K_m(f,\Omega)_p$ is defined in (1.2)*, and the constants of equivalency depend only on m and n.*

We wish to point out a recent result of Karaivanov and Petrushev [KP] who showed that if $\Delta \subset \mathbb{R}^2$ is a triangle and $f \in L_p(\Delta)$, $0 < p \le \infty$, then for any $m \in \mathbb{N}$

$$E_{m-1}(f,\Delta)_p \le C(m,p)\omega_m(f,\Delta)_p, \tag{1.12}$$

where $\omega_m(f,\Delta)_p$ is defined in (1.4). This implies that for all triangles $\Delta \subset \mathbb{R}^2$ and functions $f \in L_p(\Delta)$, $1 \le p \le \infty$, we have the equivalence

$$E_{m-1}(f,\Delta)_p \approx \omega_m(f,\Delta)_p \approx K_m(f,\Delta)_p,$$

where the constants of equivalence depend only on p and m. Indeed, it is this result that motivated us to try to find shape-independent estimates.

We also get the following formulation of the Bramble–Hilbert lemma.

COROLLARY 1.3. *Let $\Omega \subset \mathbb{R}^n$ be convex, and let l be a sublinear functional given on $W_p^m(\Omega)$, $m \in \mathbb{N}$, $1 \le p \le \infty$, with the following properties.*

(i) *There exists a constant $\widetilde{C}$ such that for all $g \in W_p^m(\Omega)$, $|l(g)| \le \widetilde{C} \sum_{k=0}^m d^k |g|_{k,p}$;*

(ii) *$l(P) = 0$ for all $P \in \Pi_{m-1}$.*

Then for all $g \in W_p^m(\Omega)$,

$$|l(g)| \le C(n,m,\widetilde{C}) d^m |g|_{m,p}.$$

Section 2 reviews the averaged Taylor polynomial approach to the classical Bramble–Hilbert lemma (see Chapter 4 in [BrSc]). In section 3 we introduce John's theorem and explain how this tool can be applied in the case of convex domains via an affine transformation argument. Finally, in section 4 we assemble all the above tools to give a constructive proof of Theorem 1.1. We also define the notion of "almost convex" domains and note that our results extend to this case too.

2. The averaged Taylor polynomial. We recall some basic definitions of multivariate polynomials, differentials, and Taylor series. Throughout this section we use the notation of Chapter 4 in [BrSc]. For a multi-index $\alpha \in \mathbb{Z}_+^n$ let $\alpha! = \prod_{i=1}^n \alpha_i!$, and denote by $x^\alpha := \prod_{i=1}^n x_i^{\alpha_i}$ the *multivariate monomial of total degree* $|\alpha|$. Denote the set of all multivariate polynomials of total degree $m-1$ by

$$\Pi_{m-1}(\mathbb{R}^n) := \Big\{ \sum_{|\alpha|\le m-1} c_\alpha x^\alpha \Big\}.$$

The classical *Taylor polynomial of order* m (*degree* $m-1$) of a function $g \in C^m(\Omega)$ at $x \in \Omega$, about the point $y \in \Omega$, is given by

$$T_y^m g(x) := \sum_{|\alpha|<m} \frac{D^\alpha g(y)}{\alpha!}(x-y)^\alpha. \tag{2.1}$$

The *Taylor remainder of order* m of a function $g \in C^m(\Omega)$ at $x \in \Omega$, about the point $y \in \Omega$, is given by

$$TR_y^m g(x) := m \sum_{|\alpha|=m} \frac{(x-y)^\alpha}{\alpha!} \int_0^1 s^{m-1} D^\alpha g\big(x + s(y-x)\big) ds. \tag{2.2}$$

It is meaningful provided the segment $[y,x]$ is contained in Ω. Then we have

$$g(x) = T_y^m g(x) + TR_y^m g(x).$$

Next we introduce the averaged Taylor polynomial. It can be shown that for a ball $B(x_0,\rho) := \{z \in \mathbb{R}^n : |z-x_0| \le \rho\}$ there exists a *cut-off function* ϕ_B with the following properties:

(i) $\int_{\mathbb{R}^n} \phi_B(x)dx = 1$,
(ii) $\operatorname{supp}(\phi_B) = B$,
(iii) $\phi_B \in C^\infty(\mathbb{R}^n)$,
(iv) $\|\phi_B\|_\infty \le \rho^{-n}$.

Given $g \in C^m(\Omega)$, the *averaged Taylor polynomial of order* m (*degree* $m-1$) (averaged over a ball $B \subseteq \Omega$) is defined by

$$Q^m g(x) := \int_B T_y^m g(x)\phi_B(y)dy, \qquad x \in \Omega. \tag{2.3}$$

We also define the *averaged Taylor remainder*, namely,

$$R^m g(x) := g(x) - Q^m g(x). \tag{2.4}$$

The following lemma is a special case of the classical Bramble–Hilbert lemma which estimates the (simultaneous) degree of approximation of the averaged Taylor polynomial in a "normalized" setting. For the proof see Theorem 4.3.8 in [BrSc]; observe that the chunkiness parameter (1.6) in this case depends only on n.

LEMMA 2.1. *Let* $B(0,1) \subseteq \Omega \subseteq B(0,n)$ *be star-shaped with respect to* $B(0,1)$. *Then for any* $g \in C^m(\Omega)$, $m \in \mathbb{N}$, *and* $1 \le p \le \infty$, *we have*

$$|g - Q^m g|_{k,p} \le C(n,m)|g|_{m,p}, \qquad k = 0,1,\dots,m,$$

where Q^m *is averaged over* $B(0,1)$.

3. John's theorem.

DEFINITION 3.1. *An ellipsoid E is the image of the closed unit ball in $\mathbb{R}^n$ under a nonsingular affine mapping $A(x) = Mx + b$, $M \in M_{n\times n}(\mathbb{R})$, $b \in \mathbb{R}^n$. The center of E is $b = A(0)$.*

The next result [J] (see also [Ba]) is the crucial ingredient that is missing in previous work. Let $c + n(E - c) := \{c + n(x - c) : x \in E\}$.

PROPOSITION 3.2 (John's theorem). *Let $\Omega \subset \mathbb{R}^n$ be convex. Then there exists an ellipsoid $E \subseteq \Omega$ such that if x_0 is the center of E, then*

$$E \subseteq \Omega \subseteq x_0 + n(E - x_0).$$

By Definition 3.1, John's theorem implies that for each convex domain Ω we can find a nonsingular affine mapping A such that

$$B(0,1) \subseteq A^{-1}(\Omega) \subseteq B(0,n).$$

It is interesting to note that John's ellipsoid is the ellipsoid $E \subseteq \Omega$ with maximal volume. In some sense this means that E "covers" Ω sufficiently well.

To use John's maximal ellipsoid (or equivalently, John's optimal affine transformation), we apply the following commutativity of Taylor polynomials and differentiation.

LEMMA 3.3. *Let $A(x) = Mx + b$, $M \in M_{n\times n}(\mathbb{R})$, $b \in \mathbb{R}^n$, be a nonsingular affine mapping, and let $g \in C^m(\Omega)$. Then for any $x \in \Omega$, $y \in A^{-1}(\Omega)$, and $\alpha \in \mathbb{Z}_+^n$, $1 \le |\alpha| \le m-1$, we have*

$$D_x^\alpha \Big[T_y^m \big(g(A\cdot)\big)(A^{-1}x) \Big] = T_y^{m-|\alpha|} \big((D^\alpha g)(A\cdot)\big)(A^{-1}x). \tag{3.1}$$

Proof. Observe that it is sufficient to prove that for any $1 \le k \le m-1$ and $1 \le s \le n$,

$$D_x^{e_s} \Big[\sum_{|\beta|=k} \frac{D_y^\beta \tilde{g}(y)}{\beta!} (A^{-1}x - y)^\beta \Big] = \sum_{|\gamma|=k-1} \frac{D_y^\gamma \widetilde{g_{x_s}}(y)}{\gamma!} (A^{-1}x - y)^\gamma, \tag{3.2}$$

where $\tilde{g} := g(A\cdot)$, $\widetilde{g_{x_s}} := g_{x_s}(A\cdot)$, $g_{x_s} := \frac{\partial g}{\partial x_s}$, and $\{e_s\}_{s=1,\dots,n}$ is the standard basis of $\mathbb{R}^n$. The case of a general multivariate derivative D_x^α follows by repeated applications of (3.2), and the Taylor series formulation (3.1) is obtained by adding all the degrees $1 \le k \le m-1$. To prove the above let $M =: (a_{i,j})_{1\le i,j\le n}$ and $M^{-1} =: (b_{i,j})_{1\le i,j\le n}$. In the calculations below, if $\beta_i = 0$, then differentiating $(A^{-1}x - y)^\beta$ with respect to x_s does not produce the term $\beta_i b_{i,s}(A^{-1}x - y)^{\beta - e_i}$; rather we have 0, and it does not appear in the summation. Hence in this case we regard $\beta_i b_{i,s}(A^{-1}x - y)^{\beta-e_i} := 0$ and $(\beta - e_i)! = \infty$, and again the term is not there. This takes care of itself automatically when we switch the summation below from β to $\gamma = \beta - e_i$.

$$\begin{aligned} D_x^{e_s} \left[\sum_{|\beta|=k} \frac{D_y^\beta \tilde{g}(y)}{\beta!} (A^{-1}x - y)^\beta \right] &= \sum_{|\beta|=k} \frac{D_y^\beta \tilde{g}(y)}{\beta!} D_x^{e_s} \big((A^{-1}x - y)^\beta\big) \\ &= \sum_{|\beta|=k} \frac{D_y^\beta \tilde{g}(y)}{\beta!} \sum_{i=1}^n \beta_i b_{i,s} (A^{-1}x - y)^{\beta - e_i} \\ &= \sum_{|\beta|=k} \sum_{i=1}^n \frac{D_y^\beta \tilde{g}(y)}{(\beta - e_i)!} b_{i,s} (A^{-1}x - y)^{\beta - e_i} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{|\gamma|=k-1} \frac{(A^{-1}x-y)^\gamma}{\gamma!} \sum_{i=1}^n b_{i,s} D_y^{\gamma+e_i} \tilde{g}(y) \\
&= \sum_{\gamma|=k-1} \frac{(A^{-1}x-y)^\gamma}{\gamma!} \sum_{i=1}^n b_{i,s} D_y^{\gamma} \Big(\sum_{j=1}^n a_{j,i} g_{x_j}(Ay) \Big) \\
&= \sum_{|\gamma|=k-1} \frac{(A^{-1}x-y)^\gamma}{\gamma!} \sum_{j=1}^n D_y^\gamma \big(g_{x_j}(Ay)\big) \sum_{i=1}^n a_{j,i} b_{i,s} \\
&= \sum_{|\gamma|=k-1} \frac{(A^{-1}x-y)^\gamma}{\gamma!} \sum_{j=1}^n D_y^\gamma \big(g_{x_j}(Ay)\big) \delta_{j,s} \\
&= \sum_{|\gamma|=k-1} \frac{D_y^\gamma (\widetilde{g_{x_s}}(y))}{\gamma!} (A^{-1}x-y)^\gamma \ . \qquad \square
\end{aligned}
$$

By (2.3), we have the following corollary.

COROLLARY 3.4. *Let $\Omega \subset \mathbb{R}^n$, and let A be a nonsingular affine mapping such that $B(0,1) \subseteq A^{-1}(\Omega)$. Then for $g \in C^m(\Omega)$ and $\alpha \in \mathbb{Z}_+^n$, $|\alpha| = k$, $1 \le k \le m-1$,*

$$
D^\alpha \Big[Q^m \big(g(A\cdot)\big)(A^{-1}x) \Big] = Q^{m-k} \big((D^\alpha g)(A\cdot) \big)(A^{-1}x), \tag{3.3}
$$

where Q^m is averaged on $B(0,1)$.

Observing that affine transformations map convex domains onto convex domains, the following argument, when combined with John's theorem, is the main tool of our approach.

LEMMA 3.5. *Let $\Omega \subset \mathbb{R}^n$, and let A be a nonsingular affine mapping such that $B(0,1) \subseteq A^{-1}(\Omega) \subseteq B(0,n)$ and $A^{-1}(\Omega)$ is star-shaped with respect to $B(0,1)$. Then for $g \in C^m(\Omega)$, $1 \le p < \infty$, and $P(x) = Q^m(g(A\cdot))(A^{-1}x)$ (where Q^m is averaged on $B(0,1)$), we have*

$$
|g-P|_{W_p^k(\Omega)} \le C(n,m) d^{m-k} |g|_{W_p^m(\Omega)}, \qquad k = 0, 1, \dots, m. \tag{3.4}
$$

Proof. Since $A(x) = Mx + b$ maps $B(0,1)$ into Ω, we conclude that $\|M\|_2 \le d$. Thus with $M = (a_{i,j})_{1 \le i,j \le n}$, we have that $\max_{1 \le i,j \le n} |a_{i,j}| \le d$. Recalling that $\tilde{g} = g(A\cdot)$, this implies that for $y \in A^{-1}(\Omega)$, $x = Ay$, and $\alpha \in \mathbb{Z}_+^n$, $|\alpha| = i$, $i = 0, \dots, m$,

$$
|D_y^\alpha \tilde{g}(y)| \le d^i \sum_{|\gamma|=i} |D^\gamma g)(Ay)|,
$$

and hence, in particular,

$$
\sum_{|\alpha|=m} \|D_y^\alpha \tilde{g}\|_{L_p(A^{-1}(\Omega))} \le C(n,m) d^m \sum_{|\alpha|=m} \|(D^\alpha g)(A\cdot)\|_{L_p(A^{-1}(\Omega))}. \tag{3.5}
$$

We can now prove (3.4) for $k = 0$. Let $\widetilde{P} := Q^m(g(A\cdot))$; then by Lemma 2.1 and (3.5)

$$
\begin{aligned}
\|g-P\|_{L_p(\Omega)} &= |\det M|^{1/p} \|\tilde{g} - \widetilde{P}\|_{L_p(A^{-1}(\Omega))} \\
&\le C(n,m) |\det M|^{1/p} |\tilde{g}|_{W_p^m(A^{-1}(\Omega))} \\
&= C(n,m) |\det M|^{1/p} \sum_{|\alpha|=m} \|D_y^\alpha \tilde{g}\|_{L_p(A^{-1}(\Omega))}
\end{aligned}
$$

$$\begin{aligned}
&\le C(n,m)|\det M|^{1/p} d^m \sum_{|\alpha|=m} \|(D^\alpha g)(A\cdot)\|_{L_p(A^{-1}(\Omega))} \\
&= C(n,m) d^m \sum_{|\alpha|=m} \|D_x^\alpha g\|_{L_p(\Omega)} \\
&= C(n,m) d^m |g|_{W_p^m(\Omega)}.
\end{aligned}$$

For $1 \le k \le m-1$ take $\alpha \in \mathbb{Z}_+^n$, $|\alpha| = k$, $1 \le k \le m-1$, and let $h := D^\alpha g$. Then (3.3) yields

$$\|D^\alpha(g-P)\|_{L_p(\Omega)} = \big\|h(x) - Q^{m-k}\big(h(A\cdot)\big)(A^{-1}x)\big\|_{L_p(\Omega)}.$$

By the case $k=0$ proved above,

$$\big\|h(x) - Q^{m-k}\big(h(A\cdot)\big)(A^{-1}x)\big\|_{L_p(\Omega)} \le C(n,m) d^{m-k} |h|_{m-k,p},$$

which in turn implies that

$$\|D^\alpha(g-P)\|_{L_p(\Omega)} \le C(n,m) d^{m-k} |g|_{m,p}. \tag{3.6}$$

Summing up (3.6) over all $\alpha \in \mathbb{Z}_+^n$, $|\alpha| = k$, we obtain the required result. The case $k = m$ is trivial. □

4. Proofs of the main results.

Proof of Theorem 1.1. The proof of (1.11) for the case $p = \infty$ can be applied to star-shaped domains with respect to a point x_0, by using the classical Taylor polynomial (2.1) at the point $y = x_0$ and estimating the remainder (2.2). We leave the details to the reader and assume $1 \le p < \infty$. Let $E \subseteq \Omega$ be John's maximal ellipsoid (see Proposition 3.2) and A the corresponding affine mapping, i.e., $A(B(0,1)) = E$. John's theorem implies that

$$B(0,1) \subseteq A^{-1}(\Omega) \subseteq B(0,n).$$

First assume that $g \in C^m(\Omega)$. By Lemma 3.5 the polynomial $P(x) = Q^m(g(A\cdot))(A^{-1}x)$ is in Π_{m-1} and satisfies

$$|g-P|_{k,p} \le C(n,m) d^{m-k} |g|_{m,p}, \quad k = 0, 1, \dots, m.$$

Since $C^\infty(\Omega)$ is dense in $W_p^m(\Omega)$ (see, e.g., Theorem 1.3.4 in [BrSc]), the proof of the general case follows from a standard density argument. □

Proof of Corollary 1.2. The method of proof is standard but we give it for the sake of completeness. Let $f \in L_p(\Omega)$ and $g \in W_p^m(\Omega)$ be such that

$$\|f-g\|_p + d^m |g|_{m,p} \le 2K_m(f,\Omega)_p.$$

By (1.9) with $k=0$, there exists $P \in \Pi_{m-1}$ such that

$$\|g-P\|_p \le C(n,m) d^m |g|_{m,p}.$$

Therefore

$$\begin{aligned}
E_{m-1}(f)_p &\le \|f-P\|_p \\
&\le \|f-g\|_p + \|g-P\|_p \\
&\le \|f-g\|_p + C(n,m) d^m |g|_{m,p} \\
&\le C(n,m) K_m(f,\Omega)_p.
\end{aligned}$$

In the other direction, it is easy to see from (1.1) that for any polynomial $Q \in \Pi_{m-1}$ and any $t > 0$,

$$K_m(f,t)_p \le \|f - Q\|_p.$$

Consequently,

$$K_m(f,\Omega)_p \le E_{m-1}(f)_p. \qquad \square$$

Proof of Corollary 1.3. Let $g \in W_p^m(\Omega)$, and let P be the polynomial for which (1.11) holds. Then by property (ii) of the sublinear functional l we have that $|l(g)| \le |l(g-P)|$. Property (i) and (1.11) yield

$$\begin{aligned} |l(g)| &\le |l(g-P)| \\ &\le \widetilde{C} \sum_{k=0}^{m} d^k |g-P|_{k,p} \\ &\le \widetilde{C} C(n,m) \sum_{k=0}^{m} d^k d^{m-k} |g|_{m,p} \\ &\le C(n,m,\widetilde{C}) d^m |g|_{m,p} \ . \qquad \square \end{aligned}$$

Finally, we would like to point out a certain natural extension of our results to slightly more general types of domains.

DEFINITION 4.1. *A compact domain $\Omega \subset \mathbb{R}^n$ with nonempty interior is almost convex if there exists a nonsingular affine mapping A, such that*

(i) $B(0,1) \subseteq A^{-1}(\Omega) \subseteq B(0,n)$,

(ii) $A^{-1}(\Omega)$ *is star-shaped with respect to* $B(0,1)$.

Indeed, John's theorem shows that every convex domain is almost convex. Furthermore, by the method used in this work (specifically Lemma 3.5) it can be seen that our main results remain valid for this type of domain.

Acknowledgments. We are grateful and indebted to the referees for bringing to our attention the paper by Verfürth. We also thank Manor Mendel for fruitful discussions of this work.

REFERENCES

[Ba] K. BALL, *Ellipsoids of maximal volume in convex bodies*, Geom. Dedicata, 41 (1992), pp. 241–250.

[BeSh] C. BENNETT AND R. SHARPLEY, *Interpolation of Operators*, Pure and Applied Mathematics 129, Academic Press, Boston, 1988.

[BrHi] J. H. BRAMBLE AND S. R. HILBERT, *Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolations*, SIAM J. Numer. Anal., 7 (1970), pp. 112–124.

[BrSc] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Texts in Appl. Math 15, Springer-Verlag, New York, 1994.

[De] R. A. DEVORE, *Nonlinear approximation*, Acta Numer., 7 (1998), pp. 51–150.

[DLS] S. DEKEL, D. LEVIATAN, AND M. SHARIR, *On bivariate smoothness spaces associated with nonlinear approximation*, Constr. Approx., to appear.

[DQ] L. T. DECHEVSKI AND E. G. QUAK, *On the Bramble-Hilbert lemma*, Numer. Funct. Anal. Optim., 11 (1990), pp. 485–495.

[H] S. M. HUDSON, *Polynomial approximation in Sobolev spaces*, Indiana Univ. Math. J., 39 (1990), pp. 199–228.

[J] J. Fritz, *Extremum problems with inequalities as subsidiary conditions*, in Studies and Essays Presented to R. Courant on his 60th Birthday, Interscience, New York, 1948, pp. 187–204.

[JS] H. Johnen and K. Scherer, *On the equivalence of the K-functional and the moduli of continuity and some applications*, in Constructive Theory of Functions of Several Variables, Lecture Notes in Math. 571, Springer-Verlag, Berlin, 1976, pp. 119–140.

[KP] B. Karaivanov and P. Petrushev, *Nonlinear piecewise polynomial approximation beyond Besov spaces*, Appl. Comput. Harmon. Anal., 15 (2003), pp. 177–223.

[V] R. Verfürth, *A note on polynomial approximation in Sobolev spaces*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 715–719.

SIAM J. MATH. ANAL.
Vol. 35, No. 5, pp. 1213–1240

DETERMINING A FUNCTION FROM ITS MEAN VALUES OVER A FAMILY OF SPHERES*

DAVID FINCH†, SARAH K. PATCH‡, AND RAKESH§

Abstract. Suppose D is a bounded, connected, open set in R^n and f is a smooth function on R^n with support in $\overline{D}$. We study the recovery of f from the mean values of f over spheres centered on a part or the whole boundary of D. For strictly convex $\overline{D}$, we prove uniqueness when the centers are restricted to an open subset of the boundary. We provide an inversion algorithm (with proof) when the mean values are known for all spheres centered on the boundary of D, with radii in the interval $[0, diam(D)/2]$. We also give an inversion formula when D is a ball in R^n, $n \geq 3$ and odd, and the mean values are known for all spheres centered on the boundary.

Key words. spherical mean values, wave equation

AMS subject classifications. 35L05, 35L15, 35R30, 44A05, 44A12, 92C55

DOI. 10.1137/S0036141002417814

1. Introduction.

Wave propagation and integral geometry are the physical and mathematical underpinnings of many medical imaging modalities. To date, standard modalities measure the same type of output energy as was input to the system. Ultrasound systems send and receive ultrasound waves; CT systems send and receive X-ray radiation. Recent work on a hybrid imaging technique, thermoacoustic tomography (TCT), uses radio frequency (RF) energy input at time t_0 and measures emitted ultrasound waves [18], [19], [20].

RF energy is deposited impulsively in time and uniformly throughout the imaging object, causing a small amount of thermal expansion. The premise is that cancerous masses absorb more RF energy than healthy tissue [17]. Cancerous masses preferentially absorb RF energy heat and expand more quickly than neighboring tissue, creating a pressure wave which is detected by ultrasound transducers at the edge of the object. Assuming constant sound speed, c, the sound waves detected at any point in time $t > t_0$ were generated by inclusions lying on the sphere of radius $c(t - t_0)$ centered at the transducer. Therefore, this imaging technique requires the inversion of a generalized Radon transform, because integrals of the tissue's RF absorption coefficient are measured over surfaces of spheres.

Figure 1.1 shows a TCT mammography system. The breast is immersed in a tank of water and transducers surround the exterior of the tank. Integrals of the RF absorption coefficient over spheres centered at each transducer are measured. Notice that only "limited angle" data may be measured, as we cannot put transducers on certain parts of the exterior of the tank.

The above motivated the study of the following mathematical problem. For a continuous, real valued function f on R^n, $n \geq 2$, p a point in R^n, and r a real

*Received by the editors November 14, 2002; accepted for publication (in revised form) July 4, 2003; published electronically January 6, 2004.
http://www.siam.org/journals/sima/35-5/41781.html
†Department of Mathematics, Oregon State University, Corvallis, OR 97331-4605 (finch@math.orst.edu).
‡GE Medical Systems, Mail Stop W-875, 3200 N. Grandview Boulevard, Waukesha, WI 53188 (sarah.patch@med.ge.com).
§Department of Mathematical Sciences, University of Delaware, Newark, DE 19716 (rakesh@math.udel.edu).

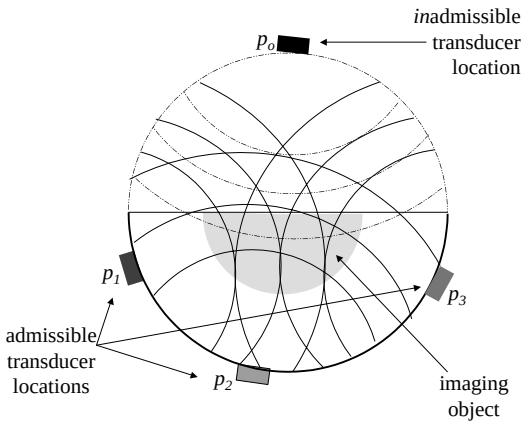


FIG. 1.1. *TCT mammography system.*

number, define the mean value operator

$$(Mf)(p,r) = \frac{1}{w_n} \int_{|\theta|=1} f(p + r\theta)\, d\theta,$$

where w_n is the surface area of the unit sphere in R^n. Let D denote a bounded, open, connected subset of R^n with a smooth boundary S. For functions f supported in D, we are interested in recovering f from the mean value of f over spheres centered on S; that is, given $(Mf)(p,r)$ for all p in S and all real numbers r, we wish to recover f. We also examine the situation where the mean values of f are given over spheres centered on an open subset of S.

In the rest of the article, $B_\rho(p)$ will represent the open ball of radius ρ centered at p, $\overline{B_\rho(p)}$ its closure, and $S_\rho(p)$ its boundary; Ω^c will represent the complement of Ω. Furthermore, all functions will be real valued.

We have the following results.

THEOREM 1 (uniqueness). *Suppose D is a bounded open subset of R^n, $n \geq 2$, with a smooth boundary S and $\overline{D}$ is strictly convex. Let Γ be any relatively open subset of S. If f is a smooth function on R^n, supported in $\overline{D}$, and $(Mf)(p,r) = 0$ for all $p \in \Gamma$ and all r, then $f = 0$.*

Here by the strict convexity of $\overline{D}$ we mean that if p, q are in $\overline{D}$, then any other point on the line segment pq is in D. Also, note that $(Mf)(p,r) = 0$ for all $p \in S$, $|r| > diam(D)$.

THEOREM 2 (reconstruction). *Suppose D is a bounded, open, connected subset of R^n, n odd and $n \geq 3$, with a smooth boundary S. If f is a smooth function on R^n, supported in $\overline{D}$, and $(Mf)(p,r)$ is known for all p in S and for all $r \in [0, diam(D)/2]$, then we may stably recover f. If $(Mf)(p,r)$ is known for all p in S and for all r, then f may be recovered by a simpler algorithm.*

Note that we do not assume $\overline{D}$ is convex, but we do need the centers to vary over all of S. If D is a ball in R^n, $n \geq 3$ and n odd, and we know the mean values for all spheres centered on the boundary of D, then we have an explicit inversion formula.

We introduce some notation to state the explicit inversion formula. Let $\tilde{C}^\infty(S_\rho(0)\times[0,\infty))$ consist of smooth functions $G(p,t)$ which are zero for t large and also $\partial_t^k G(p,t) = 0$ at $t=0$ for $k=0,1,2,\dots$ and all $p\in S_\rho(0)$. Let us define the operator

$$\begin{aligned}\mathcal{N} &: C_0^\infty(\overline{B_\rho(0)}) \to \tilde{C}^\infty\left(S_\rho(0)\times[0,\infty)\right),\\ &(\mathcal{N}f)(p,t) = t^{n-2}(Mf)(p,t), \qquad p\in S_\rho(0),\ t\ge 0\end{aligned}$$

and the operator (for odd $n\ge 3$)

$$\begin{aligned}\mathcal{D} &: \tilde{C}^\infty(S_\rho(0)\times[0,\infty)), \to \tilde{C}^\infty(S_\rho(0)\times[0,\infty)),\\ &(\mathcal{D}G)(p,t) = \left(\frac{1}{2t}\frac{\partial}{\partial t}\right)^{(n-3)/2}(G(p,t)).\end{aligned}$$

For example, $\mathcal{D}=I$ when $n=3$.

We now compute the formal L^2 adjoints of $\mathcal{N}$ and $\mathcal{D}$. For $G\in\tilde{C}^\infty(S_\rho(0)\times[0,\infty))$, using the change of variables $(t,\theta)\to y=p+t\theta$, we note that

$$\begin{aligned}\langle \mathcal{N}f\,, G\rangle &= \int_{|p|=\rho}\int_0^\infty (\mathcal{N}f)(p,t)\,G(p,t)\,dt\,dS_p\\ &= \frac{1}{\omega_n}\int_{|p|=\rho}\int_0^\infty\int_{|\theta|=1} t^{n-2}f(p+t\theta)\,G(p,t)\,d\theta\,dt\,dS_p\\ &= \frac{1}{\omega_n}\int_{R^n}\int_{|p|=\rho} f(y)\,\frac{G(p,|p-y|)}{|p-y|}\,dS_p\,dy\\ &= \langle f\,, \mathcal{N}^*G\rangle\end{aligned}$$

if we take

$$(\mathcal{N}^*G)(x) = \frac{1}{\omega_n}\int_{|p|=\rho}\frac{G(p,|p-x|)}{|p-x|}\,dS_p. \tag{1.1}$$

Note that for $G\in\tilde{C}^\infty(S_\rho(0)\times[0,\infty))$, $(\mathcal{N}^*G)(x)$ is a smooth function on R^n with compact support. The smoothness may be seen as follows: from the hypothesis on G, we may express $G(p,t)$ in the form $G(p,t)=tK(p,t^2)$ for $|p|=\rho$, $t\in[0,\infty)$ for some smooth function $K(p,s)$. Substituting this expression for G in the definition of $\mathcal{N}^*$, the smoothness of $\mathcal{N}^*G$ becomes clear.

Also

$$\begin{aligned}\langle \mathcal{D}G_1\,, G_2\rangle &= \int_{|p|=\rho}\int_0^\infty \left(\frac{1}{2t}\frac{\partial}{\partial t}\right)^{(n-3)/2}(G_1(p,t))\ G_2(p,t)\,dt\,dS_p\\ &= (-1)^{(n-3)/2}\int_{|p|=\rho}\int_0^\infty G_1(p,t)\ \left(\frac{\partial}{\partial t}\frac{1}{2t}\right)^{(n-3)/2}(G_2(p,t))\,dt\,dS_p\\ &= (-1)^{(n-3)/2}\int_{|p|=\rho}\int_0^\infty G_1(p,t)\,t\ \left(\frac{1}{2t}\frac{\partial}{\partial t}\right)^{(n-3)/2}\left(\frac{G_2(p,t)}{t}\right)\,dt\,dS_p\\ &= \langle G_1\,, \mathcal{D}^*G_2\rangle\end{aligned}$$

if we take

$$(\mathcal{D}^*G)(p,t) = (-1)^{(n-3)/2}t\,\mathcal{D}(G(p,t)/t). \tag{1.2}$$

Note that $\mathcal{D}^*$ maps functions in $\tilde{C}^\infty(S_\rho(0)\times[0,\infty))$ to functions in $\tilde{C}^\infty(S_\rho(0)\times[0,\infty))$.

THEOREM 3 (inversion formula). *If $n \geq 3$ and odd, f is a smooth function supported in $\overline{B_\rho(0)}$ and $(Mf)(p,r)$ (and hence $(\mathcal{N}f)(p,r)$) is known for all $p \in S_\rho(0)$ and all real r, then we have the explicit inversion formulas*

$$f(x) = -\frac{\pi}{2\,\rho\,\Gamma(n/2)^2}\left(\mathcal{N}^*\,\mathcal{D}^*\,\partial_t^2\,t\,\mathcal{D}\,\mathcal{N}f\right)(x), \qquad x \in B_\rho(0),$$

$$f(x) = -\frac{\pi}{2\,\rho\,\Gamma(n/2)^2}\left(\mathcal{N}^*\,\mathcal{D}^*\,\partial_t\,t\,\partial_t\,\mathcal{D}\,\mathcal{N}f\right)(x), \qquad x \in B_\rho(0),$$

$$f(x) = -\frac{\pi}{2\,\rho\,\Gamma(n/2)^2}\,\Delta_x\left(\mathcal{N}^*\,\mathcal{D}^*\,t\,\mathcal{D}\,\mathcal{N}f\right)(x), \qquad x \in B_\rho(0).$$

The inversion formulas in Theorem 3 are local in the sense that $f(x)$ is determined purely from the mean values of f over spheres centered on $S_\rho(0)$ passing through an arbitrarily small neighborhood of x. These inversion formulas also generate energy L^2 norm identities, which are a step toward a characterization of the range of the map $f \to Mf$.

There is some similarity between the inversion formula in Theorem 3 and the inversion formula for the Radon transform. The Radon transform of a function f on R^n is

$$(\mathcal{R}f)(\theta,r) = \int_{x\cdot\theta=r} f(x)\,dS_x \qquad \forall r \in (-\infty,\infty),\ \theta \in R^n,\ |\theta| = 1.$$

Its L^2 adjoint is, for every function F on $S_1(0)\times(-\infty,\infty)$,

$$(\mathcal{R}^*F)(x) = \int_{|\theta|=1} F(\theta, x\cdot\theta)\,d\theta \qquad \forall x \in R^n,$$

and the inversion formula for the Radon transform is (see [25])

$$f(x) = \frac{(-1)^{(n-1)/2}}{2(2\pi)^{n-1}}\Delta_x^{(n-1)/2}(\mathcal{R}^*\mathcal{R}f)(x) \qquad \forall x \in R^n.$$

The above theorems will be proved by converting the problem to a problem about the solutions of the wave equation. Consider the IVP

$$\Box u \equiv u_{tt} - \Delta u = 0, \qquad x \in R^n, \quad t \in R, \tag{1.3}$$

$$u(.,t{=}0) = 0,\ u_t(.,t{=}0) = f(.), \tag{1.4}$$

with f smooth and supported in $\overline{D}$. Then, from the standard theory for solutions of the wave equation, u is smooth in x,t, odd in t (because $-u(x,-t)$ is also the solution), and as shown in [7, p. 682] for $n \geq 2$,

$$u(x,t) = \frac{1}{(n-2)!}\frac{\partial^{n-2}}{\partial t^{n-2}}\int_0^t r\,(t^2-r^2)^{(n-3)/2}\,(Mf)(x,r)\,dr, \qquad t \geq 0. \tag{1.5}$$

Hence the original problem is equivalent to the problem of recovering $u_t(x,0)$ from the value of $u(x,t)$ on subsets of $S\times(-\infty,\infty)$. So Theorems 1 and 2 will follow from the following theorems.

THEOREM 4 (uniqueness). *Suppose D is a bounded open subset of R^n, $n \geq 2$, with a smooth boundary S, and $\overline{D}$ is strictly convex. Let Γ be a relatively open subset of S. Suppose f is a smooth function on R^n, supported in $\overline{D}$, and u is the solution of the IVP* (1.3) *and* (1.4). *If $u(p,t) = 0$ for all $p \in \Gamma$ and all t, then $f = 0$.*

The appropriate version of this result for $n = 1$ is also true and may be shown by arguments similar to (but simpler than) those used in proving the above theorem.

THEOREM 5 (reconstruction). *Suppose D is a bounded, open, connected subset of R^n, n odd, with a smooth boundary S. Suppose f is a smooth function on R^n, supported in $\overline{D}$, and u is the solution of the IVP* (1.3) *and* (1.4). *If $u(p,t)$ is known for all p in S and for all $t \in [0, diam(D)/2]$, then we may recover f. We have a simpler algorithm if $u(p,t)$ is known for all $t \in R$ (and all $p \in S$).*

In our reconstruction procedures, we use the fact that for n odd the fundamental solution of the wave operator is supported on the cone $t^2 = |x|^2$. This is not true in even dimensions, and so our algorithm is not valid in even space dimensions. Furthermore, the method of descent does not help, because if we consider u as a function of an additional one-dimensional variable z, of which u is independent, then the initial data of the new u is supported in an infinite cylinder in x, z space, and hence is not supported in a bounded domain.

We show, at the end of the introduction, that Theorem 3 follows from the following theorem.

THEOREM 6 (trace identity). *Suppose $n \geq 3$, n odd, $\rho > 0$, $f_i \in C_0^\infty(\overline{B_\rho(0)})$, and u_i is the solution of the IVP* (1.3) *and* (1.4) *for $f = f_i$, $i = 1, 2$. Then we have the identities*

$$\frac{1}{2}\int_{R^n} f_1(x)\, f_2(x)\, dx = \frac{-1}{\rho}\int_0^\infty \int_{|p|=\rho} t\, u_1(p,t)\; u_{2tt}(p,t)\, dS_p\, dt, \tag{1.6}$$

$$\frac{1}{2}\int_{R^n} f_1(x)\, f_2(x)\, dx = \frac{1}{\rho}\int_0^\infty \int_{|p|=\rho} t\, u_{1t}(p,t)\; u_{2t}(p,t)\, dS_p\, dt. \tag{1.7}$$

Note that (1.6) is not symmetric, so it clearly implies another similar identity. Some other interesting consequences of the nonsymmetry will be addressed elsewhere.

We do not have an inversion formula similar to the one in Theorem 3 or Theorem 6 for even dimensions. If we can prove an inversion formula or an identity of the above type for the $n = 2$ case, even when f is spherically symmetric, then we feel that the techniques used in the proof of Theorem 6 would carry over to a proof for all n even and all f (not just spherically symmetric f). However, we do not have an inversion formula in the $n = 2$ case even when f is spherically symmetric.

The identity in Theorem 3 is a step towards identifying the range of the map $f \to (Mf)(p,r)$ in the case when D is a ball in R^n, n odd. The other theorems do not attempt to specify the range of this map for the general case. The identity in Theorem 6 has important implications for optimal regularity of traces of solutions of hyperbolic partial differential equations whose principal part is the wave operator. Some of this may be seen in the proof of Theorem 6, but the general trace regularity results and their proofs will be given in [11].

Theorems 4 and 5 (and hence Theorems 1 and 2) are valid under a slightly weaker hypothesis. Let D be a bounded, open, connected subset of R^n with a smooth boundary, and U be the unbounded component of $R^n \setminus \overline{D}$ (note that $\partial U \subset S$). Then, for Theorem 4, we may replace the hypothesis that $\overline{D}$ be strictly convex by the hypothesis that $R^n \setminus U$ be strictly convex, and Γ must be a relatively open subset of ∂U

(instead of S). For Theorem 5, the reconstruction requires knowing $u(p,t)$ for all $t \in [0, diam(D)/2]$ and for all p in ∂U (instead of all p in S). This may be seen by applying Theorems 4 and 5 to the same function f, but over the region $R^n \setminus U$ instead of $\overline{D}$, because we are given that $f = 0$ on the bounded components of $R^n \setminus \overline{D}$.

The recovery of a function from its mean values over spheres centered on some surface or other families of surfaces has been studied by many authors. John [15] is a good source for the early work on recovering a function from its mean values over a family of spheres with centers on a plane. A very interesting theoretical analysis of the problem with centers restricted to a plane was provided by Bukhgeim and Kardakov in [5]; see also the work of Fawcett [9] and Andersson [4] for additional results for this problem. The difficult problem of recovering a function from integrals over a fairly general family of surfaces has also been studied; see [22] and [23] and the references there. The results in our article, for the very specialized family of surfaces we consider, are stronger.

Cormack and Quinto in [6] and Yagle in [35] studied the recovery of f from the mean values of f over spheres passing through a fixed point. Volchkov in [31] studied the injectivity issue in the problem of recovering a function from its mean values over a family of spheres. He characterizes injectivity sets which have a spherical symmetry, so these results do not cover the injectivity result in Theorem 1. Using techniques from $\mathcal{D}$-module theory, Goncharov in [13] finds explicit inversion formulas for the spherical mean value transform operator restricted to some n-dimensional varieties of spheres in R^n. The variety of spheres tangent to a hypersurface is included, but our interest, the family of spheres centered on a hypersurface, is not.

Agranovsky and Quinto in [1], [2] have proved several significant uniqueness results for the spherical mean transform, and applied them to related questions such as stationary sets for solutions of the wave equation. In [1] they give a complete characterization of sets of uniqueness (sets of centers) for the spherical mean transform on compactly supported functions in the plane, i.e., without assumption on the location of the support with respect to the set of centers. In [21] there is an announcement of a uniqueness theorem more general than our Theorem 1, which can be proved using techniques from microlocal analysis in the analytic category as exposed in section 3 of [2]. We think that our proof is still interesting. We use domain of dependence arguments and unique continuation for the time-like Cauchy problem to prove Theorem 4, and hence Theorem 1. Since the domain of dependence result and the unique continuation result for the time-like Cauchy problem are valid for very general hyperbolic operators (with coefficients independent of t), our proof of Theorem 4 is actually valid if the wave operator is replaced by a first order perturbation with coefficients which are C^1 and independent of t. Our technique may perhaps extend to solutions of more general hyperbolic operators with nonconstant reasonably smooth coefficients, whereas the methods in [2], [21] would carry over, at most, to operators with analytic coefficients.

Theorem 3 in [3] also addresses a question similar to the one dealt with in Theorem 1. There, they are interested in the uniqueness question when the mean values of f are known for all spheres centered on the boundary, but they do not require that f be supported inside the region D. They show uniqueness holds if $f \in L^q(R^n)$ as long as $q \leq 2n/(n-1)$. The theorem fails for $q > 2n/(n-1)$.

Norton in [26] derived an explicit inversion formula for the $n = 2$ case of the problem discussed in Theorem 3, using an expansion in Bessel functions. The inversion formula needs further analysis to analyze the effect of the zeros of Bessel functions

used in the formula. Norton and Linzer in [27] considered the recovery of f (supported in a ball in R^3) from the mean values of f over all spheres centered on the boundary of the ball, again by using a harmonic decomposition. They also related it to the solution of the wave equation and then transferred the problem to the frequency domain by taking the Fourier transform of the time variable. There they provide an inversion formula in the form of an integral operator whose kernel is given by an infinite sum. Then they truncated this sum to obtain an approximate inversion formula. They did not deal with the higher dimensional case. Our exact inversion formula, which is valid in all odd dimensions, seems to have a cleaner closed form. The recent articles [32], [33], [34] use the work of Norton and Linzer [27] for reconstruction in TCT.

After the presentation of some of the results of this paper at Oberwolfach, A. G. Ramm informed us that he could also invert the spherical mean transform with centers on some surfaces and sent us the preprint [28]. For the problem of inversion when centers lie on a sphere, he gives a series method whose details are given for dimension $n = 3$. In that case, his result can already be found in formulas (52) and (56) of Norton and Linzer [27]. He also establishes a uniqueness theorem whose strength in relation to prior results is not fully clear, but it does not contain our Theorem 1, for example.

We conclude the introduction by showing how Theorem 3 follows from Theorem 6. For n odd, $n \geq 3$, from page 682 of [7], we have a more convenient representation of $u(x,t)$ in terms of $(Mf)(x,r)$ than the one given earlier. We have

$$u(x,t) = \frac{\sqrt{\pi}}{2\,\Gamma(n/2)} \left(\frac{1}{2t}\frac{\partial}{\partial t}\right)^{(n-3)/2} \left(t^{n-2}(Mf)(x,t)\right) = \frac{\sqrt{\pi}}{2\,\Gamma(n/2)} \mathcal{D}\mathcal{N}(f)(x,t). \tag{1.8}$$

Hence, for all $f_1, f_2 \in C_0^\infty(\overline{B_\rho(0)})$, (1.6) and (1.7) may be rewritten as

$$\langle f_1, f_2\rangle = \frac{-\pi}{2\,\rho\,\Gamma(n/2)^2}\,\langle t\,\mathcal{D}\mathcal{N} f_1\,,\,\partial_t^2 \mathcal{D}\mathcal{N} f_2\rangle \;=\; \frac{-\pi}{2\,\rho\,\Gamma(n/2)^2}\,\langle \mathcal{N}^*\,\mathcal{D}^*\,\partial_t^2\,t\,\mathcal{D}\mathcal{N} f_1\,,\,f_2\rangle,$$

$$\langle f_1, f_2\rangle = \frac{\pi}{2\,\rho\,\Gamma(n/2)^2}\,\langle t\,\partial_t\mathcal{D}\mathcal{N} f_1\,,\,\partial_t \mathcal{D}\mathcal{N} f_2\rangle \;=\; \frac{-\pi}{2\,\rho\,\Gamma(n/2)^2}\,\langle \mathcal{N}^*\,\mathcal{D}^*\,\partial_t\,t\,\partial_t\,\mathcal{D}\mathcal{N} f_1\,,\,f_2\rangle.$$

We have an additional identity which comes from the observation that if u is a solution of (1.3), (1.4), then u_{tt} is also a solution of (1.3) but with the initial conditions $u_{tt}(.,t{=}0) = 0$ and $u_{ttt}(.,t{=}0) = \Delta f$. Hence (1.6) also implies

$$\langle f_1, f_2\rangle \;=\; \frac{-\pi}{2\,\rho\,\Gamma(n/2)^2}\,\langle t\,\mathcal{D}\mathcal{N} f_1\,,\,\mathcal{D}\mathcal{N}\Delta f_2\rangle \;=\; \frac{-\pi}{2\,\rho\,\Gamma(n/2)^2}\,\langle \Delta\mathcal{N}^*\,\mathcal{D}^*\,t\,\mathcal{D}\mathcal{N} f_1\,,\,f_2\rangle.$$

These give us the three inversion formulas of Theorem 3.

2. Proof Of Theorem 4. We will need three results in the proof of Theorem 4.

2.1. Unique continuation for time-like surfaces. The first result concerns unique continuation for the time-like Cauchy problem for the wave equation.

PROPOSITION 1. *If u is a distribution and satisfies* (1.3) *and u is zero on $B_\epsilon(p)\times(-T,T)$ for some $\epsilon > 0$, and $p \in R^n$, then u is zero on (see Figure* 2.1*)*

$$\{\ (x,t)\ :\ |x-p| + |t| < T\ \},$$

and in particular on

$$\{\ (x,t{=}0)\ :\ |x-p| < T\ \}\ .$$

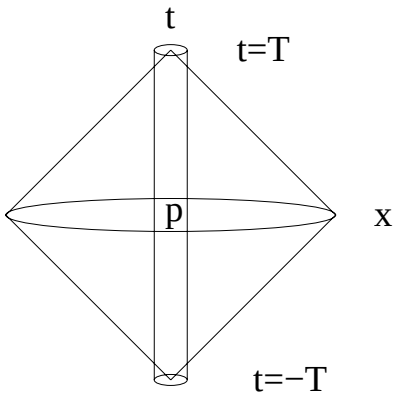


FIG. 2.1. *John's theorem.*

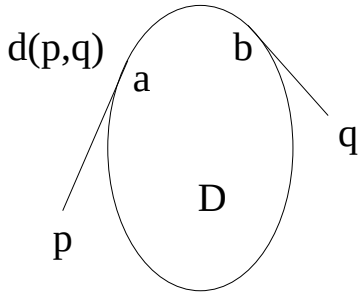


FIG. 2.2. *Shortest path between p and q.*

Proof of Proposition 1. Proposition 1 follows quickly from Theorem 8.6.8 in [14], which itself is derived from Holmgren's theorem. In that theorem, take $X_2 = \{\ (x,t)\ : |x-p|+|t| < T\ \}$ and $X_1 = X_2 \cap (B_\epsilon(p) \times (-T,T))$, and note that any characteristic hyperplane through a point in X_2 has the form $(x-x_0)\cdot\theta + (t-t_0) = 0$ for some unit vector θ and some point $(x_0,t_0) \in X_2$. This plane cuts the vertical line $x = p$ (in (x,t) space) at the point (p,t), where $t = t_0 - (p-x_0)\cdot\theta$, and hence $|t| \le |t_0| + |p-x_0| < T$. □

While Proposition 1 is well known in certain circles, we did not find a ready reference for the proof and so have included the proof above. The proposition was generalized by Robbiano and Hormander to apply to hyperbolic operators with coefficients independent of t, but the generalization was not as sharp as Proposition 1. The definitive form, due to Tataru in [30], includes Proposition 1 as a special case. The proof of Tataru's result is quite complicated, but for Theorem 4 we need only the special case above. The possible extension of Theorem 4 to more general hyperbolic operators would require the full strength of Tataru's result.

2.2. Domain of dependence for exterior problems. Let D be a bounded open subset of R^n with a smooth boundary S. For points p, q outside D, let $d(p,q)$ denote the infimum of the lengths of all the piecewise C^1 paths in $R^n \setminus D$ joining p to q (see Figure 2.2). Using ideas in Chapter 6 of [24] (where it is applied to the distance function generated by a Riemannian metric), one may show that $d(p,q)$ is a topological metric on $R^n \setminus D$.

For any point p in $R^n \setminus D$ and any positive number r, define $E_r(p)$ to be the p centered ball of radius r in $R^n \setminus D$ under this metric, that is,

$$E_r(p) = \{\ x \in R^n \setminus D\ :\ d(x,p) < r\ \}.$$

The second result we need in the proof of Theorem 4 is about the domain of dependence of solutions of the wave equation on an exterior domain. Loosely speaking, the result claims that the value of the solution of the wave equation in an exterior domain, at a point (x,s), affects the value of the solution at the point (y,t) only if $d(x,y) \le t-s$.

PROPOSITION 2 (domain of dependence). *Suppose D is a bounded, connected, open subset of R^n with a smooth boundary S. Suppose u is a smooth solution of the exterior problem*

$$u_{tt} - \Delta u = 0, \qquad x \in R^n \setminus D,\ t \in R,$$

$$u = h \qquad \text{on } S \times R\ .$$

Suppose p is not in D, and $t_0 < t_1$ are real numbers. If $u(.,t_0)$ and $u_t(.,t_0)$ are zero on $E_{t_1-t_0}(p)$ and h is zero on

$$\{(x,t)\ :\ x \in S,\ t_0 \le t \le t_1,\ d(x,p) \le t_1 - t\ \},$$

then $u(p,t)$ and $u_t(p,t)$ are zero for all $t \in [t_0,t_1)$.

In textbooks one may find proofs of this result when $D = \emptyset$ (in which case $d(x,y) = |x-y|$), whereas we are interested in the result for solutions in exterior domains. While the method of attack for proving such a result is clear enough, the details are complicated by the fact that the map $x \to d(x,p)$ is not a smooth map and hence one has to appeal to a more general version of the divergence theorem.

Proof. To prove Proposition 2 we first show that for any p outside D, the function $x \to d(x,p)$ is a locally Lipschitz function on $R^n \setminus D$; that is, for every point $q \in R^n \setminus D$, there is a ball $B_\rho(q)$ such that

$$|d(x,p) - d(y,p)| \le C|x-y| \qquad \forall x,y \in B_\rho(q) \setminus D$$

with C independent of x and y. From the triangle inequality,

$$|d(x,p) - d(y,p)| \le d(x,y),$$

so the Lipschitz nature will follow if we can show that, for every $q \in R^n \setminus D$, there is a ball $B_\rho(q)$ such that

$$d(x,y) \le C|x-y| \qquad \forall\ x,y \in B_\rho(q) \setminus D, \tag{2.1}$$

with C independent of x,y. We give the proof below.

If q is not on the boundary of D, then we can find a ball $B_\rho(q)$ contained in $R^n \setminus D$ and hence $d(x,y) = |x-y|$ for all $x,y \in B_\rho(q)$. So the challenge is to prove (2.1) for $q \in S$. We give a proof of (2.1) in the $n=3$ case (the general case is very similar, but the notation gets a little cumbersome). For $q \in S$, without loss of generality, we may find a ball $B_\rho(q)$ so that

$$D \cap B_\rho(q) = \{u = (u_1,u_2,u_3)\ :\ u_3 > \phi(u_1,u_2),\ u \in B_\rho(q)\}$$

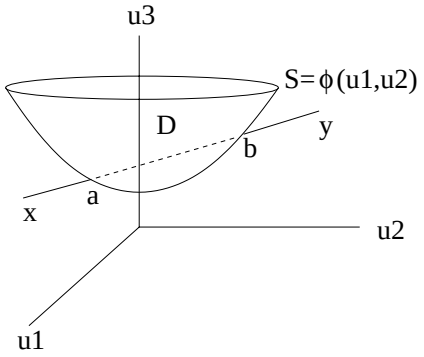


FIG. 2.3. *Lipschitz estimate.*

for some smooth function $\phi(u_1, u_2)$. Now if $x, y \in B_\rho(q) \setminus D$, then $x_3 \le \phi(x_1, x_2)$, $y_3 \le \phi(y_1, y_2)$. If the line segment xy does not intersect D, then $d(x, y) = |x - y|$ and (2.1) is valid. So assume that the segment xy enters $\overline{D}$ at a and leaves $\overline{D}$ the last time at b (see Figure 2.3). Then

$$d(x, y) \le |x - a| + d(a, b) + |b - y| \le |x - y| + d(a, b) + |x - y| = 2|x - y| + d(a, b).$$

So if we could prove $d(a, b) \le C|a - b|$ for a, b on the boundary of D, then (2.1) would follow because $|a - b| \le |x - y|$. So let us estimate $d(a, b)$. From its definition, $d(a, b)$ is not larger than the length of the projection of the line segment ab onto S. Now the projection of the segment ab onto S is

$$\begin{aligned} s \to \mathbf{r}(s) = (1 - s)[a_1, a_2, 0] + s[b_1, b_2, 0] \\ +[0, 0, \phi((1 - s)a_1 + sb_1, (1 - s)a_2 + sb_2)], \qquad 0 \le s \le 1. \end{aligned}$$

Hence

$$|\mathbf{r}'(s)| = |\,[b_1 - a_1, b_2 - a_2, \phi_1(.,.)(b_1 - a_1) + \phi_2(.,.)(b_2 - a_2)]\,| \le C|b - a|$$

because the partial derivatives of ϕ are bounded on S. Hence the length of the projection is no more than $C|b - a|$. This proves (2.1) for $x, y \in B_\rho(q)$. □

Since the map $x \to d(x, p)$ is Lipschitz, from Rademacher's theorem (see [8]), $d(x, p)$ is differentiable almost everywhere in $R^n \setminus D$. Let us estimate $|\nabla_x d(x, p)|$ (if it exists) for $x \notin \overline{D}$. For x not in $\overline{D}$, there is a ball around x which does not intersect $\overline{D}$, and hence for any y in this ball we have $d(x, y) = |x - y|$. Since $d(x, p)$ is a metric, for any y in this ball $|d(y, p) - d(x, p)| \le d(x, y) = |x - y|$, and hence

$$\frac{|d(y, p) - d(x, p)|}{|y - x|} \le 1$$

for all y in the ball. Hence the directional derivative of $d(x, p)$ at x, in any direction, does not exceed 1, and hence $|\nabla_x d(x, p)| \le 1$ for all x not in $\overline{D}$, where it exists. Actually, we believe $|\nabla d(x, p)| = 1$ almost everywhere, but we do not need this.

Proof of Proposition 2. For any real number τ in (t_0, t_1), choose $\epsilon > 0$ so that $\tau + \epsilon < t_1$. Let $\mathcal{K}$ be the backward "conical" surface $t = \tau + \epsilon - d(x, p)$ in x, t space with vertex $(p, \tau + \epsilon)$, defined by $d(p, q)$. Specifically

$$\mathcal{K} = \{\ (x, \tau + \epsilon - d(x, p))\ :\ x \in R^n \setminus D\ \}.$$

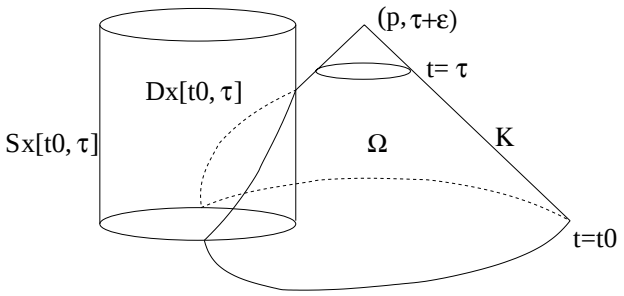


FIG. 2.4. *Domain of integration.*

Since $d(x,p)$ is a Lipschitz function, $\mathcal{K}$ is a Lipschitz surface and so has a normal almost everywhere. For points of K corresponding to x not in S, the upward pointing normal (where it exists) will be parallel to $(\nabla_x d(x,p), 1)$. So if (ν_x, ν_t) is the upward pointing unit normal to $\mathcal{K}$, then

$$|\nu_x| = \frac{|\nabla_x d(x,p)|}{\sqrt{1+|\nabla_x d(x,p)|^2}} \leq \frac{1}{\sqrt{1+|\nabla_x d(x,p)|^2}} = \nu_t.$$

To prove the domain of dependence result we imitate the proof used for such a result if the domain were the whole space. We will perform an integration over the region Ω (which is the subset of $(R^n \setminus D) \times R$) enclosed by the planes $t=\tau$, $t=t_0$, the surface $S \times [t_0, \tau]$, and the backward cone $\mathcal{K}$ (see Figure 2.4). Since Ω need not be a domain with a smooth (or even C^1) boundary, we will appeal to a generalization of the divergence theorem—the generalized Gauss–Green theorem of Federer. Please see the appendix and [8] for definitions and the statement of the results below. Let Φ be the n-dimensional Hausdorff measure on R^{n+1} (which is a regular Borel measure on R^{n+1}), and let $(\nu_x, \nu_t) = \nu = \nu(\Omega, x, t)$ be the generalized outward pointing unit normal at (x,t) associated with the region Ω.

We have

$$(u_t^2 + |\nabla u|^2)_t - 2\nabla \cdot (u_t \nabla u) = 2u_t(u_{tt} - \Delta u) = 0 \qquad \text{in } \Omega.$$

Hence from the Gauss–Green theorem (Proposition 6 in the appendix)

$$0 = \int_\Omega (u_t^2 + |\nabla u|^2)\nu_t - 2u_t \nabla u \cdot \nu_x \; d\Phi. \tag{2.2}$$

Note that to apply the Gauss–Green theorem we must make sure that the Φ measure of $\partial\Omega$ is finite. However, $\partial\Omega$ is a subset of the union of bounded parts of the surfaces of $t=t_0$, $t=\tau$, $\mathcal{K}$, and $S \times [t_0, \tau]$, and the Φ measure of these surfaces equals their surface area (for Lipschitz surfaces), and all these surface areas are finite (including $\mathcal{K} : t = \tau + \epsilon - d(x,p)$ because $|\nabla_x d(x,p)| \leq 1$). Note that all the sets entering our discussion are Borel sets.

Now, by definition (see the appendix), $\nu(x,t) = 0$ at all interior points of Ω and Ω^c. So we need $\nu(\Omega, x, t)$ for points (x,t) on the boundary of Ω. The boundary of Ω consists of a part coming from $t=\tau$, a part coming from $t=\tau_0$, a part from $\mathcal{K}$, and a part from $S \times [t_0, \tau]$. The generalized normal agrees with the usual normal to surfaces at points where the boundary is Lipschitz (so where it is smooth). The boundary is Lipschitz at all those points which lie on only one of the bounding surfaces: the

difficulties arise at points where one or more bounding surfaces of Ω meet. Hence at most points of $\partial\Omega$, on $t = \tau$ we have $\nu = (0, 1)$, on $t = t_0$ we have $\nu = (0, -1)$, on $S \times [t_0, \tau]$ we have $\nu = (\nu_x, 0)$, and on $\mathcal{K}$ we have $\nu = (\nu_x, \nu_t)$ with $|\nu_x| \leq \nu_t$.

Now we must deal with the boundary points which lie on the intersection of two or more of the bounding surfaces of Ω. If we could prove that the Φ measure of this set is zero, then we would not need to know the value of $\nu(\Omega, x, t)$ for points on this set. This is true perhaps if all surfaces were C^1, but we are not sure if this is true if one them is Lipschitz. So we must determine $\nu(\Omega, x, t)$ at these special points on the boundary. From Proposition 5 in the appendix, if the special point lies at the intersection of (two smooth nontangential) surfaces $t = t_0$ or $t = \tau$ with $S \times [t_0, \tau]$, then $\nu = 0$ at that point; if the special point lies at the intersection of $\mathcal{K}$ with $t = t_0$ or $t = \tau$ or $S \times [t_0, \tau]$, then either $\nu = 0$ at that point or ν is the normal at that point to the corresponding smooth surface $t = t_0$ or $t = \tau$ or $S \times [t_0, \tau]$ (as if $\mathcal{K}$ did not play a part).

Now we examine the contribution to the right-hand side (RHS) of (2.2) from the various parts. Based on our description of $\nu(\Omega, x, t)$ in the previous paragraph, we get nonzero contributions, at most, from points on the boundary of Ω. The contribution from the $S \times [t_0, \tau]$ part will be zero because $\nu_t = 0$ on $S \times [t_0, \tau]$ and u, and hence u_t is zero on the part of $\partial\Omega$ on $S \times [t_0, \tau]$. The contribution from the $t = \tau_0$ parts is zero because u and u_t are zero on the part of $\partial\Omega$ on $t = t_0$. The contribution from the $\mathcal{K}$ part of $\partial\Omega$ which is not on any of the other parts is nonnegative because for this part $\nu_t(x, t) \geq |\nu_x(x, t)| \geq 0$ for $x \notin \overline{D}$, and hence the integrand is nonnegative because

$$\begin{aligned}(u_t^2 + |\nabla u|^2)\nu_t - 2u_t \nabla u \cdot \nu_x &\geq (u_t^2 + |\nabla u|^2)\nu_t - 2|u_t||\nabla u||\nu_x| \\ &\geq \nu_t(u_t^2 + |\nabla u|^2 - 2|u_t||\nabla u|) = \nu_t(|u_t| - |\nabla u|)^2.\end{aligned}$$

Hence the contribution from the $t = \tau$ part (which is nonnegative because $\nu_x = 0$ and $\nu_t = 1$ on $t = \tau$) must be zero. Furthermore, the integration is over a region lying above the part of $B_\epsilon(p)$ outside D. Hence $u_t(p, \tau) = 0$ for every $\tau \in [t_0, t_1)$. Also $u(p, t_0) = 0$ by hypothesis; hence $u(p, \tau) = 0$ for all $\tau \in [t_0, t_1)$. □

2.3. Distance computation. The third intermediate result we need is the crucial computation of a certain distance. Below, when we refer to the boundary or the closure of $E_r(p)$, we mean that as a subset of R^n in the topology of R^n and not in the topology induced by $d(p, q)$.

PROPOSITION 3. *Suppose D is a bounded open subset of R^n, $n \geq 2$, S is its smooth boundary, and $\overline{D}$ is strictly convex. Suppose p is a point on S and ρ a small enough positive number less than r. If $K = \overline{D} \setminus B_r(p)$ is not empty, then the shortest distance between K and the closure of $E_\rho(p)$ is the length of some line segment joining a point on $S \cap S_r(p)$ to a point on the boundary of $E_\rho(p)$ which lies on S (see Figure 2.5).*

Proof of Proposition 3. Let δ be the shortest distance between K and the closure of $E_\rho(p)$. Let Γ be the subset of S consisting of points whose geodesic distance from p (on S) is less than or equal to ρ. The boundary of K consists of a part on S, which we call the outer boundary, and the rest (which is on $S_r(p)$), which we call the inner boundary. The boundary of $E_\rho(p)$ consists of a part to the right of the tangent plane to S at p (the part in the region $(x - p) \cdot \nu_p > 0$ where ν_p is the exterior normal to S at p), a part on S which is Γ, and the rest which we denote by C.

It is clear enough that the shortest distance will be the distance between some point on the boundary of K and some point on the boundary of $E_\rho(p)$. Furthermore,

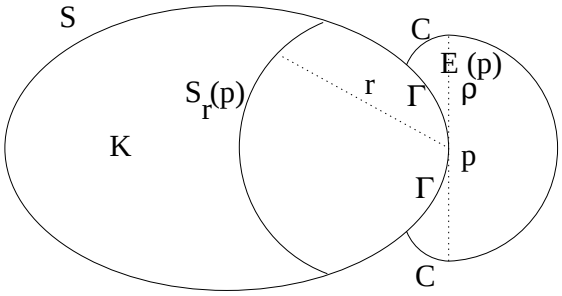


FIG. 2.5. *Distance computation.*

a shortest segment will be normal to the two boundaries provided the boundaries are smooth at the optimal points.

Because of the strict convexity of $\overline{D}$, we can find points p' on S arbitrarily close to p so that the distance between p' and the inner boundary of K is less than r, implying $\delta < r$. In fact pick a point q in the interior of the inner boundary of K so that the segment pq is not normal to S. Then we can find a direction tangential to S so that if p moves in that direction on S, then $|p-q|$ will decrease. So an optimal point on ∂K cannot be on the interior of the inner boundary of K, or else an optimal segment will be normal to $S_r(p)$ and hence would pass through p.

An optimal point on the boundary of $E_\rho(p)$ cannot be to the right of the tangent plane to S at p because then the corresponding optimal line will be normal to $S_\rho(p)$, and hence will pass through p, and then p will be a better candidate than this point.

Next we claim that no point of C is a candidate for an optimal point on the boundary of $E_\rho(p)$, unless it is on Γ. We show this by showing that for any point q on $S \setminus \Gamma$, the point on C closest to q is on Γ.

For ρ small enough, we may parameterize Γ by (s,θ), where s is the geodesic distance from p and θ is a unit vector representing the tangent vector to the geodesic at p. So the surface Γ is

$$x = \gamma(s,\theta), \qquad 0 \le s \le \rho, \quad |\theta| = 1, \ \theta \in R^{n-1},$$

and for each fixed θ the curve $s \to \gamma(s,\theta)$ with $s \in [0,\rho]$ is a geodesic on S and s is the arc length along this geodesic. So $\gamma_{ss}(s,.)$ is normal to S at $\gamma(s,.)$ and $|\gamma_s| = 1$.

Furthermore, because $\overline{D}$ is strictly convex, for ρ small enough, for any point q in the closure of $E_\rho(p)$, the $d(p,q)$ is attained either as the length of the segment pq or the length of a curve consisting of a geodesic on S, starting at p, followed by a line segment from the end point of the geodesic to q which is tangential to the geodesic; see Figure 2.2 (of course p is on S in our case). So, for ρ small enough, C is generated by the family of curves

$$s \to c(s,\theta) = \gamma(s,\theta) + (\rho - s)\gamma_s(s,\theta), \qquad 0 \le s \le \rho,$$

as θ ranges over the unit sphere in R^{n-1}.

Let us examine the distance between q and points on one of the generating curves of C. Define $h(s) = |c(s,.) - q|^2$—the square of the distance between q and a point

on a generating curve. Then, for $0 \le s \le \rho$,

$$\begin{aligned} h'(s) &= 2(c(s,.) - q) \cdot c_s(s,.) \\ &= 2(\gamma(s,.) + (\rho - s)\gamma_s(s,.) - q) \cdot \gamma_{ss}(s,.)(\rho - s) \\ &= 2(\rho - s)(\gamma(s,.) - q) \cdot \gamma_{ss}(s,.). \end{aligned}$$

Above we used $\gamma_s \cdot \gamma_{ss} = 0$ because γ_s is tangential to S and γ_{ss} is normal to S. Now $\gamma(s,.)$ is a point on S (actually on Γ) (denote it by a), and $\gamma_{ss}(s,.)$ is the inward pointing normal to S there. Hence the strict convexity of $\overline{D}$ implies (note $q \neq a$)

$$0 < (q - a) \cdot \gamma_{ss} = (q - \gamma(s,.)) \cdot \gamma_{ss}(s,.).$$

Hence $h'(s) < 0$ for $0 \le s < \rho$, and so $h(s)$, on $0 \le s \le \rho$, attains its minimum at $s = \rho$, that is, at the point $c(\rho,.)$, which is $\gamma(\rho,.)$, which lies on Γ. This proves that the point on C closest to a fixed point q on $S \setminus \Gamma$ must be on Γ.

Because D is strictly convex, the normal lines to the exterior boundary of K will have to cross the inner boundary of K before they meet Γ. To see this, suppose the normal line connects a point x on the exterior boundary to a point y on Γ. Then strict convexity of $\overline{D}$ implies that the line segment xy is in D (except for the end points). Now $|x - p| > r$ and $|y - p| < r$, so there is a point z on the line segment xy so that $|z - p| = r$, and hence z is on the inner boundary of K. Hence no interior point of the outer boundary of K can be an optimal point. This completes the proof of Proposition 3. □

2.4. Proof of Theorem 4. We now give the proof of Theorem 4. Without loss of generality, we may assume that there is a point p on S and a small positive real number ρ so that $\Gamma = \overline{E_\rho(p)} \cap S$.

Step 1. Choose any $\epsilon > 0$ smaller than ρ. Let q be any point in the hemisphere $H \cap B_\epsilon(p)$, where H is the region to the right of, and includes, the tangent plane to S at p (see Figure 2.5); so H is the half-space in x-space containing p and not intersecting D. Then $d(p,q) = |p - q| < \epsilon$, and hence from the triangle inequality, for any x in $S \setminus \Gamma$, we have $d(x,q) \ge d(x,p) - d(p,q) \ge \rho - \epsilon$.

Let $u = h$ on $S \times R$. Then u is the solution of the exterior problem

$$u_{tt} - \Delta u = 0, \qquad x \in R^n \setminus D,\ t \in R,$$

$$u(x, t{=}0) = 0,\ u_t(x, t{=}0) = 0, \qquad x \in R^n \setminus D,$$

$$u = h \qquad \text{on } S \times R.$$

Now h is supported in $(S \setminus \Gamma) \times R$ and the distance of q from $S \setminus \Gamma$ is at least $\rho - \epsilon$. Hence from Proposition 2, we have $u(q,t)$ is zero for $|t| < \rho - \epsilon$ for all $q \in H \cap B_\epsilon(p)$.

Now u is the solution of the wave equation on $R^n \times R$, so the previous result combined with Proposition 1 gives that $f(x) = u_t(x, t{=}0)$ is zero on

$$\{\ x \in R^n\ :\ |x - q^*| < \rho - \epsilon\ \}$$

for some (actually all) q^* in the interior of $H \cap B_\epsilon(p)$ for all small $\epsilon > 0$. Since $|x - q^*| \le |x - p| + |p - q^*|$, $f(x)$ is zero on

$$\{\ x \in R^n\ :\ |x - p| < \rho - 2\epsilon\ \}$$

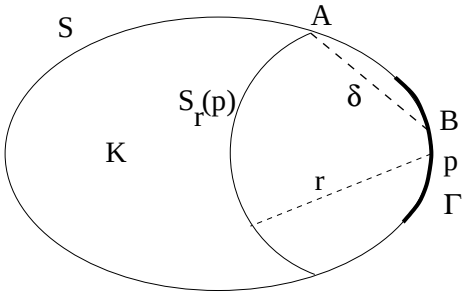


FIG. 2.6. *Triangle inequality.*

for all $\epsilon > 0$. Hence $f(x)$ is zero on

$$\{\ x \in R^n\ :\ |x-p| < \rho\ \}.$$

Step 2. We now show that $f(x)$ is zero for all x. This will follow easily if we can show the following: if $f(x)$ is zero on the region $|x-p| < r$ for some $r \geq \rho$, then $f(x)$ is zero on the region $|x-p| < r+\sigma$, where σ is a positive number independent of r.

Please refer to Figure 2.5 for a geometrical interpretation of the notation below. So suppose f is supported in the region K consisting of the part of $\overline{D}$ outside $B_r(p)$. Let $\delta > 0$ be the straight line distance between $E_\rho(p)$ and K; then we show that f is zero on $B_{\rho+\delta}(p)$. Postponing the proof of this claim, let α be the supremum of the straight line distances between p and points on Γ. Since $\overline{D}$ is strictly convex, from the definition of Γ, we have $\alpha < \rho$. From Proposition 3, δ is the length of the line segment AB for some point A on $S_r(p) \cap S$ and some point B on Γ. Then, using the triangle inequality (see Figure 2.6),

$$\begin{aligned}\rho + \delta = \rho + |AB| &= |AB| + |Bp| + (\rho - |Bp|)\\ &\geq |pA| + (\rho - |Bp|) \geq r + (\rho - \alpha),\end{aligned}$$

and we note that $\rho - \alpha$ is positive and independent of r. Hence Theorem 4 holds.

So it remains to show that if f is supported in $K = \overline{D} - B_r(p)$, then f is zero on $B_{\rho+\delta}(p)$. Since u is the solution of the IVP (1.3), (1.4), and $\delta = dist(E_\rho(p), K)$, the standard domain of dependence argument for IVPs implies that u and u_t are zero on

$$\{(x,t)\ :\ x \in E_\rho(p),\ |t| < \delta\}. \tag{2.3}$$

Fix a small $\epsilon > 0$, $\epsilon < \rho$, and let $q \in B_\epsilon(p) \cap H$; note that $q \in E_\rho(p)$. Now u may be considered as the solution of the initial boundary value problem

$$u_{tt} - \Delta u = 0, \qquad x \in R^n \setminus D,\ t \geq \delta - \epsilon,$$

$$u = f_1,\ u_t = f_2, \qquad \text{on } \{R^n \setminus D\} \times \{t = \delta - \epsilon\},$$

$$u = h \text{ on } S \times [\delta - \epsilon, \infty)$$

for some functions f_1 and f_2. Now f_1 and f_2 are zero on $E_\rho(p)$ (by hypothesis), h is zero on $\Gamma \times [\delta - \epsilon, \infty)$, and $d(q,x) \geq d(p,x) - d(p,q) \geq \rho - \epsilon$ for all $x \in D^c$ which are

not in $E_\rho(p) \cup \Gamma$ (note that Γ is the part of the boundary of $E_\rho(p)$ which lies on S). Hence from Proposition 2, $u(q,t)$ is zero for all $t \in [\delta - \epsilon, \delta - \epsilon + \rho - \epsilon)$.

Since we have already shown that u is zero on (2.3), we have that $u(q,t)$ is zero for all t in $[0, \rho + \delta - 2\epsilon)$. Since u is odd in t we have $u(q,t)$ is zero for all t with $|t| < \rho + \delta - 2\epsilon$, for all $q \in B_\epsilon(p) \cap H$. So, from Proposition 1, $u_t(x,0)$, and hence $f(x)$, is zero on

$$\{\ x\ :\ |x - q^*| < \rho + \delta - 2\epsilon\ \}$$

for all small $\epsilon > 0$ and a (actually any) q^* in the interior of $B_\epsilon(p) \cap H$. Hence $f(x)$ is zero on

$$\{\ x\ :\ |x - p| < \rho + \delta - 3\epsilon\ \}$$

for all $\epsilon > 0$, and hence $f(x)$ is zero on

$$\{\ x\ :\ |x - p| < \rho + \delta\ \}$$

and the theorem is proved.

3. Proof of Theorem 5. Since D is a bounded, open, connected subset of R^n, with a smooth boundary, it follows that the complement of $\overline{D}$ will be a disjoint union of connected open sets called components of $R^n \setminus \overline{D}$. Since D is bounded, only one of the components will be unbounded and the rest of the components will be subsets of a fixed ball in R^n. Then from the smoothness of the boundary of D and compactness, one may show that the number of components is finite, the boundaries of the components are disjoint and subsets of the boundary of D, and the boundaries are smooth.

Part 1. Let $\delta = diam(D)$ and $u = h$ on $S \times [0, \delta/2]$ (h is given to us). Since u is an odd function of t, we extend h as an odd function of t. Below $\partial_\nu u$ will represent the derivative of u on $S \times (-\infty, \infty)$ in the direction of the outward pointing normal to $S \times (-\infty, \infty)$.

Since f is supported in $\overline{D}$, we may consider u as the solution of the exterior problem

$$u_{tt} - \Delta u = 0, \qquad \text{on } (R^n \setminus D) \times [-\delta/2,\ \delta/2],$$

$$u(x, t{=}0) = 0,\ u_t(x, t{=}0) = 0, \qquad x \in R^n \setminus D,$$

$$u = h \text{ (given)} \qquad \text{on } S \times [-\delta/2,\ \delta/2].$$

This initial boundary value problem (IBVP) is well posed, and so one may obtain the value of $\partial_\nu u$ on $S \times [-\delta/2,\ \delta/2]$; this may be done numerically using finite differences (one may assume $u = 0$ for points far away from S without changing the value of $\partial_\nu u$ on $S \times [-\delta/2,\ \delta/2]$) .

Now we have u and $\partial_\nu u$ on $S \times [-\delta/2, \delta/2]$ and we show how we may recover u and u_t over the region $\overline{D} \times \{t{=}-\delta/2\}$. This is done using the Kirchhoff formula, which expresses the value of a solution of the wave equation in a cylindrical (in time) domain, at a point, purely in terms of the value of the solution and its normal derivative, on the intersection of the cylinder with the forward light cone through the point; see Figure 3.1. This may be done only in odd space dimensions as will be seen in the

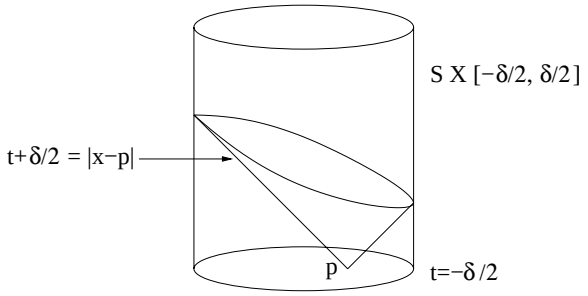


FIG. 3.1. *Kirchhoff formula.*

formal derivation below; the derivation may be made rigorous. A rigorous derivation in the three space dimensional case may be found in [12].

Let $E_+(x,t)$ be the fundamental solution of the wave operator with support in the region $t \geq 0$ (see [14, Chapter VI]). Consider a point $p \in D$. Then

$$\Box E_+(x-p, t+\delta/2) = \delta(x-p, t+\delta/2), \qquad (x,t) \in R^{n+1} \ .$$

Also, note that $E_+(x-p, t+\delta/2)$ is zero for $t < -\delta/2$ and is zero also on $\overline{D} \times (\delta/2, \infty)$ because the support of $E_+(x-p, t+\delta/2)$, *for odd* n, is on the cone $t+\delta/2 = |x-p|$ and $|x-p| \leq \delta$ for any $x \in \overline{D}$. Then, from Green's theorem,

$$\begin{aligned}
u(p, -\delta/2) &= \int_{-\infty}^{\infty} \int_D u(x,t)\, \delta(x-p, t+\delta/2)\, dx\, dt \\
&= \int_{-\infty}^{\infty} \int_D u(x,t)\, \Box E_+(x-p, t+\delta/2)\, dx\, dt \\
&= \int_{-\infty}^{\infty} \int_D \Box u(x,t)\, E_+(x-p, t+\delta/2)\, dx\, dt \\
&\quad + \int_{-\infty}^{\infty} \int_S (\partial_\nu u(x,t)\, E_+(x-p, t+\delta/2) \\
&\qquad\qquad - u(x,t)\, \partial_\nu E_+(x-p, t+\delta/2))\, dS_x\, dt \\
&= \int_{-\delta/2}^{\delta/2} \int_S (\partial_\nu u(x,t)\, E_+(x-p, t+\delta/2) \\
&\qquad\qquad - u(x,t)\, \partial_\nu E_+(x-p, t+\delta/2))\, dS_x\, dt.
\end{aligned}$$

Note that the singular set of E_+ consists of the forward light cone through $(p, -\delta/2)$ and the singular directions (the wave front set) of E_+, away from the vertex of the cone, are the normals to the cone, and so are transverse to $S \times (-\infty, \infty)$, and hence E_+ and $\partial_\nu E_+$ have traces on $S \times (-\infty, \infty)$.

Examining the definition of E_+ in [14, Chapter VI], the last integral may be written in terms of the values of u and $\partial_\nu u$ (and their time derivatives) on $S \times [-\delta/2, \delta/2]$. Hence we now have the values of u on $D \times \{t = -\delta/2\}$; using continuity we can determine the value on $\overline{D} \times \{t = -\delta/2\}$. A similar argument will recover the value of u_t on $\overline{D} \times \{t = -\delta/2\}$.

Knowing u and u_t on $\overline{D} \times \{t = -\delta/2\}$ and that u is the solution of the well-posed

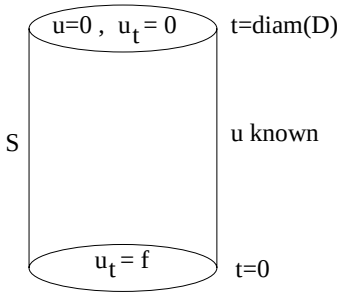


FIG. 3.2. *Backward IBVP.*

IBVP

$$\begin{aligned} u_{tt} - \Delta u = 0 \qquad & \text{on } \overline{D} \times [-\delta/2, 0], \\ u(., t = -\delta/2) = \text{known},\ u_t(., t = -\delta/2) = \text{known} \qquad & \text{on } \overline{D} \\ u = h \text{ (given)} \qquad & \text{on } S \times [-\delta/2, 0]; \end{aligned}$$

we may solve this numerically using finite differences and obtain the value of u_t on $\overline{D} \times \{t = 0\}$ and so obtain f.

Part 2. Again δ represents the diameter of D. If u is known on $S \times [0, \delta]$, then we now give a simpler inversion scheme than the one given above. The problem is the recovery of $u_t(x, 0)$ for $x \in \overline{D}$ from the values of u on $S \times [0, \delta]$.

In odd space dimensions, the domain of dependence of the value of the solution of the wave equation at a point is the sphere of intersection of the backward characteristic cone through that point with $t = t_{init}$. Since u is a smooth solution of the wave equation and the initial data is supported in $\overline{D}$, we have $u(x, t)$ is zero for $t \geq \delta$ and $x \in \overline{D}$. Hence u and u_t are zero on $\overline{D} \times \{t = \delta\}$. Now we may consider u as the solution of the backward IBVP (see Figure 3.2)

$$\begin{aligned} u_{tt} - \Delta u = 0 \qquad & \text{on } \overline{D} \times [0, \delta], \\ u(., t_{=}\delta) = 0,\ u_t(., t_{=}\delta) = 0 \qquad & \text{on } \overline{D}, \\ u = h \qquad & \text{on } S \times [0, \delta]. \end{aligned}$$

This problem is well posed, so given h one may obtain $u_t(x, 0)$ for x in $\overline{D}$ and hence recover f. □

4. Proof of Theorem 6. We first note that (1.7) follows fairly quickly from (1.6) (but not vice versa) as shown next. Noting that (1.7) is symmetric, it is enough to prove its norm form, namely

$$\frac{1}{2} \int_{R^3} |f(x)|^2 \, dx = \frac{1}{\rho} \int_0^\infty \int_{|p|=\rho} t \, |u_t(p, t)|^2 \, dS_p \, dt$$

for all $f \in C_0^\infty(\overline{B_\rho(0)})$. To prove this, we take $f_1 = f_2 = f$ in (1.6). Then, using an integration by parts,

$$\begin{aligned}\frac{1}{2}\int_{R^3} |f(x)|^2\,dx &= \frac{-1}{\rho}\int_0^\infty \int_{|p|=\rho} tu(p,t)\,u_{tt}(p,t)\,dS_p\,dt \\ &= \frac{1}{\rho}\int_0^\infty \int_{|p|=\rho} \{t\,u_t(p,t)\,u_t(p,t) + u(p,t)u_t(p,t)\}\,dS_p\,dt \\ &= \frac{1}{\rho}\int_0^\infty \int_{|p|=\rho} \left\{t\,|u_t(p,t)|^2 + \frac{1}{2}\frac{\partial}{\partial t}(u^2(p,t))\right\} dS_p\,dt \\ &= \frac{1}{\rho}\int_0^\infty \int_{|p|=\rho} t\,|u_t(p,t)|^2\,dS_p\,dt,\end{aligned}$$

where we made use of the fact that $u(p,t{=}0) = f(p) = 0$ for $|p| = \rho$ and that from Huyghen's principle (note that n is odd and $n \geq 3$) $u(p,t) = 0$ for all $t > 2\rho$ and $|p| = \rho$.

To prove (1.6), we will first prove it in the case when $n = 3$, and then we will show (with some effort) that the case for all odd $n \geq 3$ follows from this.

4.1. Proof of trace identity (1.6) when $n = 3$. Part I—An inversion formula. The proof of the three-dimensional case is actually based on proving one of the inversion formulas in Theorem 3 directly, that is without relating it to the wave equation. Note that $\mathcal{D}$ is the identity operator when $n = 3$. We will show that for every $f \in C_0^\infty(\overline{B_\rho(0)})$,

$$f(x) = -\frac{2}{\rho}\Delta\,(\mathcal{N}^*\,t\,\mathcal{N})(f)(x) \qquad \forall x \in B_\rho(0). \tag{4.1}$$

Below, we will make use of the following observation. Suppose $\mathcal{M}$ is an $n-1$ dimensional surface in R^n, given by $\phi(z) = 0$, with $\nabla\phi(z) \neq 0$ at every point of $\mathcal{M}$. Then

$$\int_{\mathcal{M}} h(z)\,dS_z = \int h(z)\,|\nabla\phi(z)|\,\delta(\phi(z))\,dz.$$

We now compute $\mathcal{N}^*(t(\mathcal{N}f)(x))$. We have

$$\begin{aligned}(\mathcal{N}^*(t\mathcal{N}f))(x) &= \frac{1}{4\pi}\int_{|p|=\rho} \frac{1}{|x-p|}\,|x-p|\,(\mathcal{N}f)(p,|x-p|)\,dS_p \\ &= \frac{1}{8\pi^2}\int_{|p|=\rho}\int_{R^3} f(y)\,\delta(|y-p|^2 - |x-p|^2)\,dy\,dS_p \\ &= \frac{1}{8\pi^2}\int_{R^3} f(y)\int_{|p|=\rho} \delta(|y-p|^2 - |x-p|^2)\,dS_p\,dy \\ &= \frac{\rho}{4\pi^2}\int_{R^3} f(y)\int_{R^3} \delta(|y-p|^2 - |x-p|^2)\,\delta(|p|^2 - \rho^2)\,dp\,dy. \qquad (4.3)\end{aligned}$$

The inner integral is an integral on the curve of intersection of the sphere $|p| = \rho$ with the plane of points equidistant from x and y. Define a characteristic function $\chi(x,y)$, for $x \neq y$, which is 1 if the above plane intersects the sphere $|p| = \rho$ in a circle of nonzero radius and zero otherwise.

Let Q be the orthogonal transformation which maps $y - x$ to $|y-x|e_3$, where $e_3 = [0,0,1]$. Then Qx and Qy differ only in the third coordinate and in the fact

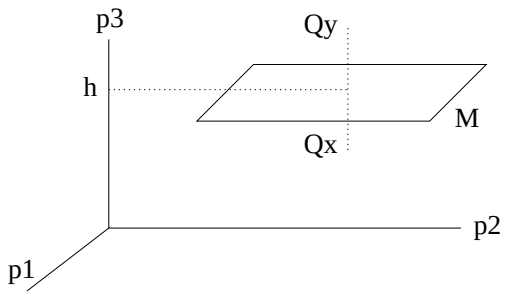


FIG. 4.1. *The plane* $\mathcal{M}$.

that $Qy = Qx + |y - x|e_3$. Then, using an orthogonal change of variables, the inner integral may be rewritten as

$$\begin{aligned}&\int_{R^3} \delta(|y - Q^T p|^2 - |x - Q^T p|^2)\, \delta(|Q^T p|^2 - \rho^2)\, dp \\ &\quad = \int_{R^3} \delta(|Qy - p|^2 - |Qx - p|^2)\, \delta(|p|^2 - \rho^2)\, dp.\end{aligned}$$

Let $\mathcal{M}$ be the plane consisting of points in p space which are equidistant from Qx and Qy (see Figure 4.1). In fact $\mathcal{M}$ is the plane $p_3 = h$, where $h = (Qx)\cdot e_3 + |y-x|/2$. Furthermore

$$\left|\nabla_p(|Qy - p|^2 - |Qx - p|^2)\right| = 2\,|(p - Qy) - (p - Qx)| = 2|Qx - Qy| = 2|x - y|.$$

Then for x, y with $\chi(x, y) = 1$, the inner integral of (4.3) equals

$$\frac{1}{2|x - y|} \int_{\mathcal{M}} \delta(|p|^2 - \rho^2)\, dS_p.$$

Now $\mathcal{M}$ may be parameterized by p_1, p_2; hence the inner integral of (4.3) is

$$\frac{1}{2|x - y|} \int \delta(p_1^2 + p_2^2 + h^2 - \rho^2)\, dp_1\, dp_2.$$

So the integral is really over the circle C centered at the origin with radius $\sqrt{\rho^2 - h^2}$. Now on $p_1^2 + p_2^2 + h^2 = \rho^2$, the magnitude squared of the gradient of $p_1^2 + p_2^2 + h^2 - \rho^2$ is

$$4(p_1^2 + p_2^2) = 4(\rho^2 - h^2).$$

Hence the integral equals

$$\frac{1}{2|x - y|} \int_C \frac{1}{2\sqrt{\rho^2 - h^2}}\, ds = \frac{\pi}{2|x - y|}.$$

Hence

$$(\mathcal{N}^*(t\mathcal{N}f))(x) = \frac{\rho}{8\pi} \int \chi(x, y) \frac{f(y)}{|x - y|}\, dy.$$

The above calculations could be done more rigorously (i.e., without the use of δ functions) with the help of the coarea formula in [8].

Now if x and y are in the open ball $B_\rho(0)$, and $x \neq y$, then $\chi(x,y) = 1$. Hence if f is a smooth function supported in the ball $\overline{B_\rho(0)}$, then

$$(\mathcal{N}^*(t\mathcal{N}f))(x) = \frac{\rho}{8\pi}\int \frac{f(y)}{|x-y|}\,dy \qquad \forall x \in B_\rho(0). \tag{4.4}$$

Hence, taking the Laplacian of both sides, we get

$$\Delta_x(\mathcal{N}^*t\mathcal{N})(f)(x) = \frac{-4\pi\rho}{8\pi}\int f(y)\delta(x-y)\,dy = \frac{-\rho f(x)}{2} \qquad \forall x \in B_\rho(0),$$

which implies

$$f(x) = \frac{-2}{\rho}\Delta_x(\mathcal{N}^*t\mathcal{N})(f)(x), \qquad x \in B_\rho(0) \tag{4.5}$$

for all smooth functions f supported in $\overline{B_\rho(0)}$.

Part II—*The identity.* We now prove (1.6) in Theorem 6 for the $n = 3$ case. For $f_i \in C_0^\infty(\overline{B_\rho(0)})$, $i = 1, 2$, let $u_i(x,t)$ be the solutions of the IVP (1.3), (1.4) with $f = f_i$. Then, from (1.8), $u_i(p,t) = (\mathcal{N}f)(p,t)$ for any $p \in S_\rho(0)$. Furthermore, u_{itt} is also a solution of (1.3) except its initial conditions are

$$u_{itt}(x,0) = \Delta_x u_i(x,0) = 0, \qquad u_{ittt}(x,0) = \Delta_x u_{it}(x,0) = \Delta f_i(x).$$

Hence $\mathcal{N}(\Delta f_i)(p,t) = u_{itt}(p,t)$ for all $p \in S_\rho(0)$ and all $t \in [0,\infty)$.

From (4.5) we have

$$\begin{aligned}
\frac{1}{2}\int_{R^3} f_1(x)\, f_2(x)\, dx &= \frac{-1}{\rho}\langle \Delta(\mathcal{N}^*t\mathcal{N}f_1), f_2\rangle \\
&= \frac{-1}{\rho}\langle t(\mathcal{N}f_1)(p,t), \mathcal{N}(\Delta f_2)(p,t)\rangle \\
&= \frac{-1}{\rho}\int_0^\infty \int_{|p|=\rho} t\,(\mathcal{N}f_1)(p,t)\,\mathcal{N}(\Delta f_2)(p,t)\,dS_p\,dt \\
&= \frac{-1}{\rho}\int_0^\infty \int_{|p|=\rho} t\,u_1(p,t)\,u_{2tt}(p,t)\,dS_p\,dt,
\end{aligned}$$

proving (1.6) for the $n = 3$ case. ∎

4.2. Proof of trace identity (1.6) for all odd $n \geq 3$. Let $\{\phi_m\}_{m=1}^\infty$ be spherical harmonics which form an orthonormal basis for $L^2(S_1(0))$; see Chapter 4 of [29]. These are restrictions to $S_1(0)$ of some harmonic homogeneous polynomials on R^n. If ϕ_m is the restriction of a homogeneous polynomial of degree $k(m)$, then that harmonic homogeneous polynomial is $r^{k(m)}\phi_m(\theta)$, where $r = |x|$ and $\theta = x/|x|$.

Suppose f is a smooth function on R^n supported in $\overline{B_\rho(0)}$. We have a decomposition of f of the form (convergence in L^2)

$$f(r\theta) = \sum_{m=1}^\infty f_m(r)\, r^{k(m)}\, \phi_m(\theta), \qquad r \geq 0,\ |\theta| = 1,$$

with

$$r^{k(m)} f_m(r) = \int_{|\theta|=1} f(r\theta)\,\phi_m(\theta)\,d\theta. \tag{4.6}$$

From the smoothness and support of $f(x)$, we may show[1] that $f_m(r)$ is a smooth, even function on $(-\infty,\infty)$, supported in $[-\rho,\rho]$.

Below we will show that the solution $u(x,t)$, of (1.3), (1.4), will have the form

$$u(x,t)=\sum_{m=1}^{\infty} a_m(r,t)\, r^{k(m)}\phi_m(\theta),$$

where $r=|x|$ and $\theta=x/|x|$. Then, from the orthonormality of $\{\phi_m\}_{m=1}^{\infty}$, the left-hand side (LHS) of the trace identity (1.6) is

$$\begin{aligned}\frac{1}{2}\int_0^{\infty}\int_{|\theta|=1} r^{n-1}\, f_1(r\theta)\, f_2(r\theta)\, d\theta\, dr &= \frac{1}{2}\sum_{m=1}^{\infty}\int_0^{\infty} r^{n-1} r^{2k(m)} f_{1m}(r)\, f_{2m}(r)\, dr\\ &= \frac{1}{2}\sum_{m=1}^{\infty}\int_0^{\infty} r^{\nu(m)-1} f_{1m}(r)\, f_{2m}(r)\, dr,\end{aligned}$$

where $\nu(m)=2k(m)+n$. The RHS of (1.6) is

$$\begin{aligned}RHS &= \frac{-1}{\rho}\int_0^{\infty}\int_{|p|=\rho} t\, u_1(p,t)\, u_{2tt}(p,t)\, dS_p\, dt\\ &= \frac{-1}{\rho}\sum_{m=1}^{\infty}\sum_{l=1}^{\infty}\int_0^{\infty}\int_{|p|=\rho} t\, a_{1m}(\rho,t)\, a_{2ltt}(\rho,t)\, \rho^{k(m)+k(l)}\, \phi_m(p/|p|)\, \phi_l(p/|p|)\, dS_p\, dt\\ &= -\sum_{m=1}^{\infty}\sum_{l=1}^{\infty}\rho^{k(m)+k(l)+n-2}\int_0^{\infty} t\, a_{1m}(\rho,t)\, a_{2ltt}(\rho,t)\, dt\int_{|\theta|=1}\phi_m(\theta)\,\phi_l(\theta)\, d\theta\\ &= -\sum_{m=1}^{\infty}\rho^{\nu(m)-2}\int_0^{\infty} t\, a_{1m}(\rho,t)\, a_{2mtt}(\rho,t)\, dt.\end{aligned}$$

So, to prove (1.6), it would be enough to prove the following: if $f_i(x)$ have the form $g_i(r)r^k\phi(\theta)$, where $g_i(r)$ are smooth, even functions of r, supported in $[-\rho,\rho]$, and $\phi(x)$ is a homogeneous harmonic polynomial on R^n of some degree k with the L^2 norm of ϕ on $S_1(0)$ equal to 1, then the solution $u_i(x,t)$ has the form $a_i(r,t)r^k\phi(\theta)$ and

$$\frac{1}{2}\int_0^{\infty} r^{\nu-1} g_1(r)\, g_2(r)\, dr = -\,\rho^{\nu-2}\int_0^{\infty} t\, a_1(\rho,t)\, a_{2tt}(\rho,t)\, dt, \tag{4.7}$$

where $\nu=n+2k$. Note that the RHS of (4.7) depends on ρ while the LHS does not seem to; but we assumed that the g_i were supported in $[-\rho,\rho]$.

[1] One may show that all derivatives of the function

$$r\to\int_{|\theta|=1} f(r\theta)\,\phi_m(\theta)\, d\theta$$

up to order $k(m)-1$ are zero at $r=0$ because these derivatives at $r=0$ will be sums of terms of the form

$$\int_{|\theta|=1}\theta^{\alpha}\,\phi_m(\theta)\, d\theta, \qquad |\alpha|<k(m),$$

and ϕ_m is orthogonal to all polynomials of degree less than $k(m)$ on the unit sphere (Theorem 2.1 and Corollary 2.4 of Chapter IV in [29]).

Since $r^k\phi(\theta)$ is harmonic, if Δ_S is the Laplace–Beltrami operator on $S_1(0)$, then, noting that

$$\Delta = \partial_r^2 + \frac{n-1}{r}\partial_r + \frac{1}{r^2}\Delta_S,$$

one may show that

$$\Delta_S\phi = -k(k+n-2)\phi \qquad \text{on } S_1(0). \tag{4.8}$$

When $f = g(r)r^k\phi(\theta)$, we seek a solution of (1.3), (1.4) of the form $u(x,t) = a(r,t)r^k\phi(\theta)$. Noting that

$$\begin{aligned}(ar^k)_r &= r^k a_r + kr^{k-1}a,\\ (ar^k)_{rr} &= r^k a_{rr} + 2kr^{k-1}a_r + k(k-1)ar^{k-2},\end{aligned}$$

if we substitute $u = a(r,t)r^k\phi(\theta)$ in (1.3) and use (4.8), we have

$$\begin{aligned}0 &= \left((ar^k)_{tt} - (ar^k)_{rr} - \frac{n-1}{r}(ar^k)_r\right)\phi_m - \frac{ar^k}{r^2}\Delta_S\phi_m\\ &= \phi_m r^k\left(a_{tt} - a_{rr} - \frac{n+2k-1}{r}a_r\right).\end{aligned}$$

Hence $a(r,t)$ must satisfy (here $\nu = n+2k$)

$$a_{tt} - a_{rr} - \frac{\nu-1}{r}a_r = 0, \qquad r \in (-\infty,\infty),\ t \geq 0, \tag{4.9}$$

$$a(.,t{=}0) = 0,\ a_t(.,t{=}0) = g. \tag{4.10}$$

This is an IVP for the Darboux equation, which is well posed and has an explicit solution given on page 700 of [7]. Essentially, one may use a method of descent to reduce the problem to the cases $\nu = 2$ and $\nu = 3$ by noting that a_r/r also satisfies (4.9) and (4.10), except with ν replaced by $\nu+2$ and g replaced by g_r/r.

Now if n is odd, then $\nu = n+2k$ is odd. So the goal is to show that for all odd $\nu = 3,5,\ldots,$ and all $g_i(r)$ which are smooth, even, and supported in $[-\rho,\rho]$, we have

$$\frac{1}{2}\int_0^\infty r^{\nu-1}\,g_1(r)\,g_2(r)\,dr = r^{\nu-2}\int_0^\infty t\,a_1(r,t)\,a_{2tt}(r,t)\,dt \qquad \forall r \geq \rho, \tag{4.11}$$

where $a_i(r,t)$, $i = 1,2$ are the solution of (4.9), (4.10) with $g = g_i$.

Now we have proved (1.6) for $n = 3$ and hence we have proved (4.11) for all $\nu = 3+2k$ with $k = 0,1,2,\ldots$. Hence, we have already proved (4.11) for all odd ν, $\nu \geq 3$ (note (4.9) depends on ν and not on n directly). So we have completed the proof of (1.6).

Remark. Another possible approach to proving (4.11) without first proving (1.6) for the $n = 3$ case is to first verify (4.11) for $\nu = 3$ (it is easy to write the explicit solution of (4.9) when $\nu = 3$), and then use a method of descent by observing that a_r/r also solves (4.9) except with ν replaced by $\nu+2$. So far, we have been unable to use the method of descent to prove (4.11). We were able to prove a symmetric version of this relation using the method of descent, and while the nonsymmetric version easily implies the symmetric version, the validity of the converse is not known.

Appendix. The material is based on [8] and [10] and is included here for the reader's convenience; just Proposition 5 is new.

We give a definition of the $m-1$ dimensional Hausdorff measure on R^m. For a subset S of R^m, define

$$\gamma(S) = vol(m-1)\ (diam(S)/2)^{m-1},$$

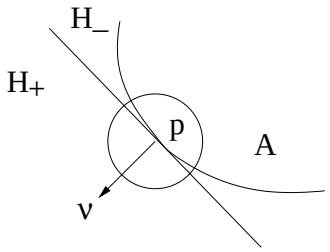
where $vol(m-1)$ is the volume of the $m-1$ dimensional unit ball. So if S were the intersection of a ball in R^m with a hyperplane, then $\gamma(S)$ would be its surface area. For any positive δ, define

$$\phi_\delta(S) = \inf_{\mathcal{F}} \sum_{U\in\mathcal{F}} \gamma(U),$$

where $\mathcal{F}$ is a countable open cover of S, with each set in $\mathcal{F}$ having diameter less than δ. Now $\phi_\delta(S)$ is a decreasing function of δ (the larger the δ the greater the number of admissible open covers and hence the smaller the infimum), so we may define

$$\Phi(S) = \lim_{\delta\to 0^+} \phi_\delta(S) = \sup_{\delta>0} \phi_\delta(S).$$

It is shown in [8] that Φ is an outer measure, the σ algebra of all Borel subsets of R^m are measurable in this outer measure, and Φ is regular. Furthermore, if a surface S is the graph of a smooth function from an open subset of R^{m-1} to R, then $\Phi(S)$ equals the usual surface area of S (Section 3.3.4 in [8]). So the Hausdorff measure generalizes the notion of surface area to Borel subsets of R^m.



Next we define the exterior normal for any subset of R^m. For a point $p \in R^m$ and a unit vector ν we define the half-planes

$$H_+(p,\nu) = \{\, x \in R^m \,:\, (x-p)\cdot\nu > 0 \,\}, \qquad H_-(p,\nu) = \{\, x \in R^m \,:\, (x-p)\cdot\nu < 0 \,\}.$$

Suppose A is a subset of R^m and p a point in R^m. A unit vector ν is defined to be an exterior normal to A at p if

$$\lim_{r\to 0^+} r^{-m}\ |A^c \cap H_-(p,\nu) \cap B_r(p)| = 0,$$
$$\lim_{r\to 0^+} r^{-m}\ |A \cap H_+(p,\nu) \cap B_r(p)| = 0.$$

Here $|\ |$ is the Lebesgue measure on R^m. It is shown in [10] that if such a unit vector exists (for a given A and p) then it is unique. We denote this unit vector by $\nu(A,p)$. If no such unit vector exists then we set $\nu(A,p) = 0$.

PROPOSITION 4. *Below, a vector $x \in R^m$ will be occasionally written as $x = [x', x_m]$.*

- *If A is a subset of R^m, then $\nu(A,p)$ is zero if p is in the interior of A or $R^m \setminus A$.*
- *If $p \in \partial A$ and for some $\rho > 0$,*

$$A \cap B_\rho(p) = \{x \in B_\rho(p) \,:\, x_m > f(x')\}$$

for some C^2 function $f(x')$ of $m-1$ variables, then $\nu(A,p)$ is a positive multiple of $[\nabla f(p'), -1]$.
- *Under the conditions in the second item, for any unit vector $\theta \neq \nu(A,p)$, there is a $c > 0$ so that for r small enough,*

$$r^{-m}\ |A^c \cap H_-(p,\theta) \cap B_r(p)| > c,$$
$$r^{-m}\ |A \cap H_+(p,\theta) \cap B_r(p)| > c.$$

So, the second item asserts that $\nu(A,x)$ extends the notion of an outward pointing unit normal to arbitrary subsets of R^n. In the third item, if $\theta \neq \nu(A,p)$, then the definition of $\nu(A,p)$ implies that at least one of the limits will be nonzero; our claim is that both of them are nonzero for A with C^2 boundary.

Proof of first item. If p is an interior point of A, then for r small enough and any unit vector θ,

$$r^{-m}\ |A \cap H_+(p,\theta) \cap B_r(p)| = r^{-m}\ |H_+(p,\theta) \cap B_r(p)| = vol(m)/2 \ > \ 0,$$

and hence θ cannot be $\nu(A,p)$. A similar argument works if p is an interior point of A^c. □

Proof of second and third items. We will prove the result in the case $m = 2$; the general case is very similar. Here points in R^2 will be denoted by (x,y).

Without loss of generality, we assume that $p = (0,0)$, that there is an $f \in C^2(R)$ with $f(0) = 0$, $f'(0) = 0$, and that

$$A = \{\,(x,y) \,:\, y > f(x)\,\}.$$

Hence we have the representation $f(x) = x^2 g(x)$ for some continuous function $g(x)$.

Since $B_r(0)$ contains the rectangle $[-r/2, r/2] \times [-r/2, r/2]$ and is contained in the rectangle $[-r, r] \times [-r, r]$, without loss of generality we may assume that $B_r(p)$ is the rectangle $[-r, r] \times [-r, r]$. Furthermore, we may take r small enough so that $|f(x)| < r$ for $|x| < r$.

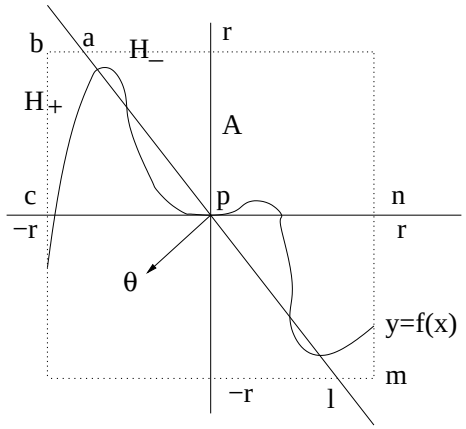
We first show that $\nu(A,p) = e_2 = (0,-1)$. Now $H_+(e_2,p)$ is the lower half-plane and $H_-(e_2,p)$ is the upper half-plane. Then

$$A \cap B_r(p) \cap H_+(e_2,p) = \{\,(x,y)\,;\, -r < x < r,\ \min(f(x),0) \le y \le 0\,\},$$
$$A^c \cap B_r(p) \cap H_-(e_2,p) = \{\,(x,y)\,;\, -r < x < r,\ 0 \le y \le \max(f(x),0)\,\}.$$

Hence

$$r^{-2}\,|A \cap B_r(p) \cap H_+(e_2,p)| \le r^{-2} \int_{-r}^{r} |f(x)|\,dx \le Cr^{-2} \int_{-r}^{r} x^2\,dx = \frac{2Cr}{3},$$
$$r^{-2}\,|A^c \cap B_r(p) \cap H_-(e_2,p)| \le r^{-2} \int_{-r}^{r} |f(x)|\,dx \le Cr^{-2} \int_{-r}^{r} x^2\,dx = \frac{2Cr}{3},$$

which proves the second item. □



We now give a proof of the third item. We will give a proof when $\theta = (\theta_1, \theta_2)$ with $\theta_1 < 0$ and $\theta_2 < 0$. The other cases are similar. Below we will talk of the quadrilaterals (or triangles) $Quad(pabc)$ and $Quad(plmn)$, which represent the intersections of $H_+(p,\theta)$ and $H_-(p,\theta)$ with the second and fourth quadrants.

We observe that

$$\begin{aligned}
&A \cap H_+(p,\theta) \cap B_r(p) \supset Quad(pabc) \setminus A^c \\
&= Quad(pabc) \setminus \{\, (x,y) \,:\, -r < x < 0,\ 0 \le y \le \max(f(x),0) \,\}, \\
&A^c \cap H_-(p,\theta) \cap B_r(p) \supset Quad(plmn) \setminus A \\
&= Quad(plmn) \setminus \{\, (x,y) \,:\, 0 < x < r,\ \min(f(x),0) \le y \le 0 \,\}.
\end{aligned}$$

Hence

$$\begin{aligned}
r^{-2}\,|A \cap H_+(p,\theta) \cap B_r(p)| &\ge r^{-2} Area(pabc) - r^{-2} \int_{-r}^{0} |f(x)|\,dx, \\
r^{-2}\,|A^c \cap H_-(p,\theta) \cap B_r(p)| &\ge r^{-2} Area(plmn) - r^{-2} \int_{0}^{r} |f(x)|\,dx.
\end{aligned}$$

Now $r^{-2}Area(pabc) = r^{-2}Area(plmn) = C$ for some constant $C > 0$ independent of r, and

$$r^{-2} \int_{-r}^{r} |f(x)|\,dx \le C_1 r^{-2} \int_{-r}^{r} x^2\,dx = \frac{2C_1 r}{3}.$$

Hence the result follows. □

For subsets A and B of R^m, let $p \in \partial(A \cap B)$. We now wish to relate $\nu(A \cap B, p)$ to $\nu(A,p)$ and $\nu(B,p)$. If $p \in \partial(A \cap B)$, then $p \in \partial A \cup \partial B$, and if p is not a boundary point of B, then it is an interior point of B, and hence $B_r(p) \cap (A \cap B) = B_r(p) \cap A$ for r small enough. Hence for boundary points p of $A \cap B$, with $p \notin \partial A \cap \partial B$, we have $\nu(A \cap B, p) = \nu(A,p)$ if $p \in \partial A$ and $\nu(A \cap B, p) = \nu(B,p)$ if $p \in \partial B$. So it remains to determine $\nu(A \cap B, p)$ when $p \in \partial A \cap \partial B$.

PROPOSITION 5. *Suppose A and B are subsets of R^m, p is a boundary point of $A \cap B$, and $p \in \partial A \cap \partial B$. Suppose, for some $\rho > 0$,*

$$A \cap B_\rho(p) = \{x \in B_\rho(p) \,:\, x_m > f(x')\}$$

for some C^2 function $f(x')$ of $m-1$ variables; then either $\nu(A\cap B,p)=\nu(A,p)$ or $\nu(A\cap B,p)=0$.

Proof. Let θ be a unit vector, $\theta\neq\nu(A,p)$. Then, from Proposition 4, there is a $c>0$ so that for small enough r

$$r^{-m}\ |A^c\cap H_-(p,\theta)\cap B_r(p)|\ >\ c.$$

Hence, for small enough r,

$$r^{-m}\ |(A\cap B)^c\cap H_-(p,\theta)\cap B_r(p)|\ >\ c.$$

So θ cannot be the normal to $A\cap B$ at p. □

Now we state the Gauss–Green theorem as stated in [10].

PROPOSITION 6 (Gauss–Green theorem). *Let A be a bounded measurable subset of R^m with $\Phi(\partial A)<\infty$, and $f\in C^1(R^m)$. Then*

$$\int_A \frac{\partial f}{\partial x_j}\,dx=\int_{R^m} f(x)\,\nu_j(A,x)\,d\Phi,\qquad\qquad j=1,2,\dots,m.$$

Here $\nu_j(A,x)$ is the jth component of $\nu(A,x)$.

Acknowledgments. A large part of this work was done during the fall of 2001, at the MSRI, Berkeley, CA, where the authors were participating in a semester-long program on inverse problems. The authors would like to thank the organizers of this program, particularly David Eisenbud, Gunther Uhlmann, and the NSF, for organizing this program and providing financial support. David Finch and Rakesh were on sabbatical and would like to thank their respective universities for providing this opportunity and financial support. Rakesh also worked on this article when visiting Vanderbilt University during the spring of 2002 and would like to thank the Mathematics Department at Vanderbilt University for its hospitality and financial support. The authors also wish to thank one of the referees for the observation, mentioned in the introduction, that Theorems 4 and 5 (and hence Theorems 1 and 2) are valid under a slightly weaker hypothesis.

REFERENCES

[1] M. L. AGRANOVSKY AND E. T. QUINTO, *Injectivity sets for the Radon transform over circles and complete systems of radial functions*, J. Funct. Anal., 139 (1996), pp. 383–414.

[2] M. L. AGRANOVSKY AND E. T. QUINTO, *Geometry of stationary sets for the wave equation in R^n; the case of finitely supported initial data*, Duke Math. J., 107 (2001), pp. 57–84.

[3] M. L. AGRANOVSKY, C. BERENSTEIN, AND P. KUCHMENT, *Approximation by spherical means in L^p spaces*, J. Geom. Anal., 6 (1996), pp. 365–383.

[4] L.-E. ANDERSSON, *On the determination of a function from spherical averages*, SIAM J. Math. Anal., 19 (1988), pp. 214–232.

[5] A. L. BUKHGEIM AND V. B. KARDAKOV, *Solution of an inverse problem for an elastic wave equation by the method of spherical means*, Siberian Math. J., 19 (1978), pp. 528–535.

[6] A. M. CORMACK AND E. T. QUINTO, *A Radon transform on spheres through the origin in R^n and applications to the Darboux equation*, Trans. Amer. Math. Soc., 260 (1980), pp. 575–581.

[7] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics, Volume* II, John Wiley, New York, 1962.

[8] L. C. EVANS AND R. F. GARIEPY, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[9] J. A. FAWCETT, *Inversion of n-dimensional spherical averages*, SIAM J. Appl. Math., 45 (1985), pp. 336–341.

[10] H. Federer, *A note on the Gauss-Green theorem*, Proc. Amer. Math. Soc., 9 (1958), pp. 447–451.
[11] D. Finch and Rakesh, *Trace Regularity of Solutions of the Wave Equation*, in preparation (2003).
[12] F. G. Friedlander, *The Wave Equation on a Curved Space-Time*, Cambridge University Press, Cambridge, UK, 1975.
[13] A. B. Goncharov, *Differential equations and integral geometry*, Adv. Math., 131 (1997), pp. 279–343.
[14] L. Hormander, *The Analysis of Linear Partial Differential Operators* I, Springer-Verlag, New York, 1983.
[15] F. John, *Plane Waves and Spherical Means*, Wiley, New York, 1955.
[16] F. John, *Partial Differential Equations*, 4th ed., Springer-Verlag, New York, 1982.
[17] W. Joines, Y. Zhang, C. Li, and R. Jirtle, *The measured electrical properties of normal and malignant human tissue from* 50 *to* 900 *mhz*, Med. Phys., 21 (1994), pp. 547–550.
[18] R. A. Kruger, K. D. Miller, H. E. Reynolds, W. L. Kiser, Jr., D. R. Reinecke, and G. A. Kruger, *Contrast enhancement of breast cancer in vivo using thermoacoustic CT at* 434 *MHz*, Radiology, 216 (2000), pp. 279–283.
[19] R. A. Kruger, K. K. Kopecky, A. M. Aisen, D. R. Reinecke, G. A. Kruger, and W. L. Kiser, Jr., *Thermoacoustic CT with radio waves: A medical imaging paradigm*, Radiology, 211 (1999), pp. 275–278.
[20] R. A. Kruger, D. R. Reinecke, and G. A. Kruger, *Thermoacoustic computed tomography*, Med. Phys., 26 (1999), pp. 1832–1837.
[21] A. K. Louis and E. T. Quinto, *Local tomographic methods in SONAR*, in Surveys on Solution Methods for Inverse Problems, D. Colton, H. W. Engl, A. K. Louis, J. R. McLaughlin, and W. Rundell, eds., Springer-Verlag, Vienna, 2000, pp. 147–154.
[22] M. M. Lavrentiev, V. G. Romanov, and V. G. Vasiliev, *Multidimensional Inverse Problems for Differential Equations*, Lecture Notes in Math. 167, Springer-Verlag, New York, 1970.
[23] M. M. Lavrentiev, V. G. Romanov, and S. P. Shishatskii, *Ill-posed Problems of Mathematical Physics and Analysis*, Transl. Math. Monogr. 64, AMS, Providence, RI, 1986.
[24] J. M. Lee, *Riemannian Manifolds*, Springer-Verlag, New York, 1997.
[25] F. Natterer, *The Mathematics of Computerized Tomography*, Teubner, Stuttgart, 1986.
[26] S. J. Norton, *Reconstruction of a two-dimensional reflecting medium over a circular domain: Exact solution*, J. Acoust. Soc. Amer., 67 (1980), pp. 1266–1273.
[27] S. J. Norton and M. Linzer, *Ultrasonic reflectivity imaging in three dimensions: Exact inverse scattering solutions for plane, cylindrical, and spherical apertures*, IEEE Trans. Biomedical Engineering, 28 (1981), pp. 200–202.
[28] A. G. Ramm, *Injectivity of the spherical mean operator*, C. R. Math. Acad. Sci. Paris, 335 (2002), pp. 1033–1038.
[29] E. M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1971.
[30] D. Tataru, *Unique continuation for partial differential operators with partially analytic coefficients*, J. Math. Pures Appl., 78 (1999), pp. 505–521.
[31] V. V. Volchkov, *Injectivity sets for the Radon transform over a sphere*, Izv. Math., 63 (1999), pp. 481–493.
[32] M. Xu, L. V. Wang, *Time-Domain Reconstruction for Thermoacoustic Tomography in a Spherical Geometry*, IEEE Trans. Medical Imaging, 21 (2002), pp. 814–822.
[33] Y. Xu, D. Feng, and L. V. Wang, *Exact frequency-domain reconstruction for thermoacoustic tomography* I: *Planar geometry*, IEEE Trans. Medical Imaging, 21 (2002), pp. 823–828.
[34] Y. Xu, M. Xu, and L. V. Wang, *Exact frequency-domain reconstruction for thermoacoustic tomography* II: *Cylindrical geometry*, IEEE Trans. Medical Imaging, 21 (2002), pp. 829–833.
[35] A. E. Yagle, *Inversion of spherical means using geometric inversion and Radon transforms*, Inverse Problems, 8 (1992), pp. 949–964.

SIAM J. MATH. ANAL.
Vol. 35, No. 5, pp. 1241–1249

ON THE SECOND EIGENVALUE FOR NONHOMOGENEOUS QUASI-LINEAR OPERATORS*

STEPHEN B. ROBINSON†

Abstract. In this paper we study a nonlinear eigenvalue problem and associated perturbations of the problem. More specifically, we generalize a variational characterization of the second eigenvalue for homogenous quasi-linear elliptic operators, such as the p-Laplacian, to a class of nonhomogeneous quasi-linear elliptic operators. Neumann boundary data is assumed throughout the paper. To demonstrate the utility of this characterization we use it to prove a generalized Fredholm alternative for nonresonant perturbations of the given eigenvalue problem.

Key words. second eigenvalue, quasi-linear, nonresonance

AMS subject classifications. 35P30, 35J20, 35J65

DOI. 10.1137/S0036141003426008

1. Introduction.

In this paper we consider the nonlinear eigenvalue problem

$$\begin{aligned} Qu - \lambda|u|^{p-2}u = 0 \text{ a.e. in } \Omega, \\ \tfrac{\partial u}{\partial \nu} = 0 \text{ on } \partial\Omega, \end{aligned} \tag{1.1}$$

where Q is a quasi-linear elliptic operator generalizing the p-Laplacian, $\Omega \subset \Re^N$ is a smooth bounded domain, ν is the unit outward normal on $\partial\Omega$, λ is a real number, and $1 < p < \infty$. A key technical challenge is that Q is not assumed to be homogeneous.

It will be clear that the principal eigenvalue is $\lambda_1 = 0$ with an associated simple eigenspace of constant functions, $W := span\{1\}$. Our interest is in establishing and exploiting an appropriate variational characterization for the second eigenvalue. For homogeneous operators, it is simplest to define λ_2 as the smallest number that is strictly larger than λ_1 such that (1.1) has a nontrivial solution, and it is well known that this definition has useful variational characterizations; see [2]. Choosing an appropriate definition becomes more subtle when dealing with nonhomogeneous operators, because the existence of a λ such that (1.1) has a nontrivial solution does not necessarily imply the existence of an unbounded set of solutions. Describing unbounded sets of solutions is fundamental to understanding perturbations of (1.1). In order to motivate a useful definition, it is helpful to review the properties of more familiar operators.

In the linear case, e.g., where $Q = -\Delta$, it is well known that

$$\lambda_2 = \inf \frac{\mathcal{Q}(v,v)}{||v||^2_{L^2}} \text{ for } v \in V_2 \setminus \{0\}, \tag{1.2}$$

where $\mathcal{Q}(u,v) := \int_\Omega \nabla u \cdot \nabla v$ is the bilinear form associated with Q, and $V_2 := W^\perp = \{u \in W^{1,2}(\Omega) : \int_\Omega u = 0\}$. An equivalent characterization, of minimax type, is given by

$$\lambda_2 = \inf_{\gamma \in \Gamma} \sup_{t \in [-1,1]} \mathcal{Q}(\gamma(t), \gamma(t)), \tag{1.3}$$

*Received by the editors April 9, 2003; accepted for publication July 18, 2003; published electronically January 6, 2004.
http://www.siam.org/journals/sima/35-5/42600.html
†Wake Forest University, Department of Mathematics, P.O. Box 7388, Winston-Salem, NC 27109 (sbr@wfu.edu).

where $\Gamma := \{\gamma : [-1,1] \to \partial B_1 : \gamma \text{ is continuous, and } \gamma(\pm 1) = \pm\phi_1\}$, $\partial B_1 := \{u \in W^{1,2}(\Omega) : ||u||_{L^2} = 1\}$, and $\phi_1 := (\frac{1}{|\Omega|})^{\frac{1}{2}}$.

These statements generalize in a natural way to certain homogeneous operators such as the p-Laplacian, i.e., $Qu := -\nabla \cdot (|\nabla u|^{p-2}\nabla u)$. To generalize (1.2) in this case, replace the bilinear form with the *quasi-linear* form $\mathcal{Q}(u,v) = \int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla v$, and replace the subspace V_2 with the surface $V_p := \{u \in W^{1,p}(\Omega) : \int |u|^{p-2}u = 0\}$. To this author's knowledge, it was not known until recently whether replacing V_2 with V_p is necessary. In fact, for the ODE case, and for the PDE case over certain simple domains such as $\Omega = [0,1]^N$, it is clear that one obtains the same value when minimizing over either V_2 or V_p. However, in [7] it is shown that, in general, a sharper characterization is obtained by minimizing over V_p.

To generalize (1.3), simply replace L^2 norms by L^p norms in appropriate places. See [2], [3], [4], and the references therein for a more detailed discussion of the p-Laplacian and its eigenvalues.

For nonhomogeneous quasi-linear operators, there are many papers in the literature describing the properties of the principle eigenvalue and corresponding principle eigenfunctions. For example, see the results and references in [5]. However, there seem to be relatively few papers that consider the second eigenvalue. A notable exception is found in the work of Shapiro et al., where a variety of resonance and nonresonance theorems involving a second eigenvalue are proved under very general circumstances. For example, see [10] and [12] and references therein. (Studying these results was a primary motivation for this paper.) In both [10] and [12], the second eigenvalue is defined as

$$\lambda_2^* := \liminf_{||v||_{L^p(\Omega)} \to \infty} \frac{Q(v,v)}{||v||^p_{L^p}(\Omega)} \text{ for } v \in V_2.$$

This definition is naturally motivated by the nearly orthogonal splitting of the Banach space $W^{1,p}(\Omega) = W + V_2$. Observe that this definition generalizes (1.2) but does not substitute V_p for V_2. The result in [7] shows that $\lambda_2^* < \lambda_2$ in general. It follows that our results, which are based upon the sharper characterization, lay the groundwork for more general existence theorems, as is demonstrated in section 4.

The paper is organized as follows. In section 2 we provide a precise description of Q along with preliminary remarks on notation and simple properties. In section 3 we define λ_2 using a natural generalization of (1.3) and then show that this is equivalent to a natural generalization of (1.2). Moreover, we establish a helpful estimate regarding the primitive of the quasi-linear form associated with Q and hint at possible generalizations. Finally, in section 4 we consider the boundary value problem

$$\begin{aligned} Qu - g(x,u) &= h \text{ a.e. in } \Omega, \\ \tfrac{\partial u}{\partial \nu} &= 0 \text{ on } \partial\Omega \end{aligned} \tag{1.4}$$

and prove an existence theorem assuming that $\frac{g(x,u)}{|u|^{p-2}u}$ lies strictly between 0 and λ_2 for large u and that $h \in (W^{1,p}(\Omega))^*$ is arbitrary. This demonstrates the usefulness of λ_2 as a bound for existence theorems and generalizes one case of the Fredholm alternative for self-adjoint linear operators. The proof obtains a solution as a saddle point over linked sets. For the relevant definitions and theorems of measure theory, Sobolev spaces, and variational theory, we refer the reader to [9], [1], and [13], respectively.

2. Preliminaries. Let $A : \Omega \times \Re^N \to \Re^N$ such that

(A–1) (*Carathéodory*) The map $x \to A(x,\xi)$ is measurable for each $\xi \in \Re^N$, and the map $\xi \to A(x,\xi)$ is continuous for a.e. $x \in \Omega$.

(A–2) (*Growth*) There exist a positive constant c_1, a constant $p \in (1,\infty)$, and a nonnegative function $\widetilde{h} \in L^{p'}(\Omega)$, where $p' = p/(p-1)$, such that

$$|A(x,\xi)| \le \widetilde{h}(x) + c_1|\xi|^{p-1}$$

for a.e. $x \in \Omega$ and for all $\xi \in \Re^N$.

(A–3) (*Ellipticity*) There exists a positive constant c_2 such that

$$\sum_{i=1}^{N} A_i(x,\xi)\cdot\xi \ge c_2 \sum_{i=1}^{N} |\xi|^p$$

for a.e. $x \in \Omega$ and for all $\xi \in \Re^N$, where p is as in (A–2).

(A–4) (*Monotonicity*) Assume that for a.e. $x \in \Omega$ and each $\xi, \xi^* \in \Re^N$ with $\xi \ne \xi^*$,

$$\sum_{i=1}^{N} [A_i(x,\xi) - A_i(x,\xi^*)](\xi_i - \xi_i^*) > 0.$$

(A–5) (*One-sided p-homogeneity*) $A(x,t\xi)\cdot\xi \le t^{p-1}A(x,\xi)\cdot\xi$ for all $t > 0$ and all $(x,\xi) \in \Omega \times \Re^N$.

Given the assumptions above, it now makes sense to formally define

$$Qu := -\nabla\cdot(A(x,\nabla u)) \tag{2.1}$$

and to define the quasi-linear Dirichlet form

$$\mathcal{Q}(u,v) := \int_\Omega A(x,\nabla u)\cdot\nabla v \qquad \forall\ u,v \in W^{1,p}(\Omega). \tag{2.2}$$

In view of (A–2), we see that $\mathcal{Q}$ is well defined on $W^{1,p}(\Omega)\times W^{1,p}(\Omega)$.

In order to impose a variational structure on Q, we assume that $A(x,\xi) = \nabla_\xi F(x,\xi)$, where $F\colon \Omega \times \Re^N \to \Re$ satisfies the following:

(F–1) (*Carathéodory*) The map $x \to F(x,\xi)$ is measurable for each $\xi \in \Re^N$, and the map $\xi \to F(x,\xi)$ is continuously differentiable for a.e. $x \in \Omega$.

(F–2) (*Growth*) There exist a positive constant c_3 and a nonnegative function $h \in L^1(\Omega)$ such that

$$|F(x,\xi)| \le h(x) + c_3|\xi|^p$$

for a.e. $x \in \Omega$ and all $\xi \in \Re^N$, where p is chosen as in (A–2).

(F–3) (*Normalization*) $F(x,0) = 0$ for a.e. $x \in \Omega$.

It follows that $u \mapsto \int_\Omega F(x,\nabla u)$ is a C^1 functional on $W^{1,p}(\Omega)$ with derivative $u \mapsto \mathcal{Q}(u,\cdot)$. Moreover, using the fundamental theorem of calculus and Fubini's theorem, we see that $\int_\Omega F(x,\nabla u) = \int_0^1 \mathcal{Q}(tu,u)dt$, which will be the more useful form for later estimates. Given this structure, we see that solutions of (1.1) and (1.4) are equivalent to critical points of the functionals

$$\begin{aligned} E_\lambda(u) &= \textstyle\int_0^1 \mathcal{Q}(tu,u)dt - \frac{\lambda}{p}\int_\Omega |u|^p \text{ and} \\ J(u) &= \textstyle\int_0^1 \mathcal{Q}(tu,u)dt - \int_\Omega G(x,u) - h(u), \end{aligned} \tag{2.3}$$

respectively, where $G(x,u) := \int_0^u g(x,t)dt$.

Throughout this paper we will use the norm in $W^{1,p}(\Omega)$ given by

$$||u||_{1,p}^p = ||u||_{L^p}^p + \sum_{i=1}^N ||\frac{\partial u}{\partial x_i}||_{L^p}^p,$$

where $||\cdot||_{L^p}$ denotes the $L^p(\Omega)$ norm. We will also be using the seminorm

$$|u|'_{1,p} = \left\{ \sum_{i=1}^N \left\| \frac{\partial u}{\partial x_i} \right\|_{L^p}^p \right\}^{1/p}.$$

Observe that by the definition of $\mathcal{Q}$ in (2.2) and (A–3) we get

$$\mathcal{Q}(u,u) \geq c_2 \int_\Omega \sum_{i=1}^N \left| \frac{\partial u}{\partial x_i} \right|^p = c_2(|u|'_{1,p})^p \tag{2.4}$$

for all $u \in W^{1,p}(\Omega)$ so that $\liminf_{||u||_{L^p}\to\infty} \frac{\mathcal{Q}(u,u)}{||u||_{L^p}^p} \geq 0$. Define

$$\lambda_1 := \liminf_{||u||_{L^p}\to\infty} \frac{\mathcal{Q}(u,u)}{||u||_{L^p}^p}$$

as in [11, p. 1821] . Since $\mathcal{Q}(u,u) = 0$ for u constant, we see that $\lambda_1 = 0$. On the other hand, for nonconstant $v \in W^{1,p}(\Omega)$ we obtain from (2.4) that $\mathcal{Q}(v,v) > 0$, so $\lambda_1 = 0$ behaves like a simple eigenvalue with constant normalized eigenfunction $\phi_1 \equiv \frac{1}{|\Omega|^{1/p}}$ and corresponding eigenspace $W := span\{1\}$.

Remark 1. Conditions (A–1) through (A–5) and (F–1) through (F–3) are not entirely independent. For example, the growth condition on F can be derived from the growth condition on A. For convenience and clarity, we chose to state these standard properties separately.

Remark 2. Condition (A–5) is only used in the proof of the Palais–Smale condition in section 4. Conditions of this type have appeared often in the literature. Recently, in [12], Shapiro used a similar condition to prove an existence theorem for (1.4), assuming that g satisfies a superlinear growth condition. (A–5) implies that $pF(x,\xi) \leq \nabla_\xi F(x,\xi)\cdot\xi$, which is closely related to the Ambrosetti–Rabinowitz condition (AR), which has appeared in a variety of contexts. However, AR is usually imposed on the nonlinear perturbation of a linear elliptic boundary value problem (see condition (p_4) on page 9 of [6]), whereas (A–5) is imposed upon the nonlinear differential operator.

3. Definition and characterization of λ_2. We begin by stating a natural generalization of (1.3).

DEFINITION 3.1. *Let* $\Gamma_R := \{\gamma : [-1,1] \to \partial B_R : \gamma$ *is continuous, and* $\gamma(\pm 1) = \pm\phi_R\}$, *where* $\partial B_R := \{u \in W^{1,p}(\Omega) : ||u||_{L^p} = R\}$, *and* $\phi_R := (\frac{R}{|\Omega|})^{\frac{1}{p}}$. *Define* $\lambda_{2,R} := \inf_{\gamma\in\Gamma_R} \sup_{t\in[-1,1]} \frac{\mathcal{Q}(u,u)}{R^p}$, *and* $\lambda_2 := \liminf_{R\to\infty} \lambda_{2,R}$.

Our first lemma establishes an equivalence between the above definition and a generalization of (1.2).

LEMMA 3.2. $\lambda_2 = \liminf_{||v||_{L^p}\to\infty} \frac{\mathcal{Q}(v,v)}{||v||_{L^p}^p}$ *for* $v \in V_p := \{u \in W^{1,p}(\Omega) : \int_\Omega |u|^{p-2}u = 0\}$.

Proof. Consider definition (3.1). Any curve in Γ_R must cross V_p, so it is clear that

$$\lambda_{2,R} \geq \inf_{v \in V_p \cap \partial B_R} \frac{\mathcal{Q}(v,v)}{R^p}.$$

Thus $\lambda_2 \geq \liminf_{||v||_{L^p} \to \infty} \frac{\mathcal{Q}(v,v)}{||v||_{L^p}^p}$ for $v \in V_p$.

On the other hand, given $\epsilon > 0$ and $R > 0$, consider $v_R \in V_p \cap \partial B_R$ such that $\frac{\mathcal{Q}(v_R,v_R)}{R^p} < \inf_{v \in \partial B_R \cap V_p} \frac{\mathcal{Q}(v,v)}{R^p} + \epsilon$. Now consider the curve

$$\eta(t) := \begin{cases} -T + tv_R \ \text{ for } 0 \leq t \leq 1, \\ \\ t - 1 - T + v_R \ \text{ for } 1 \leq t \leq 2T+1, \\ \\ T + (-t + 2T + 2)v_R \ \text{ for } 2T+1 \leq t \leq 2T+2, \end{cases}$$

where T is a large positive constant. This is essentially a long line segment paralleling W with short connections to W on either end. Also consider the set

$$\begin{aligned} \mathcal{C} := \quad & \{u \in W^{1,p}(\Omega) : \exists \text{ continuous } \gamma : [-1,1] \to \partial B_{||u||_{L^p}} \text{ such that} \\ & \gamma(-1) = -\frac{||u||_{L^p}}{|\Omega|^{\frac{1}{p}}} \text{ and } \mathcal{Q}(\gamma(t),\gamma(t)) < \lambda_{2,||u||_{L^p}} ||u||_{L^p}^p \ \forall t\}. \end{aligned}$$

$\mathcal{C}$ is the set of points in $W^{1,p}(\Omega)$ that can be connected to a negative constant function by a curve which stays on the surface of an L^p sphere, ∂B_R, without crossing a point where $\frac{\mathcal{Q}(u,u)}{R^p} \geq \lambda_{2,R}$. A straightforward argument shows that $\mathcal{C}$ is open. It is clear that $\eta(0) = -T \in \mathcal{C}$. Thus $\eta^{-1}(\mathcal{C})$ is a nonempty open subset of $[0, 2T+2]$. If $\eta^{-1}(\mathcal{C}) = [0, 2T+2]$, then $\eta(2T+2) = T$ is in $\mathcal{C}$, so T can be connected to $-T$ by a curve on ∂B_R, where $\frac{\mathcal{Q}(u,u)}{R^p} < (\lambda_{2,R})R^p$, for $R = \frac{T}{|\Omega|^{\frac{1}{p}}}$, which contradicts the definition of $\lambda_{2,R}$. Therefore, there is a maximal $t' \in (0, 2T+2)$ such that $\eta(t) \in \mathcal{C}$ for $t \in [0, t')$. Let $u' = \eta(t')$ and $R' = ||u'||_{L^p}$. We see that $\mathcal{Q}(u', u') \geq \lambda_{2,R'}(R')^p$, or else we could move a little farther along η while remaining in $\mathcal{C}$, which would contradict our choice of t'. For large T, we argue that $t' \in (1, 2T+1)$. For any $t \in [0,1] \bigcup [2T+1, 2T+2]$, we have $\frac{\mathcal{Q}(\eta(t),\eta(t))}{||\eta(t)||_{L^p}^p} = \frac{\mathcal{Q}(t^* v_R, t^* v_R)}{||\eta(t)||_{L^p}^p}$, where $t^* \in [0,1]$ and $||\eta(t)||_{L^p} \to \infty$ as $T \to \infty$, so $\frac{\mathcal{Q}(\eta(t),\eta(t))}{||\eta(t)||_{L^p}^p} << \lambda_2$ for large T. This estimate leads to the fact that $\eta(t) \in \mathcal{C}$ for any $t \in [0,1]$ and $\eta(t) \notin \mathcal{C}$ for any $t \in [2T+1, 2T+2]$, and thus $t' \in (1, 2T+1)$. Along the segment where $1 < t < 2T+1$, we see that $\mathcal{Q}(\eta(t), \eta(t)) \equiv \mathcal{Q}(v_R, v_R)$. Also, since for any $u \in W^{1,p}(\Omega)$ the function $\tau \to \int_\Omega |u + \tau|^p$ achieves a unique minumum for s such that $u + s \in V_p$, we see that $||\eta(t)||_{L^p}^p$ achieves its minimum at $t = T+1$, where $\eta(t) = v_R$. Hence $R' > R$ and $\lambda_{2,R'} \leq \frac{\mathcal{Q}(u',u')}{(R')^p} \leq \frac{\mathcal{Q}(v_R,v_R)}{R^p} \leq \inf_{V_p \cap \partial B_R} \frac{\mathcal{Q}(v,v)}{R^p} + \epsilon$. The lemma follows. □

Of course, if the given characterizations lead to $\lambda_2 = \lambda_1 = 0$, then the existence theorem in section 4 would not be very interesting. Thus we should take a moment to mention the following.

LEMMA 3.3. $0 < \lambda_2 < \infty$.

Proof. The first inequality follows from the ellipticity condition and a Poincare–type inequality. For the details of a more general estimate see Lemma 3.2 in [8]. The second inequality follows from the growth condition on A. □

In the homogeneous case the relationship between $\mathcal{Q}$ and its primitive is trivial, because $\int_0^1 \mathcal{Q}(tu, u)dt = \mathcal{Q}(u,u)\int_0^1 t^{p-1}dt = \frac{1}{p}\mathcal{Q}(u,u)$, but in the nonhomogenous

case this relationship is more subtle. The following lemma provides a useful estimate and a relationship between $\mathcal{Q}$ and its primitive.

LEMMA 3.4. $\lambda_2 \leq \liminf_{||v||_{L^p}\to\infty} \frac{p\int_0^1 \mathcal{Q}(tv,v)dt}{||v||_{L^p}^p}$ *for* $v \in V_p$.

Proof. Given $\epsilon > 0$ and $\delta > 0$, choose $R > 0$ such that $\inf_{\partial B_r \cap V_p} \mathcal{Q}(v,v) \geq (\lambda_2 - \epsilon)r^p$ for all $r > \delta R$. Hence for $v \in (\partial B_R \cup V_p)$,

$$p\int_0^1 \mathcal{Q}(tv,v)dt \geq p(\lambda_2-\epsilon)\int_\delta^1 t^{p-1}dt \int_\Omega |v|^p = (\lambda_2-\epsilon)(1-\delta^p)\int_\Omega |v|^p.$$

The result follows. □

Remark 3. An interesting possibility would be to base our definition of λ_2 on the generalized Raleigh quotient

$$\frac{p\int_0^1 \mathcal{Q}(tv,v)dt}{\int_\Omega |v|^p}$$

rather than on

$$\frac{\mathcal{Q}(u,u)}{\int_\Omega |u|^p}.$$

This is not an issue in the homogenous case, but might be of interest for certain nonhomogeneous operators. Notice that a straightforward Lagrange multipliers argument shows that a critical point of $\int_0^1 \mathcal{Q}(tv,v)dt$ constrained to a sphere ∂B_R must occur on $V_p \bigcap \partial B_R$. Hence a minimax characterization similar to Definition 3.1 would reduce to the lim inf characterization in Lemma 3.4. Moreover, it would be of interest to compare these variational definitions to an arguably more natural definition generalizing $\lambda_2 := \inf\{\lambda > 0 : (1.1) \text{ has a nontrivial solution}\}$, as in [2].

Remark 4. In order to simplify notation and clarify exposition, we have limited ourselves to second order operators. However, the notation and the results in this paper generalize in a straightforward way to operators of order $2m$. See [10] for details.

4. A nonresonance theorem. In this section we consider the boundary value problem (1.4), where $\frac{g(x,u)}{|u|^{p-2}u}$ is bounded strictly between the eigenvalues λ_1 and λ_2, and where $h \in (W^{1,p}(\Omega))^*$. This is called a nonresonance problem and we should expect, as in the Fredholm alternative, that the problem will be solvable for any choice of h. Proving this theorem verifies the practicality of Definition 3.1.

First we set the stage for a variational proof. Let $G(x,u) := \int_0^u g(x,t)dt$ and let

$$J(u) := \int_0^1 \mathcal{Q}(tu,u)dt - \int_\Omega G(x,u) - h(u) \text{ for } u \in W^{1,p}(\Omega).$$

J is a C^1 functional with

$$J'(u)v = \mathcal{Q}(u,v) - \int_\Omega g(x,u)v - h(v).$$

Critical points of J correspond to weak solutions of (1.4).

Our proof establishes the existence of a critical point using a saddle point theorem over linked sets. See [13, Theorem 8.4] for details. We will show that J has a saddle

geometry over the linked sets W and V_p, and then we will show that J satisfies the Palais–Smale condition, i.e., that if $\{u_n\} \subset W^{1,p}(\Omega)$ such that $\{J(u_n)\}$ is bounded and $J'(u_n) \to 0$ in $(W^{1,p}(\Omega))^*$, then $\{u_n\}$ has a converging subsequence.

THEOREM 4.1. *Assume* (A–1), (A–2), (A–3), (A–4), (A–5), (F–1), (F–2), *and* (F–3) *and that* $Q(u) := -\nabla \cdot (A(x, \nabla u))$. *In addition assume that* $g : \Omega \times \Re \to \Re$ *is a Carathéodory function satisfying*

$$0 \leq \epsilon \leq \frac{g(x,u)}{|u|^{p-2}u} \leq \lambda_2 - \epsilon \tag{4.1}$$

for some $\epsilon > 0$. *Assume that* $h \in (W^{1,p}(\Omega))^*$. *Then* (1.4) *has at least one weak solution in* $W^{1,p}(\Omega)$.

Proof. It is clear that, for any constant $C > 0$, if $\gamma : [-1, 1] \to W^{1,p}(\Omega)$ is a continuous curve such that $\gamma(\pm 1) = \pm C$, then there is at least one $t_0 \in [-1, 1]$ such that $\gamma(t) \in V_p$. Thus $\{\pm C\}$ and V_p link.

Consider J restricted to W. Since $\mathcal{Q}(u, c) = 0$ for any $u \in W^{1,p}(\Omega)$ and any constant c, we have

$$J(c) = -\int_\Omega G(x, c) - h(c).$$

Using (4.1) we see that $G(x, c) \geq \frac{\epsilon}{p}|c|^p$ for all c, so it follows that $\lim_{||w||_{W^{1,p}(\Omega)} \to \infty} J(w) = -\infty$.

Consider J restricted to V_p. First we choose $\delta > 0$ and apply ellipticity to get

$$J(v) \geq \delta c_2 \int_\Omega |\nabla v|^p + (1 - \delta) \int_0^1 \mathcal{Q}(tv, v)dt - \int_\Omega G(x, v) - h(v).$$

By Lemma 3.4 there is an $R > 0$ such that $\int_0^1 \mathcal{Q}(tv, v) \geq (\lambda_2 - \frac{\epsilon}{2})||v||^p_{L^p}$ for $||v||_{L^p} > R$. Applying (4.1) again, we see that $G(x, v) \leq \frac{\lambda_2 - \epsilon}{p}|v|^p$ for all v, so

$$J(v) \geq \delta c_2 \int_\Omega |\nabla v|^p + (1 - \delta)\left(\frac{\lambda_2 - \frac{\epsilon}{2}}{p}\right)||v||^p_{L^p} - \frac{\lambda_2 - \epsilon}{p}||v||^p_{L^p} - ||h||_{W^{1,p}(\Omega)^*}||v||_{1,p}.$$

For δ small enough, there is a constant $c' > 0$ such that

$$J(v) \geq c'||v||^p_{1,p} - ||h||_{W^{1,p}(\Omega)^*}||v||_{1,p}.$$

Hence, for J restricted to V_p, we have $\lim_{||v||_{1,p} \to \infty} J(v) = \infty$.

We have shown that J has a saddle geometry over the linked sets W and V_p. It remains to prove the Palais–Smale condition. Suppose that $\{u_n\} \in W^{1,p}(\Omega)$ such that $|J(u_n)| \leq K$ for all n for some $K > 0$, and such that $J'(u_n) \to 0$ in $(W^{1,p}(\Omega))^*$. We must show that $\{u_n\}$ has a converging subsequence. We note that it suffices to show that there is a bounded subsequence. See [10]. Suppose that $||u_n||_{1,p} \to \infty$. First, using ellipticity and (4.1), we see that

$$J'(u_n) \cdot u_n \geq c_2(|u_n|'_{1,p})^p - (\lambda_2 - \epsilon)||u_n||^p_{L^p}.$$

Since $J'(u_n) \to 0$, we can divide the given inequality by $||u_n||_{L^p}$ and discover that $|u_n|'_{1,p} \leq c'||u_n||_{L^p}$ for some $c' > 0$, and thus $||u_n||_{1,p} \leq c''||u_n||_{L^p}$ for some $c'' > 0$.

Without loss of generality, the sequence $\{\frac{u_n}{||u_n||_{L^p}}\}$ converges weakly in $W^{1,p}(\Omega)$ and strongly in $L^p(\Omega)$ to some function u such that $||u||_{L^p} = 1$. Now consider

$$J'(u_n)\cdot 1 = \int_\Omega g(x,u_n) = \int_\Omega g(x,u_n^+) + \int_\Omega g(x,-u_n^-).$$

Equation (4.1) implies

$$\epsilon\frac{\int_\Omega |u_n^\pm|^{p-1}}{||u_n||_{L^p}^{p-1}} \le \pm\frac{\int_\Omega g(x,\pm u_n^\pm)}{||u_n||_{L^p}^{p-1}} \le (\lambda_2-\epsilon)\frac{\int_\Omega |u_n^\pm|^{p-1}}{||u_n||_{L^p}^{p-1}},$$

and $J'(u_n)\to 0$ implies that

$$\lim_{n\to\infty}\frac{\int_\Omega g(x,u_n^+)}{||u_n||_{L^p}^{p-1}} = -\lim_{n\to\infty}\frac{\int_\Omega g(x,-u_n^-)}{||u_n||_{L^p}^{p-1}},$$

where we can assume that the given limits exist by passing to a subsequence. If this limit is 0, then $\int_\Omega |u^+|^{p-1} = \int_\Omega |u^-|^{p-1} = 0$, which contradicts the fact that u is nontrivial. Thus the limit is positive and it follows that both u^+ and u^- are nontrivial. Now consider

$$J'(u_n)\cdot u_n^+ = \mathcal{Q}(u_n,u_n^+) - \int_\Omega g(x,u_n)u_n^+ = \mathcal{Q}(u_n^+,u_n^+) - \int_\Omega g(x,u_n^+)u_n^+.$$

Dividing through by $||u_n||_{L^p}^p$, using (4.1), and using the fact that $J'(u_n)\cdot\frac{u_n}{||u_n||_{L^p}}\to 0$, we see that

$$\mathcal{Q}(u_n^+,u_n^+) \le \left(\lambda-\frac{\epsilon}{2}\right)\int_\Omega |u_n^+|^p$$

for n large. A similar estimate holds for u_n^-. Hence, for large n, we have that u_n^+ and u_n^- are nontrivial and satisfy the previous inequality. Using such a u_n we construct the curve $\gamma(\alpha,\beta) = \alpha u_n^+ - \beta u_n^-$ such that α and β are nonnegative and $\alpha^p||u_n^+||_{L^p}^p + \beta^p||u_n^-||_{L^p}^p = ||u_n||_{L^p}^p$. It is now straightforward to check that this curve lives on the L^p ball of radius $||u_n||_{L^p}$ and crosses V_p. Using (A–5) we get the estimate

$$\begin{aligned}\mathcal{Q}(\alpha u_n^+ - \beta u_n^-, \alpha u_n^+ - \beta u_n^-) &= \mathcal{Q}(\alpha u_n^+,\alpha u_n^+) + \mathcal{Q}(-\beta u_n^-\beta u_n^-)\\ &\le (\alpha^p||u_n^+||_{L^p}^p + \beta^p||u_n^-||_{L^p}^p)\,(\lambda_2-\tfrac{\epsilon}{2})\\ &\le ||u||_{L^p}(\lambda_2-\tfrac{\epsilon}{2}).\end{aligned}$$

However, for large $||u_n||_{L^p}$, we must have $\mathcal{Q}(\gamma(\alpha,\beta)) > (\lambda_2-\frac{\epsilon}{2})||u_n||_{L^p}^p$ at the point where this curve crosses V_p. Thus we have reached a contradiction and the proof is complete. □

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] A. Anane and N. Tsouli, *On the second eigenvalue of the p-Laplacian*, in Nonlinear Partial Differential Equations (Fés, 1994), Pitman Res. Notes Math. Ser. 343, Longman, Harlow, UK, 1996, pp. 1–9.

[3] M. Cuesta, D. G. de Figueiredo, and J.-P. Gossez, *The beginning of the Fučik spectrum for the p-Laplacian*, J. Differential Equations, 159 (1999), pp. 212–238.

[4] P. Drábek and S. B. Robinson, *Resonance Problems for the p-Laplacian*, J. Funct. Anal., 169 (1999), pp. 189–200.

[5] P. Drábek, A. Kufner, and F. Nicolosi, *Quasilinear Elliptic Equations with Degenerations and Singularities*, de Gruyter Ser. Nonlinear Anal. Appl. 5, de Gruyter, Berlin, New York, 1997.

[6] P. Rabinowitz, *Minimax Methods in Critical Point Theory with Applications to Differential Equations*, CBMS Reg. Conf. Ser. Math. 65, AMS, Providence, RI, 1986.

[7] S. B. Robinson, *On the average value for nonconstant eigenfunctions of the p-Laplacian assuming Neumann boundary data*, in Proceedings of the Fifth Mississippi State Conference on Differential Equations and Computation Simulations, Electron. J. Differ. Equ. Conf. 10, Southwest Texas State University, San Marcos, TX, 2002, pp. 251–256.

[8] S. B. Robinson, A. J. Rumbos, and V. L. Shapiro, *One-sided resonance problems for quasilinear elliptic operators*, J. Math. Anal. Appl., 256 (2001), pp. 636–649.

[9] W. Rudin, *Real and Complex Analysis*, 2nd ed., McGraw–Hill, New York, 1974.

[10] A. J. Rumbos and V. L. Shapiro, *One-sided resonance for a quasilinear variational problem*, in Harmonic Analysis and Nonlinear Differential Equations, Contemp. Math. 208, M. L. Lapidus, L. H. Harper, and A. J. Rumbos, eds., AMS, Providence, RI, 1997, pp. 277–299.

[11] V. L. Shapiro, *Quasilinear ellipticity and the first eigenvalue*, Comm. Partial Differential Equations, 16 (1991), pp. 1819–1855.

[12] V. L. Shapiro, *Superlinear quasilinearity and the second eigenvalue*, Nonlinear Anal., 44 (2001), pp. 81–96.

[13] M. Struwe, *Variational Methods: Applications to Nonlinear Partial Differential Equations and Hamiltonian Systems*, Springer-Verlag, New York, 1990.

SIAM J. MATH. ANAL.
Vol. 35, No. 5, pp. 1250–1310

STOCHASTIC NAVIER–STOKES EQUATIONS FOR TURBULENT FLOWS*

R. MIKULEVICIUS† AND B. L. ROZOVSKII†

Abstract. This paper concerns the fluid dynamics modelled by the stochastic flow

$$\begin{cases} \dot{\boldsymbol{\eta}}(t,x) = \boldsymbol{u}(t,\boldsymbol{\eta}(t,x)) + \boldsymbol{\sigma}(t,\boldsymbol{\eta}(t,x)) \circ \dot{W}, \\ \boldsymbol{\eta}(0,x) = x, \end{cases}$$

where the turbulent term is driven by the white noise $\dot{W}$. The motivation for this setting is to understand the motion of fluid parcels in turbulent and randomly forced fluid flows. Stochastic Euler equations for the undetermined components $\boldsymbol{u}(t,x)$ and $\boldsymbol{\sigma}(t,x)$ of the spatial velocity field are derived from the first principles. The resulting equations include as particular cases the deterministic and randomly forced counterparts of these equations.

In the second part of the paper, we prove the existence and uniqueness of a strong local solution to the stochastic Navier–Stokes equation in $W_p^1(\boldsymbol{R}^d), d > 1, p > d$. In the two-dimensional case, the existence and uniqueness of a global strong solution is shown.

In the third part, we deal with the propagation of Wiener chaos by the stochastic Navier–Stokes equation and its relation to statistical moments of the solution.

Key words. stochastic Navier–Stokes, turbulence, Kraichnan's turbulence, Wiener chaos, moments

AMS subject classifications. 60H15, 35R60, 76M35

DOI. 10.1137/S0036141002409167

1. Introduction.

The relation of the Navier–Stokes equation to the phenomenon of hydrodynamic turbulence is widely regarded as one of the most fascinating problems of fluid mechanics. The onset of turbulence is often related to the randomness of background movement. One way to model this is to consider a randomly forced Navier–Stokes equation. Bensoussan and Temam [3] have pioneered the analytical study of a Navier–Stokes equation driven by white noise type random force. Later, this approach was substantially developed and extended by many authors (see, e.g., [4], [5], [7], [13], [15], [21], [35], [41], [51], [52], etc.).

These papers postulated some form of randomly forced Navier–Stokes equation at the inception point. A somewhat different approach was taken in the recent paper [40]. This paper assumed that the dynamics of the fluid particle was given by the stochastic diffeomorphism

$$\dot{\boldsymbol{\eta}}(t,x) = \boldsymbol{u}(t,\boldsymbol{\eta}(t,x)) + \boldsymbol{\sigma}(t,\boldsymbol{\eta}(t,x)) \circ \dot{W}, \boldsymbol{\eta}(0,x) = x \tag{1.1}$$

with undetermined local characteristics $\boldsymbol{u}(t,x)$, and $\boldsymbol{\sigma}(t,x)$.[1] In this setting, $\dot{W}$ is a time derivative of a Hilbert space valued Brownian motion (e.g., space-time white

*Received by the editors June 3, 2002; accepted for publication (in revised form) April 18, 2003; published electronically January 30, 2004. This work was partially supported by NSF grant DMS-98-02423, ONR grant N00014-97-1-0229, and ARO grant DAAG55-98-1-0418.

http://www.siam.org/journals/sima/35-5/40916.html

†Department of Mathematics and Center of Applied Mathematical Sciences, USC, Los Angeles, CA 90089-1113 (mikulvcs@math.usc.edu, rozovski@math.usc.edu).

[1]Here and throughout the rest of the paper, vector fields on $\boldsymbol{R}^d$ are denoted by boldface letters. This convention also applies if the entries of the vector field are taking values in a Hilbert space.

noise) and the stochastic integral is understood in the Stratonovich sense. The generalized random field $\sigma(t,x)\circ\dot{W}$ models the turbulent part of the velocity field, while $\boldsymbol{u}(t,x)$ models its regular component.

The idea of splitting up the velocity field into a sum of slow oscillating (deterministic) and fast oscillating (stochastic) components has often been entertained in fluid mechanics; important developments along these lines may be traced to the work of Reynolds in the 1880s. Our interest in stochastic flows of the form (1.1) stems in part from recent developments in modelling a turbulent velocity field by a generalized Gaussian field $\boldsymbol{V}(t,x)$ with zero mean and covariance $C(x-y,t-s)=K(x-y)\delta(t-s)$ such that the spatial part is of the form

$$K^{ij}(x-y)=A^{ij}+D^{ij}|x-y|^{\kappa} \text{ for } |x-y|<<1,$$

where $\kappa\in(0,2)$ and decays rapidly as $|x-y|\longrightarrow\infty$. This model was pioneered by Kraichnan in his work on turbulent transport [26] and substantially developed later in a series of works by Gawedzki et al. [16], [17] and other authors. The velocity field $\boldsymbol{V}(t,x)$ can be realized by way of its identification with a random field of the form $\boldsymbol{\sigma}(x)\cdot\dot{W}(t)$ (see [2], [31], and section 2.2).

Relating the Kraichnan velocity field to classic fluid mechanics might naturally lead us to ask: "Can we compensate $\boldsymbol{V}(t,x)$ by a field $\boldsymbol{u}(t,x)$ that is more regular with respect to time variable so that there is a balance of momentum for the resulting field $\boldsymbol{U}(t,x)=\boldsymbol{u}(t,x)+\boldsymbol{\sigma}(x)\circ\dot{W}(t)$ or, equivalently, that the motion of a fluid particle modelled by (1.1) satisfies the Newton's second law?"[2]

The answer to this question is positive. Moreover, it turns out that following the classic scheme of Newtonian fluid mechanics (i.e., coupling (1.1) with Newton's second law), a quite general stochastic Navier–Stokes equation,

$$\begin{aligned}\partial_t\boldsymbol{u}&=\Delta\boldsymbol{u}-(\boldsymbol{u},\boldsymbol{\nabla})\boldsymbol{u}-\nabla p+\boldsymbol{f}(\boldsymbol{u})\\&\quad+[(\boldsymbol{\sigma},\boldsymbol{\nabla})u-\nabla\tilde{p}+\boldsymbol{g}(\boldsymbol{u})]\circ\dot{W},\end{aligned}\tag{1.2}$$

may be derived for $\boldsymbol{u}(t,x)$ (see [40] and section 2.2). Special cases of this equation include the standard deterministic Navier–Stokes and Euler equation as well as many other variations of the stochastic Navier–Stokes equation considered in the literature. A more detailed treatment of this subject is given in section 2. To emphasize the relation of equation (1.2) to the flow (1.1) involving the (short time) turbulent component $\boldsymbol{\sigma}(x)\circ\dot{W}(t)$, we will often refer to it as a turbulent stochastic Navier–Stokes equation.

Section 3 deals with the analytical theory of the stochastic Navier–Stokes equation (1.2) and some generalizations of this equation. One technically challenging feature of the SNS equation (1.2) is that it involves multiplicative noise with the diffusion coefficient depending on $\boldsymbol{\nabla u}$. The existence and uniqueness of a maximal local solution in the Sobolev space $W_p^1(\boldsymbol{R}^d)$ for arbitrary $d>1$ and $p>d$ is shown (Theorem 3). Note that owing to the embedding $C^{1-d/p}(\boldsymbol{R}^d)\subset W_p^1(\boldsymbol{R}^d)$, the solution is Hölder continuous. The maximal solution is understood in the (probabilistic) strong sense, e.g., pathwise rather than as a solution of a martingale problem. For the latter, see, e.g., [4], [15], [41], [51]. In the case of $d=2$, it is proved in Theorem 4 that there exists a unique global solution to (1.2).

[2]A priori, it is not clear in what sense the motion described by Kraichnan's velocity might fit into the paradigm of Newtonian mechanics.

We remark that the results of section 3 do not cover the case of only Hölder-continuous $\boldsymbol{\sigma}(x)$, that assumption being important for Kraichnan's turbulent velocity model. This case is addressed in the forthcoming paper [43].

The L_p-theory of strong solutions of SNS equations was studied in [5] (see also references therein). In that paper, the local (global in two dimensions) existence and uniqueness were proved for a randomly forced Navier–Stokes equation

$$\partial_t \boldsymbol{u} = \Delta \boldsymbol{u} - (\boldsymbol{u}, \boldsymbol{\nabla})\boldsymbol{u} - \nabla p + \boldsymbol{f}(\boldsymbol{u}) + \boldsymbol{g}(\boldsymbol{u}) \circ \dot{W} \tag{1.3}$$

in a smooth bounded domain of $\boldsymbol{R}^d$ ($d = 2$ or 3). In this equation, the noise influences the motion of the fluid only by the velocity, rather than by the velocity and its gradient, as is the case in [4], [15], [39], and the present paper. Consequently, it does not cover the case of turbulent stochastic flow.

For a substantial body of related work on L_p-solutions of deterministic Navier–Stokes equations, see, e.g., [18], [24], [25], etc.

Section 4 deals with the propagation of Wiener chaos and moment theory for SNS equations. In this section, we derive a deterministic parabolic system for the Hermite–Fourier coefficients in a Wiener chaos expansion of $\boldsymbol{u}(t,x)$, which we refer to as the "propagator." We show that the statistical moments of the velocity field $\boldsymbol{u}(t,x)$ can be directly expressed via the solution of the propagator. While still an infinite-dimensional system, the propagator for the SNS equation is a much simpler object than the related Kolmogorov equation. On the other hand, it is quite sufficient for dealing with the basic statistical properties of solutions to the SNS equation.[3]

2. Phenomenology of stochastic Navier–Stokes and Euler equations.

2.1. Preliminaries. Classic fluid mechanics deals with two essentially equivalent approaches to modelling the motion of fluid, namely Euclidean and Lagrangian formalisms. The centerpiece of the former is the Navier–Stokes equation for the fluid velocity $\boldsymbol{u}(t,x)$. This equation is expressed in Euclidean coordinates as

$$\begin{cases} \partial_t \boldsymbol{u} + u^l \boldsymbol{u}_{x_l} - \nu \Delta \boldsymbol{u} + \frac{1}{\rho} \boldsymbol{\nabla} p = \boldsymbol{f}, \text{ in } [0, \infty) \times \boldsymbol{R}^d, \\ \\ \boldsymbol{u}(0) = \boldsymbol{u}_0. \end{cases} \tag{2.1}$$

In the case of ideal fluid, (2.1) reduces to Euler equation

$$\begin{cases} \partial_t \boldsymbol{u} + u^l \boldsymbol{u}_{x_l} + \frac{1}{\rho} \boldsymbol{\nabla} p = \boldsymbol{f}, \text{ in } [0, \infty) \times \boldsymbol{R}^d, \\ \\ \boldsymbol{u}(0) = \boldsymbol{u}_0. \end{cases} \tag{2.2}$$

In the case of incompressible fluid, both equations have to be complemented by the equation

$$\operatorname{div} \boldsymbol{u}(t,x) = 0.$$

The Lagrangian formalism emphasizes the dynamics of fluid particles. Let us write $\boldsymbol{\eta}(t,x)$ for the trajectory followed by the fluid particle that is at point x at time $t = 0$. Obviously, $\boldsymbol{\eta}(t) = (\eta^i(t,x), i = 1, \dots, d)$ verifies the equation

$$\partial_t \eta^i(t,x) = u^i(t, \boldsymbol{\eta}(t,x)), \eta^i(0,x) = x^i. \tag{2.3}$$

[3] The main results of the paper were announced in [39].

The function $\boldsymbol{\eta}(t,x)$ is usually referred to as a fluid flow or fluid flow map. Equation (2.3) yields that the fluid flow is defined by $\boldsymbol{u}(t,x)$, a solution of the Navier–Stokes (Euler) equation. On the other hand, one could argue that fluid flow $\boldsymbol{\eta}$ is an equally or even more basic notion than velocity $\boldsymbol{u}$. Indeed, the classic Euclidean approach postulates that the fluid particle motion is given by (2.3) with unknown smooth velocity field $\boldsymbol{u}$; it then shows that this equation, together with Newton's second law, yields (2.1) (see, e.g., [30], [8]). A more recent approach to fluid mechanics pioneered by Arnold, Marsden, and Ebin (see [1], [14]) treats the fluid flow as an intrinsically defined infinite-dimensional dynamical system.

In this paper we consider a flow similar to (2.3) but make the fluid particle subject to turbulent diffusion. The motivation for this setting is to understand the motion of fluid parcels in turbulent and randomly forced fluid flows.

More specifically, we postulate that the fluid particle's motion is given by the equation

$$\dot{\boldsymbol{\eta}}(t,x) = \boldsymbol{u}(t,\boldsymbol{\eta}(t,x)) + \boldsymbol{\sigma}(t,\boldsymbol{\eta}(t,x)) \circ \dot{W}, \boldsymbol{\eta}(0,x) = x, \tag{2.4}$$

where $W(t)$ is a cylindrical Brownian motion in some Hilbert space Y (see [41]), $\dot{W} = \partial W(t)/\partial t$, and $\boldsymbol{u}(t,x)$ and $\boldsymbol{\sigma}(t,x)$ are unknown random fields.

The fluid flow map (2.4) corresponds to the velocity field

$$\boldsymbol{U}(t,x) = \boldsymbol{\sigma}(t,x) \circ \dot{W} + \boldsymbol{u}(t,x). \tag{2.5}$$

The singular term of this field, $\boldsymbol{\sigma}(t,x) \circ \dot{W}$, is referred to as the "turbulent component." If W and σ are statistically independent, e.g., if σ is nonrandom, $\boldsymbol{\sigma}(t,x) \circ \dot{W} := \boldsymbol{\sigma}(t,x) \cdot \dot{W} + \frac{1}{2}\boldsymbol{\sigma}_{x_p}(t,x)\,\sigma^p(t,x)$.

We remark that Kraichnan's turbulence model is an interesting example of the turbulent component in (2.5).

In the generalization of Kraichnan's model introduced in [16] (see also [17]), the turbulent component $\boldsymbol{V}(t,x)$ is modelled by a homogeneous, isotropic, and stationary Gaussian random field with zero mean and covariance

$$EV^i(t,x)\,V^j(s,x) = K^{ij}(x-y)\,\delta(t-s),$$

where $K^{ij}(x-y) = C_0^{ij}\delta_{ij} - D^{ij}(x-y)$. The following asymptotic properties were assumed:

(i) The spatial covariance $K^{ij}(x-y)$ decays fast for $|x-y| > 1$.

(ii) For $|x-y| << 1$,

$$D^{ij}(x-y) = D\left((d+\kappa-1)\,\delta_{ij} - \kappa x_i x_j/|x|^2\right)|x|^\kappa.$$

As illustrated in [31], one possible construction of a homogeneous Gaussian random field with the Kraichnan-type covariance is given by

$$\boldsymbol{\sigma}(x)\cdot\dot{W}(t) = \frac{d}{dt}\left(\sum_{i=1}^{\infty}\boldsymbol{\sigma}_k(x)\,w^k(t)\right),$$

where $w^k(t)$ are independent one-dimensional Brownian motions and $\boldsymbol{\sigma}_k(x)$ are Hölder continuous with an exponent $\kappa/2$ and so that $div\boldsymbol{\sigma}_k(x) = 0$ and $\sum_{ik}|\sigma^{ik}(x)|^2 \le K < \infty$.

In this section we will derive equations for $\boldsymbol{u}(t,x)$ and $\boldsymbol{\sigma}(t,x)$. This will be done by coupling (2.4) with Newton's second law, in much the same way as it is done in classic macroscopic fluid dynamics.

The equations obtained include as particular cases the deterministic Navier–Stokes equation (2.1), as well as the Navier–Stokes equation with stochastic forcing (see [3], [4], [5], [7], [15], [41], [51], etc.).

2.2. Balance of momentum. Let $(\Omega, \mathcal{F}, \boldsymbol{P})$ be a complete probability space and let Y be a separable Hilbert space. Let W be Y-valued cylindrical Brownian motions on $(\Omega, \mathcal{F}, \boldsymbol{P})$. Write $\mathcal{F}_t^W = \sigma(W(s), s \le t)$.

Consider the equation

$$d\boldsymbol{\eta}(t,x) = \boldsymbol{u}(t,\boldsymbol{\eta}(t,x))dt + \boldsymbol{\sigma}(t,\boldsymbol{\eta}(t,x)) \circ dW(t), \boldsymbol{\eta}(0,x) = x, \tag{2.6}$$

where $\circ$ indicates the Stratonovich version of the stochastic integral.

Let us assume the following:

(H1) $\boldsymbol{u}$ is continuous semimartingale given by

$$d\boldsymbol{u}\,(t,x) = \boldsymbol{\alpha}\,(t,x)\,dt + \boldsymbol{\beta}\,(t,x) \circ dW(t), \tag{2.7}$$

where $\boldsymbol{\alpha} : \Omega \times [0,\infty) \times R^d \mapsto R^d$ and $\boldsymbol{\beta} : \Omega \times [0,\infty) \times R^d \mapsto Y^d$ are measurable and $\mathcal{F}_t^W$-adapted functions for every x; $\sigma : [0,\infty) \times R^d \mapsto Y^d$ is a nonrandom measurable function.

In what follows, we shall also assume that, for fixed $t, \boldsymbol{\eta}$ is an invertible mapping; $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and $\boldsymbol{\sigma}$ are appropriately integrable and smooth so that the stochastic integrals are defined and the following manipulations are legitimate. In particular, to define the Stratonovich integral, one needs to assume the existence and some regularity of the joint quadratic variation $\langle \boldsymbol{\beta}\,(\cdot,x)\,\boldsymbol{,}W\rangle_t$ (see [29]).

One fundamental postulate of fluid mechanics (see, e.g., [8]) is Newton's second law: "the rate of change of momentum of a fluid particle equals the force applied to it"; that is,

$$\frac{d}{dt}\dot{\boldsymbol{\eta}}\,(t) = \frac{\boldsymbol{F}\,(t,\boldsymbol{\eta}\,(t))}{\rho(t,\boldsymbol{\eta}(t))}, \tag{2.8}$$

where $\boldsymbol{F}(t,x)$ is the total force applied to the fluid particle and $\rho(t,x)$ is the mass density. For the sake of simplicity, in this paper, we assume that $\rho = 1$.

In our case the acceleration,

$$\frac{d}{dt}\dot{\boldsymbol{\eta}}\,(t) = \frac{d}{dt}\left(\boldsymbol{\sigma}\,(t,\boldsymbol{\eta}\,(t)) \circ \dot{W}\right) + \frac{d}{dt}\boldsymbol{u}\,(t,\boldsymbol{\eta}\,(t))\,,$$

is highly irregular. Thus (2.8) shall be interpreted in the sense of distributions; i.e., for every $\varphi \in C_0^\infty\left(\boldsymbol{R}^1\right)$,

$$\int \varphi\,(t)\,\boldsymbol{F}\,(t,\boldsymbol{\eta}\,(t))\;dt = -\int \varphi'\,(t)\,\boldsymbol{\sigma}\,(t,\boldsymbol{\eta}\,(t)) \circ dW(t) + \int \varphi\,(t)\,d\boldsymbol{u}\,(t,\boldsymbol{\eta}\,(t))\,. \tag{2.9}$$

Obviously, both sides of (2.8) must have the same structure. Hence, formulas (2.9) and (2.11) yield that there exist $\mathcal{F}_t^W$-adapted functions $\boldsymbol{f} : \Omega \times [0,\infty) \times \boldsymbol{R}^d \mapsto \boldsymbol{R}^d, \boldsymbol{g} : \Omega \times [0,\infty) \times \boldsymbol{R}^d \mapsto Y^d$, and $\boldsymbol{d} : \Omega \times [0,\infty) \times \boldsymbol{R}^d \mapsto Y^d$ so that

$$\begin{aligned} &\textstyle\int \varphi\,(t)\,\boldsymbol{F}\,(t,\boldsymbol{\eta}\,(t))\;dt = -\int \varphi'\,(t)\,\boldsymbol{d}\,(t,\boldsymbol{\eta}\,(t)) \circ dW(t) \\ &\textstyle+\int \varphi\,(t)\,(\boldsymbol{f}\,(t,\boldsymbol{\eta}\,(t))\,dt + \boldsymbol{g}(t,\boldsymbol{\eta}\,(t)) \circ dW(t)) \end{aligned} \tag{2.10}$$

By the Itô–Wentzell formula (see, e.g., Theorem 3.3.2 in [29] or Theorem 1.4.9 in [47]),

$$\begin{aligned} d\boldsymbol{u}\left(t,\boldsymbol{\eta}\left(t\right)\right) &= \boldsymbol{\alpha}\left(t,\boldsymbol{\eta}\left(t\right)\right)dt + \boldsymbol{\beta}\left(t,\boldsymbol{\eta}\left(t\right)\right)\circ dW(t) \\ &+ \boldsymbol{u}_{x_i}u^i(t,\boldsymbol{\eta}(t))dt + \boldsymbol{u}_{x_i}\sigma^i(t,\boldsymbol{\eta}(t))\circ dW(t). \end{aligned} \tag{2.11}$$

By matching terms in (2.9) and (2.10) and taking into account (2.11), we obtain the following equalities:

$$\boldsymbol{d} = \boldsymbol{\sigma}, \boldsymbol{g} = \boldsymbol{\beta} + \boldsymbol{u}_{x_i}\sigma^i, \tag{2.12}$$

$$\boldsymbol{\alpha} = -\boldsymbol{u}_{x_j}u^j + \boldsymbol{f}. \tag{2.13}$$

Thus, we arrive at the following equation for the regular velocity component $\boldsymbol{u}$:

$$d\boldsymbol{u} = \left[-\boldsymbol{u}_{x_j}u^j + \boldsymbol{f}\right]dt + \left[\boldsymbol{g}(t,x) - \boldsymbol{u}_{x_p}\sigma^p\left(t,x\right)\right]\circ dW(t). \tag{2.14}$$

2.3. Derivation of stochastic Euler and Navier–Stokes equations.

2.3.1. Incompressible stochastic fluids and Euler equation. A fluid characterized by flow $\boldsymbol{\eta}$ given by (2.4) is incompressible if $\boldsymbol{\eta}(t,x)$ is a volume preserving map. It can be shown that the latter holds iff

$$\operatorname{div}\boldsymbol{\sigma}\left(t,x\right) = \operatorname{div}\boldsymbol{u}\left(t,x\right) = 0 \tag{2.15}$$

Indeed, we may easily see that the Jacobian of $\boldsymbol{\eta}$ verifies the equation

$$\begin{aligned} dJ\eta\left(t\right) &= J\boldsymbol{\eta}\left(t\right)\left\{\operatorname{div}\boldsymbol{\sigma}\left(t,\boldsymbol{\eta}\left(t\right)\right)\cdot dW\left(t\right) + \operatorname{div}\boldsymbol{u}\left(t,\boldsymbol{\eta}\left(t\right)\right)\,dt\right. \\ &\left. + \left(1/2\right)\left[\left|\operatorname{div}\boldsymbol{\sigma}\left(t,\boldsymbol{\eta}\left(t\right)\right)\right|_Y^2 + \left(\partial_j\operatorname{div}\boldsymbol{\sigma}\right)(t,\boldsymbol{\eta}(t))\cdot\sigma^j(t,\boldsymbol{\eta}(t))\right]dt\right\}. \end{aligned} \tag{2.16}$$

The rest of the proof is similar to the case of $\boldsymbol{\sigma} = 0$ (see, e.g., [8]).

Suppose that the fluid is ideal (nonviscous). Similar to that found in the classic setting, we can assume that the force acting on the fluid particle is of the form $\boldsymbol{F} = -\boldsymbol{\nabla}P + \bar{\boldsymbol{F}}$, where P is the (unknown) pressure and $\bar{\boldsymbol{F}}$ is the given body force. More specifically, we assume that $\boldsymbol{f} = -\boldsymbol{\nabla}P^a + \bar{\boldsymbol{f}}$, $\boldsymbol{g} = -\boldsymbol{\nabla}P^d + \bar{\boldsymbol{g}}$, and $\boldsymbol{d} = -\boldsymbol{\nabla}P^t + \bar{\boldsymbol{d}}$. The body force components are considered to be given, while those of the pressure are subject to determination.

$$\begin{cases} d\boldsymbol{u} = [-\boldsymbol{u}_{x_i}u^i - \boldsymbol{\nabla}P^a + \bar{\boldsymbol{f}}]dt + \left(\bar{\boldsymbol{g}} - \boldsymbol{\nabla}P^d - \boldsymbol{u}_{x_i}\sigma^i\right)\circ dW; \\ \boldsymbol{\sigma}\left(t,x\right) = -\boldsymbol{\nabla}P^t\left(t,x\right) + \bar{\boldsymbol{d}}\left(t,x\right),\ \operatorname{div}\boldsymbol{u} = 0, \operatorname{div}\boldsymbol{\sigma} = 0; \\ \boldsymbol{u}(0,x) = \boldsymbol{u}_0(x). \end{cases} \tag{2.17}$$

Since $\operatorname{div}\boldsymbol{u}=\operatorname{div}\boldsymbol{\sigma} = 0$, we have $\Delta P^t = \operatorname{div}\bar{\boldsymbol{d}}$, $\Delta P^d = \operatorname{div}\bar{\boldsymbol{g}}$, and

$$\Delta P^a = \operatorname{div}\left[\bar{\boldsymbol{f}} - \boldsymbol{u}_{x_i}u^i\right].$$

The number of equations equals the number of unknown functions, and so mathematically this is a reasonable system.

Write $2a^{ij} = \sigma^i \cdot \sigma^j$. Since $\operatorname{div}\boldsymbol{\sigma} = 0$, the first equation in (2.17) can be rewritten in the Itô form as follows:

$$d\boldsymbol{u} = \left[(a^{ij}\boldsymbol{u}_{x_i})_{x_j} - \boldsymbol{u}_{x_j}u^j - \frac{1}{2}\left(\boldsymbol{g}_{x_p}\sigma^p - \boldsymbol{G}\right) + \boldsymbol{f}\right]dt$$

$$+ \left[\boldsymbol{g} - \boldsymbol{u}_{x_p}\sigma^p\right] \cdot dW.$$

In spite of the presence of the "effective viscosity" term $(a^{ij}\boldsymbol{u}_{x_i}(t,x))_{x_j}$, which is induced by the turbulent term, we shall still regard (2.17) as a *stochastic Euler* equation. First, it was derived for the ideal fluid. Second, (2.17) passes the ultimate test for Euler type equations, namely, it conserves the energy. Specifically, it can be easily shown that if there are no free forces, $\bar{\boldsymbol{f}} = \bar{\boldsymbol{g}} = \bar{\boldsymbol{h}} = 0$, then

$$\int |\boldsymbol{u}(t,x)|^2 dx = \int |\boldsymbol{u}(0,x)|^2 dx \quad P\text{-a.s.} \tag{2.18}$$

Besides, in "appearance,"(2.17) bears more of a resemblance to the deterministic Euler equation, since it does not contain the second order term. A special case of equation (2.17) was derived (very informally) in [21] using the variational formulation of the Euler equation. In that paper it was assumed that $\boldsymbol{\sigma} = \text{const}$, $\bar{g} = \bar{h} = 0$ and W was a one-dimensional Brownian motion.

Now let us consider some comparatively straightforward generalizations of the setup considered above. Let V be a Y-valued cylindrical Brownian motion independent of W . Write $\mathcal{F}_t^{W,V} = \sigma(W(s), V(s), s \le t)$.

Assume

(H1') $d\boldsymbol{u}(t,x) = \boldsymbol{\alpha}(t,x)\,dt + \boldsymbol{\beta}(t,x) \cdot dW(t) + \boldsymbol{\gamma}(t,x) \cdot dV(t)$,

where $\boldsymbol{\alpha} : \Omega \times [0,\infty) \times R^d \mapsto R^d, \boldsymbol{\beta} : \Omega \times [0,\infty) \times R^d \mapsto Y^d$, and $\boldsymbol{\gamma} : \Omega \times [0,\infty) \times R^d \mapsto Y^d$ are $\mathcal{F}_t^{W,V}$-adapted functions.

The stochastic integrals in (H1') are understood in the Itô sense.

REMARK 1. *The Itô setting has its advantages and disadvantages. One advantage of the Itô formulation is that it does not require the existence of joint quadratic variations* $\langle \boldsymbol{g}(\cdot,x), W\rangle_t$, $\langle \boldsymbol{\beta}(\cdot,x), W\rangle_t$, *and* $\langle \boldsymbol{\gamma}(\cdot,x), V\rangle_t$, *which is a necessary assumption for the existence of the related Stratonovich integrals. On the other hand, if the fluid particle's motion was given by the Itô equation*

$$d\boldsymbol{\eta}(t,x) = \boldsymbol{u}(t,\boldsymbol{\eta}(t,x))dt + \boldsymbol{\sigma}(t,\boldsymbol{\eta}(t,x)) \cdot dW(t), \tag{2.19}$$

the equations (2.15) *would no longer guarantee that* $\boldsymbol{\eta}(t,x)$ *is a volume preserving map. Instead, a more cumbersome condition would be needed. Therefore, we will continue with the Stratonovich form* (2.6).

Of course, owing to (H1'), the balance of momentum considerations yield that the force must be of the form

$$\boldsymbol{F} = \frac{d}{dt}\left(\boldsymbol{d} \circ \dot{W}\right) + \boldsymbol{f} + \boldsymbol{g} \cdot \dot{W}(t) + \boldsymbol{h} \cdot \dot{V}(t).$$

The interested reader could prove that, in the new setting, we have

$$\begin{cases} d\boldsymbol{u}\,(t,x) = (\left(a^{ij}\boldsymbol{u}_{x_i}\right)_{x_j} - \boldsymbol{u}_{x_i}u^i - \boldsymbol{g}_{x_i}\cdot\sigma^i - \boldsymbol{\nabla}P^a + \bar{\boldsymbol{f}})dt \\ + \left(\bar{\boldsymbol{g}} - \boldsymbol{\nabla}P^d - \boldsymbol{u}_{x_i}\sigma^i\right)\cdot dW(t) + \left(\bar{\boldsymbol{h}} - \boldsymbol{\nabla}\tilde{P}^d\right)\cdot dV(t); \\ \boldsymbol{\sigma}\,(t,x) = -\boldsymbol{\nabla}P^t\,(t,x) + \bar{\boldsymbol{d}}\,(t,x)\,,\ \operatorname{div}\boldsymbol{u} = 0, \operatorname{div}\boldsymbol{\sigma} = 0; \\ \boldsymbol{u}(0,x) = \boldsymbol{u}_0(x), \end{cases}$$

where $\boldsymbol{h} = -\boldsymbol{\nabla}\tilde{P}^d + \bar{\boldsymbol{h}}$ (for details see [40]) .

2.3.2. Stochastic Navier–Stokes equation. Let us now drop assumption (2.17) and assume that the fluid we are dealing with is viscous. This requires the appropriate modification of the structure of forces acting on the fluid particle. Because of the molecular motion of particles, the force exerted per unit area on an arbitrary surface S in the fluid has a component of the form

$$\nu\nabla\boldsymbol{U}\,(x,t)\,\vec{n} = \nu[\nabla\boldsymbol{\sigma}\,(t,x)\circ\dot{W} + \nabla\boldsymbol{u}\,(t,x)]\vec{n},$$

where $\vec{n}$ is the unit normal to S (see [8]). By the divergence theorem, this implies the following structure of forces: $\boldsymbol{f} = -\boldsymbol{\nabla}P^a + \nu\Delta\boldsymbol{u} + \frac{1}{2}\Delta\boldsymbol{\sigma}_{x_p}\,(t,x)\,\Delta\sigma^p\,(t,x) + \bar{\boldsymbol{f}}$, $\boldsymbol{g} = -\boldsymbol{\nabla}P^d + \Delta\sigma + \bar{\boldsymbol{g}}$, $\boldsymbol{h} = -\boldsymbol{\nabla}\tilde{P}^d + \bar{\boldsymbol{h}}$. The resulting stochastic Navier–Stokes equation for the components of the velocity field (2.5) is as follows:

$$\begin{cases} d\boldsymbol{u} = [\nu\Delta\boldsymbol{u} + (a^{ij}\boldsymbol{u}_{x_i})_{x_j} - \boldsymbol{u}_{x_i}u^i - \left(\bar{\boldsymbol{g}}_{x_i} - \boldsymbol{\nabla}P^d_{x_i}\right)\cdot\sigma^i \\ -\boldsymbol{\nabla}P^a + \bar{\boldsymbol{f}} + \frac{1}{2}\Delta\boldsymbol{\sigma}_{x_p}\,(t,x)\cdot\Delta\sigma^p\,(t,x)]dt \\ + \left(\bar{\boldsymbol{h}} - \boldsymbol{\nabla}\tilde{P}^d\right)\cdot dV(t) + \left(\bar{\boldsymbol{g}} + \nu\Delta\sigma - \boldsymbol{\nabla}P^d - \boldsymbol{u}_{x_i}\sigma^i\right)\cdot dW; \\ \boldsymbol{\sigma}\,(t,x) = -\boldsymbol{\nabla}P^t\,(t,x) + \bar{\boldsymbol{d}}\,(t,x)\,,\ \operatorname{div}\boldsymbol{u} = 0, \operatorname{div}\boldsymbol{\sigma} = 0; \boldsymbol{u}(0,x) = \boldsymbol{u}_0(x), \end{cases} \tag{2.20}$$

where ν is the viscosity coefficient.

2.3.3. Special cases. Now let us review several important particular cases.

1. Assume that the stochastic components of the force $\boldsymbol{g} = \boldsymbol{h} = \boldsymbol{d} = 0$. Then, by (2.12) , $\boldsymbol{\sigma} = \boldsymbol{\gamma} = \boldsymbol{\beta} = 0$, and we arrive at the standard deterministic Euler equation.

2. Assume that there is no turbulent component in (2.4) and the force has no turbulent component, i.e., $\boldsymbol{\sigma} = \boldsymbol{g} = 0$. Then, by (2.12) , $\boldsymbol{d} = \boldsymbol{\beta} = 0$, and the stochastic Navier–Stokes equation reduces to the following Navier–Stokes equation with random forcing:

$$\begin{cases} d\boldsymbol{u} = [\nu\Delta\boldsymbol{u} - \boldsymbol{u}_{x_i}u^i - \boldsymbol{\nabla}P^a + \bar{\boldsymbol{f}}]dt + \boldsymbol{h}\cdot dV(t); \\ \operatorname{div}\boldsymbol{u} = 0, \boldsymbol{u}(0,x) = \boldsymbol{u}_0(x). \end{cases}$$

3. Assume that $\nu = 0$, $\boldsymbol{\beta} = 0$ and $\boldsymbol{\alpha}, \boldsymbol{\gamma}$ are $\mathcal{F}^V_t$-adapted. Then, by (2.12) , we have

$$\boldsymbol{\alpha} = -\boldsymbol{u}_{x_i}u^j + \boldsymbol{f} - \left(\boldsymbol{u}_{x_i}a^{ij}\right)_{x_j}\,,\ \boldsymbol{g} = \boldsymbol{u}_{x_p}\sigma^p. \tag{2.21}$$

We then arrive at the following equation for $\boldsymbol{u}$:

$$\begin{cases} d\boldsymbol{u} = [-\boldsymbol{u}_{x_i} u^i - \boldsymbol{\nabla} P^a + \bar{\boldsymbol{f}} \\ \\ - \left(\bar{\boldsymbol{g}} - \boldsymbol{\nabla} P^d\right)_{x_j} \cdot \left(\bar{d}^j - P^t_{x_j}\right)]dt + \left(\bar{\boldsymbol{h}} - \boldsymbol{\nabla}\tilde{P}^d\right) \cdot dV(t); \\ \\ \operatorname{div} \boldsymbol{u} = 0, \boldsymbol{u}(0,x) = \boldsymbol{u}_0(x). \end{cases} \tag{2.22}$$

In addition, (2.21) yields

$$\Delta P^d = u^i_{x_p} \sigma^p_{x_i} - \operatorname{div}\bar{\boldsymbol{g}}. \tag{2.23}$$

Obviously, if $\boldsymbol{h} = 0$, the dynamics of the nonturbulent component $\boldsymbol{u}$ of the velocity field $\boldsymbol{U} = \boldsymbol{\sigma} \circ \dot{W} + \boldsymbol{u}$ is given by a deterministic Euler type equation.

We remark that the above results rectify the statement in [40] that (2.17) is ill-posed if $\boldsymbol{\beta} = 0$.

3. Analytical theory of turbulent stochastic Navier–Stokes equations.

3.1. Preliminaries.

3.1.1. Notation. We begin by outlining some of the notation that will be used in the paper.

$\boldsymbol{R}^d$ denotes a d-dimensional Euclidean space with elements $x = (x_1, \dots, x_d)$; if $x, y \in \boldsymbol{R}^d$, we write

$$(x,y) = \sum_{i=1}^d x_i y_i,\ |x| = \sqrt{(x,x)}.$$

Let us fix a separable Hilbert space Y. The scalar product of $x, y \in Y$ will be denoted by $x \cdot y$.

If u is a function on $\boldsymbol{R}^d$, the following notational conventions will be used for its partial derivatives: $\partial_i u = \partial u/\partial x_i, \partial^2_{ij} = \partial^2 u/\partial x_i \partial x_j$, $\partial_t u = \partial u/\partial t$, and $\nabla u = \partial u = (\partial_1 u, \dots, \partial_d u)$, and $\partial^2 u = (\partial^2_{ij} u)$ denotes the Hessian matrix of second derivatives. Let $\alpha = (\alpha_1, \dots, \alpha_d)$ be a multi-index; then $\partial^\alpha_x = \Pi_{i=1}^d \partial^{\alpha_i}_{x_i}$.

Let $C_0^\infty = C_0^\infty(\boldsymbol{R}^d)$ be the set of all infinitely differentiable functions on $\boldsymbol{R}^d$ with compact support.

For $s \in (-\infty, \infty)$, write $\Lambda^s = \Lambda^s_x = (1 - \sum_{i=1}^d \partial^2/\partial x_i^2)^{s/2}$.

For $p \in [1, \infty]$ and $s \in (-\infty, \infty)$, we define the space $H^s_p = H^s_p(\boldsymbol{R}^d)$ as the space of generalized functions u with the finite norm

$$|u|_{s,p} = |\Lambda^s u|_p,$$

where $|\cdot|_p$ is the L_p norm. Obviously, $H^0_p = L_p$. Note that if $s \geq 0$ is an integer, the space H^s_p coincides with the Sobolev space $W^s_p = W^s_p(\boldsymbol{R}^d)$.

If $p \in [1, \infty]$ and $s \in (-\infty, \infty)$, $H^s_p(Y) = H^s_p(\boldsymbol{R}^d, Y)$ denotes the space of Y-valued functions on $\boldsymbol{R}^d$ so that the norm $||g||_{s,p} = |\,|\Lambda^s g|_Y\,|_p < \infty$. We also write $L_p(Y) = L_p(\boldsymbol{R}^d, Y) = H^0_p(Y) = H^0_p(\boldsymbol{R}^d, Y)$. Let $C_0^\infty(Y)$ be the space of Y-valued infinitely differentiable functions on $\boldsymbol{R}^d$ with compact support.

Obviously, the spaces $C_0^\infty, C_0^\infty(Y)$, $H_p^s(\boldsymbol{R}^d)$, and $H_p^s(\boldsymbol{R}^d, Y)$ can be extended to vector functions (denoted by bold-faced letters). For example, the space of all vector functions $\boldsymbol{u} = (u^1, \dots, u^d)$ such that $\Lambda^s u^l \in L_p$, $l = 1, \dots, d$, with the finite norm

$$|\boldsymbol{u}|_{s,p} = \left(\sum_l |u^l|_{s,p}^p \right)^{1/p},$$

we denote by $\mathbb{H}_p^s = \mathbb{H}_p^s(\boldsymbol{R}^d)$. Similarly, we denote by $\mathbb{H}_p^s(Y) = \mathbb{H}_p^s(\boldsymbol{R}^d, Y)$ the space of all vector functions $g = (g^l)_{1 \le l \le d}$, with Y-valued components g^l, $1 \le l \le d$, so that $||g||_{s,p} = (\sum_l |g^l|_{s,p}^p)^{1/p} < \infty$. The set of all infinitely differentiable vector functions $u = (u^1, \dots, u^d)$ on $\boldsymbol{R}^d$ with compact support will be denoted by $\mathbb{C}_0^\infty$. We denote by $\mathbb{C}_0^\infty(Y)$ the set of all infinitely differentiable vector functions $\boldsymbol{u} = (u^1, \dots, u^d)$ on $\boldsymbol{R}^d$ with compact support (all u^l are Y-valued).

When $s = 0$, $\mathbb{H}_p^s(Y) = \mathbb{L}_p(Y) = \mathbb{L}_p(\boldsymbol{R}^d, Y)$. Also, in this case, the norm $||g||_{0,p}$ is denoted more briefly by $||g||_p$. To forcefully distinguish L_p-norms in spaces of Y-valued functions, we write $||\cdot||_p$, while in all other cases a norm is denoted by $|\cdot|$.

The duality $\langle \cdot, \cdot \rangle_s$ between $\mathbb{H}_q^s(\boldsymbol{R}^d)$ and $\mathbb{H}_p^{-s}(\boldsymbol{R}^d), p \ge 2, s \in (-\infty, \infty)$, and $q = p/(p-1)$ is defined by

$$\langle \boldsymbol{\phi}, \boldsymbol{\psi} \rangle_s = \langle \boldsymbol{\phi}, \boldsymbol{\psi} \rangle_{s,p} = \sum_{i=1}^d \int_{\boldsymbol{R}^d} \left[\Lambda^s \phi^i \right](x) \, \Lambda^{-s} \psi^i(x) \, dx, \boldsymbol{\phi} \in \mathbb{H}_q^s, \boldsymbol{\psi} \in \mathbb{H}_p^{-s}.$$

3.1.2. Solenoidal and gradient projections of Hilbert-valued vector fields. In this section we present some facts about solenoidal and gradient projections of vector fields, most of which were proved in [37].

We will use the Riesz transform for the definition of the projections. We set for $f \in L_2(\boldsymbol{R}^d, Y)$,

$$R_j(f)(x) = \lim_{\varepsilon \to 0} c_* \int_{|y| \ge \varepsilon} \frac{y_j}{|y|^{d+1}} f(x-y) \, dy, \ j = 1, \dots, d,$$

with $c_* = G(\frac{n+1}{2})/\pi^{(n+1)/2}$ (G is the gamma function). R_j is called a Riesz transform. According to [49] (see Chapter III, formula (8), p. 58),

$$\widehat{R_j f}(x) = -i \frac{\xi_j}{|\xi|} \widehat{f},$$

where

$$\widehat{f}(\xi) = \mathcal{F}(f) = (2\pi)^{-d/2} \int e^{-i(\xi, x)} f(x) \, dx.$$

Given a function $f \in L_p(\boldsymbol{R}^d, Y)$, we define a vector Riesz transform $Rf = (R_1 f, \dots, R_d f)$.

For $\boldsymbol{v} \in \mathbb{L}_2(Y)$, set (see [24], [25])

$$\mathcal{G}(\boldsymbol{v}) = -R R_j v^j, \mathcal{S}(\boldsymbol{v}) = \boldsymbol{v} - \mathcal{G}(\boldsymbol{v}).$$

Then (see [24], [25], and Lemma 2.7 in [37]), $\mathbb{L}_2(Y)$ is a direct sum

$$\mathbb{L}_2(Y) = \mathcal{G}(\mathbb{L}_2(Y)) \oplus \mathcal{S}(\mathbb{L}_2(Y)),$$
$$\mathcal{S}(\mathbb{L}_2(Y)) = \{ \boldsymbol{g} \in \mathbb{L}_2(Y) : \operatorname{div} \boldsymbol{g} = 0 \},$$

and $\mathcal{G}(\mathbb{L}_2(Y))$ is a Hilbert subspace orthogonal to $\mathcal{S}(\mathbb{L}_2(Y))$.

REMARK 2. *If $f \in C_0^\infty(\boldsymbol{R}^d)$, it is known (see, e.g., [44]) that the classical solution to*

$$\Delta u(x) = f(x),\ x \in \boldsymbol{R}^d \tag{3.1}$$

is given by the formula

$$u(x) = \int \Gamma(x-y) f(y)\, dy, \tag{3.2}$$

where

$$\Gamma(x-y) = \begin{cases} |x-y|^{2-d}/d(2-d)\omega_d, & d > 2, \\ \frac{1}{2\pi} \ln|x-y|, & d = 2 \end{cases}$$

and ω_d is the volume of the unit ball in $\boldsymbol{R}^d$. If $\boldsymbol{f} \in \mathbb{C}_0^\infty(Y)$, we may easily show that

$$\mathcal{G}(\boldsymbol{f}) = \boldsymbol{\nabla} \int \Gamma_{x_i}(x-y) f^i(y)\, dy = -RR_j f^j. \tag{3.3}$$

The functions $\mathcal{G}(\boldsymbol{v})$ and $\mathcal{S}(\boldsymbol{v})$ are usually referred to as the potential and the solenoidal projections, respectively, of the vector field $\boldsymbol{v}$.

The following statement holds.

LEMMA 1 (see [24], [25], and Lemmas 2.11 and 2.12 in [37]). *$\mathcal{G}, \mathcal{S}$ can be extended continuously to all $\mathbb{H}_p^s(Y), s \in (-\infty, \infty)$: there is a constant C so that for all $\boldsymbol{v} \in \mathbb{H}_p^s(Y)$*

$$||\mathcal{G}(\boldsymbol{v})||_{s,p} \le C||\boldsymbol{v}||_{s,p}, ||\mathcal{S}(\boldsymbol{v})||_{s,p} \le C||\boldsymbol{v}||_{s,p}.$$

Moreover, the space $\mathbb{H}_p^s(Y)$ can be decomposed into the direct sum

$$\mathbb{H}_p^s(Y) = \mathcal{G}(\mathbb{H}_p^s(Y)) \oplus \mathcal{S}(\mathbb{H}_p^s(Y)),$$

and, if $(1/p) + (1/q) = 1$, $\boldsymbol{f} \in \mathcal{G}(\mathbb{H}_p^s(Y))$, $\boldsymbol{g} \in \mathcal{S}(\mathbb{H}_q^{-s}(Y))$, then

$$\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathbb{H}_p^s(Y), \mathbb{H}_q^{-s}(Y)} = 0. \tag{3.4}$$

Also,

$$\mathcal{S}(\mathbb{H}_p^s(Y)) = \{\boldsymbol{v} \in \mathbb{H}_p^s(Y) : \operatorname{div} \boldsymbol{v} = \boldsymbol{0}\}. \tag{3.5}$$

3.2. Strong solutions of the Navier–Stokes equation in $\boldsymbol{R^d}$.

3.2.1. Main results. Let $(\Omega, \mathcal{F}, \boldsymbol{P})$ be a probability space with a filtration $\mathbb{F}$ of right continuous σ-algebras $(\mathcal{F}_t)_{t \ge 0}$. All the σ-algebras are assumed to be $\boldsymbol{P}$-completed. Let $W(t)$ be an $\mathbb{F}$-adapted cylindrical Brownian motion in Y.

For $\boldsymbol{v} \in \mathbb{H}_p^1$, let $\boldsymbol{G}(\boldsymbol{v}, t) = \boldsymbol{G}(\boldsymbol{v}, t, x)$ be a predictable $\mathbb{L}_p(Y)$-valued function and $\boldsymbol{F}(\boldsymbol{v}, t) = \boldsymbol{F}(\boldsymbol{v}, t, x)$ a predictable $\mathbb{L}_p$-valued function. Let us consider the following Navier–Stokes equation:

$$\begin{cases} \partial_t u^l(t,x) = \partial_i \left(a^{ij}(t,x)\, \partial_j u^l(t,x) \right) - u^k(t,x)\, \partial_k u^l(t,x) \\ -\partial_l P(t,x) + b^i(t,x) \partial_i u(t,x) + F^l(\boldsymbol{u}(t), t, x) \\ +[\sigma^i(t,x) \partial_i u^l(t,x) + G^l(\boldsymbol{u}(t), t, x) - \partial_l \tilde{P}(t,x)]\, \dot{W}_t, \\ \operatorname{div} \boldsymbol{u} = 0, \boldsymbol{u}(0,x) = \boldsymbol{u}_0(x), \ \ l = 1, \dots d. \end{cases} \tag{3.6}$$

In vector form, the equation would be

(3.7)
$$\partial_t \boldsymbol{u}(t) = \partial_i \left(a^{ij}(t)\,\partial_j \boldsymbol{u}(t)\right) - u^k(t)\,\partial_k \boldsymbol{u}(t) - \boldsymbol{\nabla} P(t) + b^i(t)\partial_i \boldsymbol{u}(t) + \boldsymbol{F}(\boldsymbol{u}(t),t)$$
$$+\ \boldsymbol{F}(\boldsymbol{u}(t),t) + \left[\sigma^i(t)\partial_i \boldsymbol{u}(t) + \boldsymbol{G}(\boldsymbol{u}(t),t) - \boldsymbol{\nabla}\tilde{P}(t)\right]\dot{W}_t,$$
$$\boldsymbol{u}(0) = \boldsymbol{u}_0, \operatorname{div} \boldsymbol{u} = 0.$$

Of course, the unknowns in the equation (3.7) are the functions $\boldsymbol{u} = (u^l)_{1\le l\le d}$, P, and $\tilde{P}$.

Everywhere in this section it is assumed that $p \ge 2$. The vector field $\boldsymbol{u}_0$ is always $\mathcal{F}_0$-measurable and $\operatorname{div} \boldsymbol{u}_0 = 0$.

It is assumed that a^{ij}, b^i are measurable $\mathbb{F}$-adapted functions on $[0,\infty)\times \boldsymbol{R}^d$ and the matrix $\left(a^{ij}\right)$ is symmetric. Let us assume also that σ^i are Y-valued measurable $\mathbb{F}$-adapted functions on $[0,\infty)\times \boldsymbol{R}^d$.

In addition, we will need the following assumptions.

B1. $\boldsymbol{P}$-a.s.

$$\sum_{k=0}^{1}(|\partial^k a^{ij}| + |\partial^k b^i| + |\partial^k \sigma^i|_Y) \le K;$$

for all $t \ge 0$, $x, \lambda \in \boldsymbol{R}^d$, we have

$$K|\lambda|^2 \ge \left[a^{ij}(t,x) - \frac{1}{2}\sigma^i(t,x)\cdot\sigma^j(t,x)\right]\lambda^i\lambda^j \ge \delta|\lambda|^2,$$

where K, δ are fixed strictly positive constants (notice that this assumption excludes the Euler equation).

B2(p). For all $\boldsymbol{v} \in \mathbb{H}^1_p, t > 0$,

$$|\boldsymbol{F}(\boldsymbol{v},t) - \boldsymbol{F}(\bar{\boldsymbol{v}},t)|_p \le C|\boldsymbol{v}-\bar{\boldsymbol{v}}|_p, ||\boldsymbol{G}(\boldsymbol{v},t) - \boldsymbol{G}(\bar{\boldsymbol{v}},t)||_p \le C|\boldsymbol{v}-\bar{\boldsymbol{v}}|_p,$$

and for all $t > 0, \boldsymbol{v} \in \mathbb{H}^1_p$,

$$|\,|\boldsymbol{G}(\boldsymbol{v},t)|\,|_{1,p} \le |\,|\boldsymbol{G}(\boldsymbol{0},t)|\,|_{1,p} + C|\boldsymbol{v}|_{1,p}, |\boldsymbol{F}(\boldsymbol{v},t)|_{1,p} \le |\boldsymbol{F}(\boldsymbol{0},t)|_{1,p} + C|\boldsymbol{v}|_{1,p}.$$

Suppose also that

$$\int_0^t (|\,|\boldsymbol{G}(\boldsymbol{0},r)|\,|^p_{1,p} + |\boldsymbol{F}(\boldsymbol{0},r)|^p_{1,p})\,dr < \infty$$

$\boldsymbol{P}$-a.s. for all t.

B3(p). For each M, there is a constant C such that for all $\boldsymbol{v}, \bar{\boldsymbol{v}} \in B_{M,p} = \{\boldsymbol{v} \in \mathbb{H}^1_p : |\boldsymbol{v}|_{1,p} \le M\}, t > 0$

$$||\nabla(\boldsymbol{G}(\boldsymbol{v},t) - \boldsymbol{G}(\bar{\boldsymbol{v}},t))||_p \le C|\boldsymbol{v}-\bar{\boldsymbol{v}}|_{1,p}.$$

Since $\operatorname{div} \boldsymbol{u} = 0$, we have

(3.8)
$$\operatorname{div}\left(\sigma^i(t)\partial_i \boldsymbol{u}(t) + \boldsymbol{G}(\boldsymbol{u}(t),t) - \boldsymbol{\nabla}\tilde{P}(t)\right) = 0$$
and
$$\operatorname{div}\left[-u^k(t)\,\partial_k \boldsymbol{u}(t)) + \partial_i\left(a^{ij}(t)\,\partial_j \boldsymbol{u}(t)\right) + \boldsymbol{F}(\boldsymbol{u}(t),t) - \boldsymbol{\nabla} P(t)\right] = 0.$$

Then, if the expressions in the left-hand sides of both equations in (3.8) belong to $\mathbb{H}^1_p$ for some $p > 1$, by Remark 2 we have

$$\boldsymbol{\nabla}\tilde{P}(t,x) = \mathcal{G}\left(\sigma^i(t)\partial_i \boldsymbol{u}\,(t) + \boldsymbol{G}(\boldsymbol{u}(t),t)\right),$$

and

$$\boldsymbol{\nabla}P(t,x) = \mathcal{G}[-u^k\,(t))\partial_k \boldsymbol{u}(t)) + \partial_i\left(a^{\,ij}\,(t)\,\partial_j \boldsymbol{u}\,(t)\right) + \boldsymbol{F}\,(\boldsymbol{u}(t),t)].$$

So, in $\mathbb{L}_p$-theory, instead of (3.7), we can and will consider its equivalent form

$$\begin{cases} \partial_t \boldsymbol{u}\,(t) = \mathcal{S}[\partial_i\left(a^{\,ij}\,(t)\,\partial_j \boldsymbol{u}\,(t)\right) - u^k\,(t)\,\partial_k \boldsymbol{u}(t) + b^i(t)\partial_i \boldsymbol{u}(t) + \boldsymbol{F}\,(\boldsymbol{u}\,(t)\,,t)] \\ \\ +\mathcal{S}[\sigma^i(t)\partial_i \boldsymbol{u}\,(t) + \boldsymbol{G}\,(\boldsymbol{u}\,(t)\,,t)]\,\dot{W}_t, \boldsymbol{u}\,(0) = \boldsymbol{u}_0. \end{cases} \tag{3.9}$$

Given a stopping time τ, we define a stochastic interval

$$[[0,\tau(\omega)]] = \begin{cases} [0,\tau(\omega)] & \text{if } \tau(\omega) < \infty, \\ [0,\infty), & \text{otherwise.} \end{cases}$$

Let $s \in \{0,1\}$.

DEFINITION 2. *Given a stopping time τ, an $\mathbb{H}^s_p(\boldsymbol{R}^d)$-valued $\mathbb{F}$-adapted function $\boldsymbol{u}(t)$ on $[0,\infty)$ is called an $\mathbb{H}^s_p$-solution of equation* (3.7) *(or* (3.9)*) in $[[0,\tau]]$ if it is strongly continuous in t with probability* 1*,*

$$\boldsymbol{u}(t) = \boldsymbol{u}(t\wedge\tau) \text{ and } \int_0^{t\wedge\tau} |\boldsymbol{u}(r)|^p_{s+1,p}\,dr < \infty \ \forall t > 0, \boldsymbol{P}-a.s., \tag{3.10}$$

and the equality

$$\begin{array}{c} \boldsymbol{u}(t) = \boldsymbol{u_0} + \int_0^{t\wedge\tau} \mathcal{S}[-u^i\,(r)\,\partial_i \boldsymbol{u}\,(r) + \partial_i(a^{ij}(r)\partial_j \boldsymbol{u}\,(r)) + \boldsymbol{F}(\boldsymbol{u}(r),r)]dr \\ + \int_0^{t\wedge\tau} \mathcal{S}(\sigma^k(r)\partial_k \boldsymbol{u}\,(r) + \boldsymbol{G}(\boldsymbol{u}(r),r))\cdot dW\,(r) \end{array} \tag{3.11}$$

holds in $\mathbb{H}^{s-1}_p(\boldsymbol{R}^d)$ for every $t > 0$, $\boldsymbol{P}$-a.s.

If $\tau = \infty$, we simply say $\boldsymbol{u}$ is an $\mathbb{H}^s_p$-solution of equation (3.6). The stochastic integral in (3.11) is defined in the appendix.

Sometimes, when the context is clear, instead of "$\mathbb{H}^s_p$-solution" we will just say "solution."

If an $\mathbb{H}^s_p$-solution in $[[0,\tau]]$ is also an $\mathbb{H}^s_q$-solution in $[[0,\tau]]$, we call it an $\mathbb{H}^s_p \cap \mathbb{H}^s_q$-solution in $[[0,\tau]]$.

EXAMPLE 1. *Let f^l be measurable $\mathbb{F}$-adapted functions on $[0,\infty)\times\boldsymbol{R}^d\times\boldsymbol{R}^d$. Let $h^{l,i}$ be Y-valued measurable $\mathbb{F}$-adapted functions on $[0,\infty)\times\boldsymbol{R}^d$, and g^l be Y-valued measurable $\mathbb{F}$-adapted functions on $[0,\infty)\times\boldsymbol{R}^d\times\boldsymbol{R}^d$. Given $\boldsymbol{v}\in\mathbb{H}^1_p$, define*

$$\begin{aligned} \boldsymbol{G}(\boldsymbol{v},t) &= (g^l(t,x,\boldsymbol{v}(x)))_{1\le l\le d} \\ \boldsymbol{F}(\boldsymbol{v},t) &= (f^l(t,x,\boldsymbol{v}(x))_l + (h^{l,j}(t,x)L_j(t,x,\boldsymbol{v}))_{1\le l\le d}, \end{aligned} \tag{3.12}$$

where $\boldsymbol{L}(t,x,\boldsymbol{v}) = (L_l(t,x,\boldsymbol{v}))_{1\le l\le d} = \mathcal{G}[\sigma^k(t)\partial_k\boldsymbol{v} + \boldsymbol{G}(\boldsymbol{v},t)]$.

Assume that, for each $t \geq 0, x \in \boldsymbol{R}^d$, $u \in \boldsymbol{R}^d$,

$$\sum_{k=0}^{1} |\partial_x^k \boldsymbol{g}(s,x,u)| \leq G_1(s,x) + K|u|,$$

$$\sum_{k=0}^{1} |\partial_x^k \boldsymbol{f}(s,x,u)| \leq F_1(s,x) + K|u|, \tag{3.13}$$

$$|\partial_u \boldsymbol{g}| + |\partial_u \boldsymbol{f}| + \sum_{k=0}^{1} (|\partial_x^k h^{l,j}|_Y + |\partial_x^k \sigma^j|_Y) \leq C,$$

and $\boldsymbol{P}$*-a.s. for all* t

$$\int_0^t (|G_1(r)|_p^p + |F_1(r)|_p^p)\, dr < \infty. \tag{3.14}$$

The assumptions (3.13), (3.14) *imply the assumption* $\mathbf{B2}(p)$ *for* $\boldsymbol{G}, \boldsymbol{F}$ *defined by* (3.12).

The assumptions (3.13), (3.14) *and the boundedness of* $\partial_u^2 \boldsymbol{g}(t,x,u)$ *imply the assumption* $\mathbf{B3}(p)$ *for* $\boldsymbol{G}$.

Now we can formulate the main theorems on local and global existence and uniqueness.

THEOREM 3. (a) *Let* $\mathbf{B1}$, $\mathbf{B2}(p)$, $\mathbf{B3}(p)$ *be satisfied* $(p > d)$ *and* $\boldsymbol{E}(|\boldsymbol{u}_0|_{1,p}^p) < \infty$.

Then there is a unique predictable stopping time ζ, $\boldsymbol{P}(\zeta > 0) = 1$ *such that for each stopping time* $S, [0,S] \subseteq [0,\zeta)$ *if and only if there is a* $\mathbb{H}_p^1$*-valued continuous* $\mathbb{L}_p$*-solution to* (3.7) *in* $[[0,S]]$.

Also, there is a unique $\mathbb{H}_p^1$*-valued continuous process* $\boldsymbol{u}(t)$ *on* $[0,\zeta)$ *such that* $\limsup_{t\uparrow\zeta} |\boldsymbol{u}(t)|_{1,p} = \infty$ *on* $\{\zeta < \infty\}$, *and* $\boldsymbol{u}(t \wedge S)$ *is an* $\mathbb{L}_p$*-solution of* (3.7) *in* $[[0,S]]$ *for each* S *so that* $[0,S] \subseteq [0,\zeta)$.

Moreover, if $\boldsymbol{E}(|\boldsymbol{u}_0|_{2-2/p,p}^p) < \infty$, *then* $\boldsymbol{u}(t)$ *is also an* $\mathbb{H}_p^1$*-solution of* (3.7) *in* $[[0,S]]$ *for all stopping times* S *such that* $[0,S] \subseteq [0,\zeta)$ *and* $\lim_{t\uparrow\zeta} |\boldsymbol{u}(t)|_{1,p} = \infty$ *on* $\{\zeta < \infty\}$ $\boldsymbol{P}$*-a.s.*

(b) *Let* $\mathbf{B1}$, $\mathbf{B2}(p)$, $\mathbf{B3}(p)$, $\mathbf{B2}(2)$, $\mathbf{B3}(2)$ *be satisfied, and*

$$E(|\boldsymbol{u}_0|_{1,p}^p + |\boldsymbol{u}_0|_{1,2}^p) < \infty,$$

$$\int_0^t (||\boldsymbol{G}(\boldsymbol{0},r)||_{1,2}^p + |\boldsymbol{F}(\boldsymbol{0},r)|_{1,2}^p)\, dr < \infty$$

$\boldsymbol{P}$*-a.s. for all* t. *Then there is a unique predictable stopping time* ζ, $\boldsymbol{P}(\zeta > 0) = 1$ *such that, for each stopping time* $S, [0,S] \subseteq [0,\zeta)$ *if and only if there is an* $\mathbb{H}_p^1 \cap \mathbb{H}_2^1$*-valued continuous* $\mathbb{L}_p \cap \mathbb{L}_2$*-solution to* (3.7) *in* $[[0,S]]$.

Also, there is a unique $\mathbb{H}_p^1 \cap \mathbb{H}_2^1$*-valued continuous process* $\boldsymbol{u}(t)$ *on* $[0,\zeta)$ *such that* $\limsup_{t\uparrow\zeta} (|\boldsymbol{u}(t)|_{1,p} + |\boldsymbol{u}(t)|_{1,2}) = \infty$ *on* $\{\zeta < \infty\}$, *and* $\boldsymbol{u}(t \wedge S)$ *is an* $\mathbb{L}_p \cap \mathbb{L}_2$*-solution to* (3.7) *on* $[0,S]$ *for each* S *so that* $[0,S] \subseteq [0,\zeta)$.

Moreover, if $\boldsymbol{E}(|\boldsymbol{u}_0|_{2-2/p,p}^p) < \infty$, *then* $\boldsymbol{u}(t)$ *is also an* $\mathbb{H}_p^1 \cap \mathbb{H}_2^1$*-solution of* (3.7) *in* $[[0,S]]$ *for all stopping times* S *such that* $[0,S] \subseteq [0,\zeta)$.

In both cases, $(u(t),\zeta)$ is called a maximal solution to (3.7) and ζ is called its explosion time.

If $d = 2$, a stronger result holds. Specifically, there is a unique global solution.

THEOREM 4. *Let* **B1**, **B2**(p), **B3**(p), **B2**(2), **B3**(2) *be satisfied,* $p > d = 2$, *and*

$$E(|\boldsymbol{u}_0|^p_{2-2/p,p} + |\boldsymbol{u}_0|^p_{1,2}) < \infty,$$

$$\int_0^t (||\boldsymbol{G}(\boldsymbol{0}, r)||^p_{1,2} + |\boldsymbol{F}(\boldsymbol{0}, r)|^p_{1,2})\, dr < \infty$$

$\boldsymbol{P}$*-a.s. for all* t.

Then there is a maximal unique $\mathbb{H}^1_p \cap \mathbb{H}^1_2$*-solution* $(\boldsymbol{u}(t), \zeta)$ *of* (3.7) *and* $\boldsymbol{P}(\zeta = \infty) = 1$.

Moreover, for each $T > 0$ *there is a constant* C *such that, for all stopping times* $\tau \le T$,

$$E \sup_{s\le\tau}(|\boldsymbol{u}(t)|^p_{1,p} + |\boldsymbol{u}(t)|^p_{1,2}) \le C[E(|\boldsymbol{u}_0|^p_{1,p} + |\boldsymbol{u}_0|^p_{1,2})$$

$$+ \textstyle\int_0^\tau (||\boldsymbol{G}(\boldsymbol{0}, r)||^p_{1,p} + |\boldsymbol{F}(\boldsymbol{0}, r)|^p_{1,p} + ||\boldsymbol{G}(\boldsymbol{0}, r)||^p_{1,2} + |\boldsymbol{F}(\boldsymbol{0}, r)|^p_{1,2})\, dr].$$

Proof of these theorems will be given in sections 3.3–3.6.

3.3. Mollified Navier–Stokes equation. In this section we consider an auxiliary equation obtained from (3.7) by applying the standard mollifier to the first term of the Navier–Stokes nonlinearity $(\boldsymbol{u} \cdot \boldsymbol{\nabla})\, \boldsymbol{u}$.

Let $\psi(x) \in C_0^\infty(\boldsymbol{R}^d), \psi \ge 0, \int \psi\, dx = 1$. Given a scalar function v on $\boldsymbol{R}^d$, we define

$$\Psi^\varepsilon(v)(x) = \begin{cases} \int v(x-y)\psi_\varepsilon(y)\, dy, & \varepsilon > 0, \\ v, & \varepsilon = 0, \end{cases}$$

where $\psi_\varepsilon(x) = \varepsilon^{-d}\psi(x/\varepsilon)$, $\varepsilon > 0$. Similarly, for a vector function $\boldsymbol{v}$,

$$\Psi^\varepsilon(\boldsymbol{v})(x) = \begin{cases} (\int v^l(x-y)\psi_\varepsilon(y)\, dy)_l, & \varepsilon > 0, \\ \boldsymbol{v}(x), & \varepsilon = 0. \end{cases}$$

For a fixed $\varepsilon \ge 0$, we consider the equation for $\boldsymbol{u} = \left(u^l\right)_{1\le l\le d}$, P, $\tilde{P}$

$$\begin{aligned} \partial_t \boldsymbol{u}(t,x) &= \partial_i(a^{ij}(t,x)\partial_j \boldsymbol{u}) - \Psi^\varepsilon(u^k(t))\partial_k \boldsymbol{u}(t) + \boldsymbol{D}(\boldsymbol{u}(t), t, x) - \boldsymbol{\nabla} P(t,x) \\ &\quad + [\sigma^k(t,x)\partial_k \boldsymbol{u}(t,x) + \boldsymbol{G}(\boldsymbol{u}(t), t, x) - \boldsymbol{\nabla}\tilde{P}(t,x)] \cdot \dot{W}, \\ \boldsymbol{u}(0,x) &= \boldsymbol{u}_0(x),\ \operatorname{div} \boldsymbol{u} = 0, \end{aligned} \tag{3.15}$$

where $\boldsymbol{u}(t) = \boldsymbol{u}(t,x) = (u^k(t,x))_{1\le k\le d}$ and $\boldsymbol{D}(\boldsymbol{v}, t) = b^i(t)\partial_i \boldsymbol{v} + \boldsymbol{F}(\boldsymbol{v}, t)$.

Obviously, if $\varepsilon = 0$, (3.15) coincides with (3.7).

Similarly to (3.7), (3.15) is equivalent to

$$\begin{aligned} \partial_t \boldsymbol{u}(t) &= \mathcal{S}[\partial_i(a^{ij}(t)\partial_j \boldsymbol{u}(t)) - \Psi^\varepsilon(u^k(t))\partial_k \boldsymbol{u}(t) + \boldsymbol{D}(\boldsymbol{u}(t), t))] \\ &\quad + \mathcal{S}[\sigma^k(t)\partial_k \boldsymbol{u}(t) + \boldsymbol{G}(\boldsymbol{u}(t), t)] \cdot \dot{W}, \\ \boldsymbol{u}(0) &= \boldsymbol{u}_0. \end{aligned} \tag{3.16}$$

For $\varepsilon > 0$, we will solve (3.16) in $\mathbb{H}^s_p, s \in (-\infty, \infty), p \ge 2$.

DEFINITION 5. *Given a stopping time* τ, *an* $\mathbb{H}^s_p(\boldsymbol{R}^d)$*-valued* $\mathbb{F}$*-adapted function* $\boldsymbol{u}(t)$ *on* $[0, \infty)$ *is called an* $\mathbb{H}^s_p$*-solution of equation* (3.15) *(or* (3.16)*) in* $[[0, \tau]]$ *if it is strongly continuous in* t *with probability* 1;

$$\boldsymbol{u}(t) = \boldsymbol{u}(t \wedge \tau),\ \int_0^{t\wedge\tau} |\boldsymbol{u}(r)|^p_{s+1,p}\, dr < \infty\ \forall t > 0, \boldsymbol{P} - a.s., \tag{3.17}$$

and the equality

$$\begin{aligned}\boldsymbol{u}(t) &= \boldsymbol{u_0} + \int_0^{t\wedge\tau} \mathcal{S}[-\Psi^\varepsilon(u^i)\partial_i \boldsymbol{u} + \partial_i(a^{ij}(r)\partial_j \boldsymbol{u}) + \boldsymbol{D}(\boldsymbol{u})]dr \\ &\quad + \int_0^{t\wedge\tau} \mathcal{S}(\sigma^k \partial_k \boldsymbol{u} + \boldsymbol{G}(\boldsymbol{u})) \cdot dW(r) \end{aligned} \tag{3.18}$$

holds in $\mathbb{H}_p^{s-1}(\boldsymbol{R}^d)$ *for every* $t > 0$, $\boldsymbol{P}$*-a.s.*

If $\tau = \infty$, we simply say $\boldsymbol{u}$ is an $\mathbb{H}_p^s$-solution of (3.15).

If an $\mathbb{H}_p^s$-solution in $[[0,\tau]]$ is also an $\mathbb{H}_q^s$-solution in $[[0,\tau]]$, we call it an $\mathbb{H}_p^s \cap \mathbb{H}_q^s$-solution in $[[0,\tau]]$.

In this subsection, we fix $\varepsilon > 0$ and consider the corresponding equation (3.15) (equivalently (3.16)).

For an integer $s > 0$, we denote

$$\mathcal{C}^s(Y) = \left\{ u \in C^{s-1} : ||u||_{\mathcal{C}^s} = \sum_{|a|\le s-1} ||\partial^\alpha u||_\infty + \sum_{|\alpha|=s-1} \sup_{x\neq y} \frac{|\partial^\alpha u(x) - \partial^\alpha u(y)|_Y}{|x-y|} < \infty \right\}.$$

Define

$$B^s(Y) = \begin{cases} H_\infty^s(Y) & \text{if } s > 0 \text{ is not an integer,} \\ \mathcal{C}^s(Y) & \text{if } s > 0 \text{ is an integer,} \\ L_\infty(Y) & \text{if } s = 0, \end{cases}$$

and denote the corresponding norms by $|\cdot|_{B^s}$.

The following assumptions will often be used in the future.

A. For all $t \ge 0$, $x, \lambda \in R^d$,

$$K|\lambda|^2 \ge \left[a^{ij}(t,x) - \frac{1}{2}\sigma^i(t,x)\cdot\sigma^j(t,x) \right] \lambda^i\lambda^j \ge \delta|\lambda|^2,$$

where K, δ are fixed strictly positive constants.

A1(s,p). For all t, x, y, $\boldsymbol{P}$-a.s.

$$|a^{ij}(t,x) - a^{ij}(t,y)| + |\sigma^i(t,x) - \sigma^i(t,y)|_Y \le K|x-y|$$

and

$$\begin{cases} |a^{ij}(t)|_{B^s} \le K & \text{if } s \ge 1, \\ |a(t,x)| \le K & \text{if } -1 < s < 1, \\ |a^{ij}(t)|_{B^{-s+\varepsilon}} \le K & \text{if } s \le -1. \end{cases}$$

where $\varepsilon \in (0,1)$.

For all i, t, x,

$$\begin{cases} ||\sigma^i(t)||_{B^s} \le K & \text{if } s \ge 1, \\ |\sigma^i(t,x)|_Y \le K & \text{if } s \in (-1,1), \\ ||\sigma^i(t)||_{B^{-s+\varepsilon}} \le K & \text{if } s \le -1, \end{cases}$$

where $\varepsilon \in (0,1)$.

A2(s,p). For $\boldsymbol{v}\in\mathbb{H}^{s+1}_p$, $\boldsymbol{G}(\boldsymbol{v},t)=\boldsymbol{G}(\boldsymbol{v},t,x)$ is a predictable $\mathbb{H}^s_p(Y)$-valued function and $\boldsymbol{D}(\boldsymbol{v},t)=\boldsymbol{D}(\boldsymbol{v},t,x)$ is a predictable $\mathbb{H}^{s-1}_p$-valued function, and $\boldsymbol{P}$-a.s. for each t

$$\int_0^t (|\boldsymbol{D}(\boldsymbol{0},r)|^p_{s-1,p}+||\boldsymbol{G}(\boldsymbol{0},r)||^p_{s,p})\,dr<\boldsymbol{\infty}\ \forall t>0,\boldsymbol{P}\text{-a.s.},$$

where $\boldsymbol{0}=(0,\dots,0)$.

A3(s,p). For every $\varepsilon>0$, there exists a constant K_ε such that, for any $\boldsymbol{u},\boldsymbol{v}\in\mathbb{H}^{s+1}_p$,

$$|\boldsymbol{D}(\boldsymbol{u},t,x)-\boldsymbol{D}(\boldsymbol{v},t,x)|_{s-1,p}+||\boldsymbol{G}(\boldsymbol{u},t,x)-\boldsymbol{G}(\boldsymbol{v},t,x)||_{s,p}$$

$$\le\varepsilon|\boldsymbol{u}-\boldsymbol{v}|_{s+1,p}+K_\varepsilon|\boldsymbol{u}-\boldsymbol{v}|_{s-1,p}\quad\boldsymbol{P}-a.s.$$

We start with the following statement.

PROPOSITION 6. *Let $s\in(-\infty,\infty),p\in[2,\infty)$. Assume* **A**, **A1**$(s,p)$–**A3**$(s,p)$ *are satisfied and $\boldsymbol{E}(|\boldsymbol{u}_0|^p_{s+1-2/p,p})<\infty$. Then there is a unique predictable stopping time ζ, $\boldsymbol{P}(\zeta>0)=1$ such that, for each stopping time S, $[0,S]\subseteq[0,\zeta)$ if and only if there is a unique $\mathbb{H}^s_p$-solution to* (3.15) *in* $[[0,S]]$;

Also, there is a unique $\mathbb{H}^s_p$-valued continuous process $\boldsymbol{u}(t)$ on $[0,\zeta)$ such that $\boldsymbol{P}$-a.s. $\limsup_{t\uparrow\zeta}|\boldsymbol{u}(t)|_{s,p}=\infty$ on $\{\zeta<\infty\}$, and $\boldsymbol{u}(t\wedge S)$ is a solution to (3.15) *in* $[[0,S]]$ *for each S so that $[0,S]\subseteq[0,\zeta)$. Moreover, for each $T>0,M>1$ there is a constant C, such that for each stopping time $\tau\le T\wedge\tau_M$*

$$E\left[\sup_{r\le\tau}|\boldsymbol{u}(r)|^p_{s,p}+\int_0^\tau|\partial^2\boldsymbol{u}(r)|^p_{s-1,p}\,dr\right]\le CE\left[|\boldsymbol{u}_0|^p_{s+1-2/p,p}+\int_0^\tau(|\boldsymbol{D}(\boldsymbol{0},r)|^p_{s-1,p}\right.$$

$$\left.+|\boldsymbol{G}(\boldsymbol{0},r)|^p_{s,p})\,dr\right],\tag{3.19}$$

where $\tau_M=\inf(t:|\boldsymbol{u}(t)|_{s,p}\ge M)$.

(The pair $(\boldsymbol{u}(t),\zeta)$ is called a maximal $\mathbb{H}^s_p$-solution of (3.15)*.)*

Proof. For each $M>0$, we define a function on $\mathbb{H}^s_p$

$$\varphi_M(\boldsymbol{u})=\begin{cases}\boldsymbol{u} & \text{if } |\boldsymbol{u}|_{s,p}\le M,\\ M|\boldsymbol{u}|^{-1}_{s,p}\boldsymbol{u}, & \text{otherwise.}\end{cases}$$

For every $\boldsymbol{u},\bar{\boldsymbol{u}}\in\mathbb{H}^s_p$, we obviously have $|\varphi_M(\boldsymbol{u})|_{s,p}\le M$ and

$$|\varphi_M(\boldsymbol{u})-\varphi_M(\bar{\boldsymbol{u}})|_{s,p}\le 2|\boldsymbol{u}-\bar{\boldsymbol{u}}|_{s,p}.$$

Define a function $\boldsymbol{B}^k_M(\boldsymbol{u})=u^k_\varepsilon\varphi_M(\boldsymbol{u}),\boldsymbol{u}\in\mathbb{H}^s_p$, where $u^k_\varepsilon=\Psi^\varepsilon(u^k)$. There is a constant C so that for each $\boldsymbol{u},\boldsymbol{v}\in\mathbb{H}^s_p$

$$|\boldsymbol{B}^k_M(\boldsymbol{u})-\boldsymbol{B}^k_M(\boldsymbol{v})|_{s,p}\le CM|\boldsymbol{u}-\boldsymbol{v}|_{s,p}.\tag{3.20}$$

Indeed, if $|\boldsymbol{u}|_{s,p}\le M,|\boldsymbol{v}|_{s,p}\le M$, then

$$\boldsymbol{B}^k_M(\boldsymbol{u})-\boldsymbol{B}^k_M(\boldsymbol{v})=u^k_\varepsilon\boldsymbol{u}-v^k_\varepsilon\boldsymbol{v}=(u^k_\varepsilon-v^k_\varepsilon)\boldsymbol{u}+v^k_\varepsilon(\boldsymbol{u}-\boldsymbol{v}),$$

and, by Lemma 7 in [38],

$$|\boldsymbol{B}_M^k(\boldsymbol{u})-\boldsymbol{B}_M^k(\boldsymbol{v})|_{s,p}\le |u_\varepsilon^k-v_\varepsilon^k|_{B^{|s|}}|\boldsymbol{u}|_{s,p}+|v_\varepsilon^k|_{B^{|s|}}|\boldsymbol{u}-\boldsymbol{v}|_{s,p}\le CM|\boldsymbol{u}-\boldsymbol{v}|_{s,p}.$$

If $|\boldsymbol{u}|_{s,p}\le M, |\boldsymbol{v}|_{s,p}>M$, then

$$\begin{aligned}\boldsymbol{B}_M^k(\boldsymbol{u})-\boldsymbol{B}_M^k(\boldsymbol{v})&=u_\varepsilon^k\boldsymbol{u}-v_\varepsilon^k\boldsymbol{v}M|\boldsymbol{v}|_{s,p}^{-1}\\&=|\boldsymbol{v}|_{s,p}^{-1}[(u_\varepsilon^k-v_\varepsilon^k)\boldsymbol{v}M+Mu_\varepsilon^k(\boldsymbol{u}-\boldsymbol{v})+u_\varepsilon^k\boldsymbol{u}(|\boldsymbol{v}|_{s,p}-M)],\end{aligned}$$

and, by Lemma 7 in [38],

$$\begin{aligned}|\boldsymbol{B}_M^k(\boldsymbol{u})-\boldsymbol{B}_M^k(\boldsymbol{v})|_{s,p}&\le C[|u_\varepsilon^k-v_\varepsilon^k|_{B^{|s|}}M+M|u_\varepsilon^k|_{B^{|s|}}|\boldsymbol{u}-\boldsymbol{v}|_{s,p}\\&\quad+|u_\varepsilon^k|_{B^{|s|}}|\boldsymbol{u}-\boldsymbol{v}|_{s,p}]\le CM|\boldsymbol{u}-\boldsymbol{v}|_{s,p}.\end{aligned}$$

Similarly, if $|\boldsymbol{u}|_{s,p}>M, |\boldsymbol{v}|_{s,p}>M$, then

$$\boldsymbol{B}_M^k(\boldsymbol{u})-\boldsymbol{B}_M^k(\boldsymbol{v})=u_\varepsilon^k\boldsymbol{u}M|\boldsymbol{u}|_{s,p}^{-1}-v_\varepsilon^k\boldsymbol{v}M|\boldsymbol{v}|_{s,p}^{-1}$$

$$=M|\boldsymbol{u}|_{s,p}^{-1}|\boldsymbol{v}|_{s,p}^{-1}[(u_\varepsilon^k-v_\varepsilon^k)\boldsymbol{u}|\boldsymbol{v}|_{s,p}+v_\varepsilon^k(\boldsymbol{u}-\boldsymbol{v})|\boldsymbol{u}|_{s,p}+v_\varepsilon^k(\boldsymbol{u}|\boldsymbol{v}|_{s,p}-|\boldsymbol{u}|_{s,p})],$$

and

$$\begin{aligned}|\boldsymbol{B}_M^k(\boldsymbol{u})-\boldsymbol{B}_M^k(\boldsymbol{v})|_{s,p}&\le C[|u_\varepsilon^k-v_\varepsilon^k|_{B^{|s|}}M+M|v_\varepsilon^k|_{B^{|s|}}|\boldsymbol{u}-\boldsymbol{v}|_{s,p}|\boldsymbol{v}|_{s,p}^{-1}\\&\quad+|v_\varepsilon^k|_{B^{|s|}}|\boldsymbol{u}-\boldsymbol{v}|_{s,p}|\boldsymbol{v}|_{s,p}^{-1}]\le CM|\boldsymbol{u}-\boldsymbol{v}|_{s,p}.\end{aligned}$$

So, (3.20) holds and therefore

$$|\boldsymbol{B}_M^k(\boldsymbol{u})|_{s,p}\le CM|\boldsymbol{u}|_{s,p}. \tag{3.21}$$

Since (3.20), (3.21) hold, then, according to Theorem 3.3 in [37] and Remark 5.5 in [28], for each M, there is a unique $\mathbb{H}_p^s$-solution $\boldsymbol{u}=\boldsymbol{u}_M$ of the equation

$$\begin{aligned}\partial_t\boldsymbol{u}(t)&=\partial_i(a^{ij}(t)\partial_j\boldsymbol{u}(t))-\partial_k\boldsymbol{B}_M^k(\boldsymbol{u}(t))+\boldsymbol{D}(\boldsymbol{u}(t),t)+\nabla p(t)\\&\quad+[\sigma^k(t)\partial_k\boldsymbol{u}(t)+\boldsymbol{G}(\boldsymbol{u}(t),t)+\nabla\tilde{p}(t)]\cdot\dot{W},\\\boldsymbol{u}(0)&=\boldsymbol{u}_0,\operatorname{div}\boldsymbol{u}=0.\end{aligned}$$

Let $\tau_M=\inf\{t:|\boldsymbol{u}_M(t)|_{s,p}\ge M\}$. By Corollary 3.6 in [37], $\boldsymbol{P}$-a.s.

$$\boldsymbol{u}_M(t\wedge\tau_M)=\boldsymbol{u}_{M'}(t\wedge\tau_M)\text{ for all }t \tag{3.22}$$

if $M'>M$. Following the proof of Theorem 14.21 in [22], we consider the set $\mathcal{S}$ of all stopping times S such that a $\mathbb{H}_p^s$-solution to (3.15) exists on $[0,S]$. Obviously, $\mathcal{S}$ is not empty ($\tau_M\in\mathcal{S}$ for all M). It is closed with respect to the finite minimum and finite maximum operations. Let ζ be the essential upper bound of the set $\mathcal{S}$. So, there is a sequence $T_n\in\mathcal{S}$ increasing to ζ. Let $\boldsymbol{U}_n$ be a corresponding sequence of solutions in $[0,T_n]$. The sequence $\boldsymbol{U}_n$ defines a solution $\boldsymbol{u}$ on $\cup_n[0,T_n]$. Let $y_t=|\boldsymbol{u}(t)|_{s,p}^p$, $R_m=\zeta\wedge\inf(t:y_t\ge m)$. Then $\boldsymbol{u}(\cdot\wedge T_q\wedge R_m)$ is a solution in $[0,T_q\wedge R_m]$. Passing to a limit as $q\to\infty$, we obtain that $\boldsymbol{u}(\cdot\wedge R_m)$ is a solution in $[0,R_m]$. If $\boldsymbol{P}(R_m=\zeta<\infty)>0$, then (3.22) would imply that there is a stopping time $S\in\mathcal{S}$ such that $S\ge R_m$ and $\boldsymbol{P}(R_m=\zeta<S)>0$. This would contradict the definition

of ζ. Thus $\boldsymbol{P}$-a.s. $R_m < \zeta$ on $\{\zeta < \infty\}$, and $\limsup_{t\uparrow\zeta} y_t = \infty$ on $\{\zeta < \infty\}$. So the sequence (R_m) "announces" ζ and ζ is a predictable stopping time. Let S be a stopping time such that $\boldsymbol{P}$-a.s. $S < \zeta$. Then $\boldsymbol{u}(\cdot \wedge S)$ is a solution in $[0, S]$: it is enough to notice that $\boldsymbol{u}(\cdot \wedge R_q \wedge S)$ is a solution in $[0, R_q \wedge S]$ and pass to the limit as $q \to \infty$.

Let $\tau_M = \inf(t : y_t \geq m)$. Since $\boldsymbol{u}(\cdot \wedge \tau_M) = \boldsymbol{u}_M(\cdot \wedge \tau_M)$, it follows by Theorem 3.3 in [37] that for each T and M there is a constant C so that for each stopping time $\tau \leq \tau_M \wedge T$ and all t,

$$
\begin{aligned}
E\left[\sup_{r\leq\tau} |\boldsymbol{u}(r\wedge\tau)|^p_{s,p} + \int_0^{t\wedge\tau} |\partial^2\boldsymbol{u}(r)|^p_{s-1,p}\,dr\right] \\
\leq CE\left[|\boldsymbol{u}_0|^p_{s+1,p} + \int_0^{t\wedge\tau} (|\boldsymbol{D}(\boldsymbol{0},r)|^p_{s-1,p} + |u^k_\varepsilon(r)\,\boldsymbol{u}(r)|^p_{s,p} + |\boldsymbol{G}(\boldsymbol{0},r)|^p_{s,p})\,dr\right] \\
\leq CE\left[|\boldsymbol{u}_0|^p_{s+1,p} + \int_0^{t\wedge\tau} (|\boldsymbol{D}(\boldsymbol{0},r)|^p_{s-1,p} \quad + |\boldsymbol{u}(r)|^p_{s,p} + |\boldsymbol{G}(\boldsymbol{0},r)|^p_{s,p})\,dr\right].
\end{aligned}
$$

So, inequality (3.19) follows by the Gronwall lemma. □

COROLLARY 7. *Let $s \in (-\infty,\infty), p \in [2,\infty)$. Assume* **A,** $\mathbf{A1}(s,p)$–$\mathbf{A3}(s,p)$. *Assume further* $\mathbf{A1}(s,q)$–$\mathbf{A3}(s,q)$ *for $q \geq 2$, and suppose that $\boldsymbol{E}(|\boldsymbol{u}_0|^p_{s+1-2/p,p} + |\boldsymbol{u}_0|^q_{s+1-2/q,q}) < \infty$. Then the maximal unique $\mathbb{H}^s_p$-solution $(\boldsymbol{u},\boldsymbol{\zeta})$ of* (3.15) *defined in Proposition* 6 *is also a maximal unique $\mathbb{H}^s_q$-solution of the equation. Moreover, for each $T > 0, M > 1$, there is a constant C such that for each stopping time $\tau \leq T \wedge \tau_M$,*

$$
\begin{aligned}
E\left[\sup_{r\leq\tau} |\boldsymbol{u}(r)|^l_{s,l} + \int_0^\tau |\partial^2\boldsymbol{u}(r)|^l_{s-1,l}\,dr\right] \leq CE\Bigg[|\boldsymbol{u}_0|^l_{s+1,l} + \int_0^\tau (|\boldsymbol{D}(\boldsymbol{0},r)|^l_{s-1,l} \\
+ |\boldsymbol{G}(\boldsymbol{0},r)|^l_{s,l})\,dr\Bigg],\ l = p, q,
\end{aligned}
$$

where $\tau_M = \inf(t : |\boldsymbol{u}(t)|_{s,p} \geq M)$.

(The pair $(\boldsymbol{u}(t),\zeta)$ is called a maximal $\mathbb{H}^1_p \cap \mathbb{H}^1_q$-solution.)

Proof. Let $(\boldsymbol{u}(t),\zeta)$ be the maximal $\mathbb{H}^s_p$-solution $\boldsymbol{u}$ of (3.15), $\tau_M = \inf\{t : |\boldsymbol{u}(t)|_{s,p} \geq M\}$. Consider the equation for $\boldsymbol{\xi}$:

$$
\begin{aligned}
\partial_t\boldsymbol{\xi}(t) = \mathcal{S}[\partial_i(a^{ij}(t)\partial_j\boldsymbol{\xi}(t)) - \partial_k(\Psi^\varepsilon(u^k(t\wedge\tau_M))\boldsymbol{\xi}(t)) + \boldsymbol{D}(\boldsymbol{\xi}(t),t)] \\
+ \mathcal{S}[\sigma^k(t)\partial_k\boldsymbol{\xi}(t) + \boldsymbol{G}(\boldsymbol{\xi}(t),t)]\cdot\dot{W},\quad \boldsymbol{\xi}(0) = \boldsymbol{u}_0(x).
\end{aligned}
$$

By Theorem 3.3, Corollary 3.7, and Corollary 3.6 in [37], $\boldsymbol{\xi}(t) = \boldsymbol{u}(t)$ is also a unique $\mathbb{H}^s_q$-solution of (3.15) in $[[0,\tau_M]]$, and the statement obviously follows. □

PROPOSITION 8. *Assume that for each $\boldsymbol{v} \in \mathbb{H}^{s+1}_p$, $\boldsymbol{G}(\boldsymbol{v},t)$ is a predictable $\mathbb{H}^{s+1}_p$-valued process and $\boldsymbol{D}(\boldsymbol{v},t)$ is a predictable $\mathbb{H}^s_p$-valued process. Let* $\mathbf{A}$, $\mathbf{A1}(s,p)$–$\mathbf{A3}(s,p)$, $\mathbf{A1}(s+1,p)$, $\mathbf{A2}(s+1,p)$ *be satisfied, $E|\boldsymbol{u}_0|^p_{s+2-2/p,p} < \infty$, and for all $t > 0, \boldsymbol{v} \in \mathbb{H}^{s+1}_p$,*

$$
\begin{aligned}
|\,|\boldsymbol{G}(\boldsymbol{v},t)|\,|_{s+1,p} \leq |\,|\boldsymbol{G}(\boldsymbol{0},t)|\,|_{s+1,p} + C|\boldsymbol{v}|_{s+1,p}, \\
|\boldsymbol{D}(\boldsymbol{v},t)|_{s,p} \leq |\boldsymbol{D}(\boldsymbol{0},t)|_{s,p} + C|\boldsymbol{v}|_{s+1,p}.
\end{aligned}
$$

Suppose also that

$$
\int_0^t (|\,|\boldsymbol{G}(\boldsymbol{0},r)|\,|^p_{s+1,p} + |\boldsymbol{D}(\boldsymbol{0},r)|^p_{s,p})\,dr < \infty
$$

$\boldsymbol{P}$*-a.s. for all* t.

Then the unique maximal $\mathbb{H}_p^s$*-solution of* (3.15) *is also a unique maximal* $\mathbb{H}_p^{s+1}$*-solution.*

Moreover, for each $T>0, M>1$, *there is a constant* C *such that for each stopping time* $\tau \le T\wedge\tau_M$,

$$E\bigg[\sup_{r\le\tau}|\boldsymbol{u}(r)|_{s+1,p}^p+\int_0^\tau|\partial^2\boldsymbol{u}(r)|_{s,p}^p\,dr\bigg]\le CE\bigg[|\boldsymbol{u}_0|_{s+2-2/p,p}^p+\int_0^\tau(|\boldsymbol{D}(\boldsymbol{0},r)|_{s,p}^p+|\,|\boldsymbol{G}(\boldsymbol{0},r)|\,|_{s+1,p}^p)\,dr\bigg],$$

where $\tau_M=\inf(t:|\boldsymbol{u}(t)|_{s,p}\ge M)$.

Proof. Since the assumptions **A**, **A1**(s,p)–**A3**(s,p) are satisfied, the existence and uniqueness of maximal $\mathbb{H}_p^s$-solution is guaranteed by Proposition 6. Let $\tau_M=\inf\{t:|\boldsymbol{u}(t)|_{s,p}\ge M\}$. Consider a linear equation

$$\begin{aligned}\partial_t\boldsymbol{\xi}(t)&=\mathcal{S}(\partial_i(a^{ij}(t)\partial_j\boldsymbol{\xi}(t))-\Psi^\varepsilon(u^k(t\wedge\tau_M))\partial_k u^l(t)+\boldsymbol{D}(\boldsymbol{u}(t),t)\\&\quad+\mathcal{S}[\sigma^k(t)\partial_k\boldsymbol{\xi}(t)+\boldsymbol{G}(\boldsymbol{u}(t),t)]\cdot\dot W,\quad\boldsymbol{\xi}(0)=\boldsymbol{u}_0.\end{aligned}$$

By Proposition 3.8 in [37], the linear equation has a unique $\mathbb{H}_p^{s+1}$-solution in $[[0,\tau_M]]$, which is also a unique $\mathbb{H}_p^s$-solution. Thus, $\boldsymbol{\xi}=\boldsymbol{u}$ $\boldsymbol{P}$-a.s. on $[[0,\tau_M]]$. Moreover, for each T, there is a constant C such that for all stopping times $\tau\le T\wedge\tau_M$,

$$E\bigg[\sup_{r\le t\wedge\tau}|\boldsymbol{u}(r)|_{s+1,p}^p+\int_0^{t\wedge\tau}|\partial^2\boldsymbol{u}(r)|_{s,p}^p\,dr\bigg]\le CE\bigg[|\boldsymbol{u}_0|_{s+2-2/p,p}^p+\int_0^{t\wedge\tau}(|\boldsymbol{u}(r)|_{s+1,p}^p+|\Psi^\varepsilon(u^k(r))\boldsymbol{u}(r)|_{s+1,p}^p+|\boldsymbol{D}(\boldsymbol{0},r)|_{s,p}^p+||\boldsymbol{G}(\boldsymbol{0},r)||_{s+1,p}^p)\,dr\bigg]$$

for all t. Since

$$|\Psi^\varepsilon(u^k(r\wedge\tau_M))\boldsymbol{u}(r)|_{s+1,p}^p\le CM|\boldsymbol{u}(r)|_{s+1,p}^p,$$

we have

$$E\bigg[\sup_{r\le t\wedge\tau}|\boldsymbol{u}(r)|_{s+1,p}^p+\int_0^{t\wedge\tau}|\partial^2\boldsymbol{u}(r)|_{s,p}^p\,dr\bigg]\le CE\bigg[|\boldsymbol{u}_0|_{s+2-2/p,p}^p+\int_0^{t\wedge\tau}(|\boldsymbol{u}(r)|_{s+1,p}^p+|\boldsymbol{D}(\boldsymbol{0},r)|_{s,p}^p+||\boldsymbol{G}(\boldsymbol{0},r)||_{s+1,p}^p)\,dr\bigg].$$

Now the estimate of the statement follows by the Gronwall lemma. □

COROLLARY 9. *Assume* **A**, **A1**$(s,2)$–**A3**$(s,2)$, $p\ge2$. *Suppose* $E|\boldsymbol{u}_0|_{s,2}^p<\infty$. *Assume further that*

$$\int_0^t(|\boldsymbol{D}(\boldsymbol{0},r)|_{s-1,2}^p+|\,|\boldsymbol{G}(\boldsymbol{0},r)|\,|_{s,2}^p)\,dr<\infty$$

$\boldsymbol{P}$*-a.s. for all* t. *Let* $(\boldsymbol{u},\zeta)$ *be the maximal* $\mathbb{H}_2^s$*-solution to* (3.15).

Then for each $T>0, M>1$, *there is a constant* C *such that for each stopping time* $\tau\le T\wedge\tau_M$,

$$\boldsymbol{E}\bigg[\sup_{r\le\tau}|\boldsymbol{u}(r)|_{s,2}^p+\int_0^\tau|\boldsymbol{u}(r)|_{s,2}^{p-2}|\nabla\boldsymbol{u}(r)|_{s,2}^2\,dr\bigg]\le CE\bigg[|\boldsymbol{u}_0|_{s,2}^p+\int_0^\tau(|\boldsymbol{D}(\boldsymbol{0},r)|_{s-1,2}^p+|\,|\boldsymbol{G}(\boldsymbol{0},r)|\,|_{s,2}^p)\,dr\bigg],$$

where $\tau_M = \inf(t : |\boldsymbol{u}(t)|_{s,p} \geq M)$.

Proof. Let $M > 0, \tau_M = \inf\{t : |\boldsymbol{u}(t)|_{s,2} \geq M\}$. We easily obtain the statement by Proposition 2 in [38] applied to the $\mathbb{H}^s_2$-solution $\boldsymbol{\xi}$ of the equation

$$\begin{aligned}\partial_t \boldsymbol{\xi}(t) &= \partial_i(a^{ij}(t)\partial_j \boldsymbol{\xi}(t)) - \Psi^\varepsilon(u^k(t\wedge\tau_M))\partial_k\boldsymbol{\xi}(t) + \boldsymbol{D}(\boldsymbol{\xi}(t),t)\\ &\quad - \mathcal{G}[\partial_i(a^{ij}(t)\partial_j\boldsymbol{\xi}(t)) - \Psi^\varepsilon(u^k(t\wedge\tau_M))\partial_k\boldsymbol{\xi}(t) + \boldsymbol{D}(\boldsymbol{\xi}(t),t)]\\ &\quad [\sigma^k(t)\partial_k\boldsymbol{\xi}(t) + \boldsymbol{G}(\boldsymbol{\xi}(t),t) - \mathcal{G}(\sigma^k(t)\partial_k\boldsymbol{\xi}(t) + \boldsymbol{G}(\boldsymbol{\xi}(t),t))]\cdot\dot{W},\\ \boldsymbol{\xi}(0) &= \boldsymbol{u}_0,\ \operatorname{div}\boldsymbol{\xi} = 0.\end{aligned}$$

According to Proposition 2 in [38],

$$\begin{aligned}E\sup_{r\leq\tau}|\boldsymbol{u}(r)|^p_{s,2} \leq CE\Big[|\boldsymbol{u}_0|^p_{s,2} + \int_0^\tau (|\boldsymbol{D}(\boldsymbol{0},r)|^p_{s-1,2}\\ + |\Psi^\varepsilon(u^k(r))\boldsymbol{u}(r)|^p_{s,2} + |\,|\boldsymbol{G}(\boldsymbol{0},r)|\,|^p_{s,2})\,dr\Big].\end{aligned}$$

Since

$$|\Psi^\varepsilon(u^k(r))\boldsymbol{u}(r)|^p_{s,2} \leq CM|\boldsymbol{u}(r)|^p_{s,2},$$

the statement follows by the Gronwall lemma. □

COROLLARY 10. *Let* $s \in \{0,1,\dots\}, q \geq 2, E|\boldsymbol{u}_0|^q_{s+1-2/q,q} < \infty$, *and* $\mathbf{A}$, $\mathbf{A1}(s,q)$–$\mathbf{A3}(s,q)$ *hold. Assume further that* $a^{ij} \in B^{s\vee 2}$ *if* $s \geq 1$, *and*

$$|\,|\boldsymbol{G}(\boldsymbol{v},t)|\,|_{s,q} \leq |\,|\boldsymbol{G}(\boldsymbol{0},t)|\,|_{s,q} + C|\boldsymbol{v}|_{s,q},\quad |\boldsymbol{D}(\boldsymbol{v},t)|_{s-1,q} \leq |\boldsymbol{D}(\boldsymbol{0},t)|_{s-1,q} + C|\boldsymbol{v}|_{s,q},$$

$$\textstyle\int_0^t (|\boldsymbol{D}(\boldsymbol{0},r)|^p_{s-1,q} + |\,|\boldsymbol{G}(\boldsymbol{0},r)|\,|^p_{s,q})\,dr < \infty$$

$\boldsymbol{P}$*-a.s. for all* t. *Let* $(\boldsymbol{u},\zeta)$ *be a maximal* $\mathbb{H}^s_q$*-solution to* (3.15).

Then for each $T > 0, M > 1$, *there is a constant* C *such that for each stopping time* $\tau \leq T \wedge \tau_M$,

$$E\sup_{r\leq\tau}|\boldsymbol{u}(r)|^p_{s,q} \leq CE\left[|\boldsymbol{u}_0|^p_{s,q} + \int_0^\tau (|\boldsymbol{D}(\boldsymbol{0},r)|^p_{s-1,q} + |\,|\boldsymbol{G}(\boldsymbol{0},r)|\,|^p_{s,q})\,dr\right],$$

where $\tau_M = \inf\{t : |\boldsymbol{u}(t)|_{s,q} \geq M\}$.

Proof. Let $M > 1, \tau_M = \inf\{t : |\boldsymbol{u}(t)|_{s,q} \geq M\}$. We easily obtain the statement by Proposition 3 in [38] applied to the $\mathbb{H}^s_q$-solution $\boldsymbol{\xi}$ of the equation

$$\begin{aligned}\partial_t \boldsymbol{\xi}(t) &= \partial_i(a^{ij}(t)\partial_j \boldsymbol{\xi}(t)) - \Psi^\varepsilon(u^k(t\wedge\tau_M))\partial_k\boldsymbol{\xi}(t) + \boldsymbol{D}(\boldsymbol{\xi}(t),t)\\ &\quad - \mathcal{G}(\partial_i(a^{ij}(t)\partial_j\boldsymbol{\xi}(t)) - \Psi^\varepsilon(u^k(t\wedge\tau_M))\partial_k\boldsymbol{\xi}^l(t) + \boldsymbol{D}(\boldsymbol{\xi}(t),t))\\ &\quad [\sigma^k(t)\partial_k\boldsymbol{\xi}(t) + \boldsymbol{G}(\boldsymbol{\xi}(t),t) - \mathcal{G}(\sigma^k(t)\partial_k\boldsymbol{\xi}(t) + \boldsymbol{G}(\boldsymbol{\xi}(t),t))]\cdot\dot{W},\\ \boldsymbol{\xi}(0) &= \boldsymbol{u}_0,\ \operatorname{div}\xi = 0.\end{aligned}$$

According to Proposition 2 in [38],

$$\begin{aligned}E\sup_{r\leq\tau\wedge t}|\boldsymbol{u}(r)|^p_{s,q} \leq C\boldsymbol{E}\Big[|\boldsymbol{u}_0|^p_{s,q} + \int_0^{t\wedge\tau} (|\boldsymbol{D}(\boldsymbol{0},r)|^p_{s-1,q}\\ + |\Psi^\varepsilon(u^k(r))\boldsymbol{u}(r)|^p_{s,q} + |\,|\boldsymbol{G}(\boldsymbol{0},r)|\,|^p_{s,q})\,dr\Big].\end{aligned}$$

Also,

$$|\Psi^{\varepsilon}(u^k(r))\boldsymbol{u}(r)|^p_{s,q} \le CM|\boldsymbol{u}(r)|^p_{s,q},$$

and the statement follows by the Gronwall lemma. □

Now we point out a simple case when $\zeta = \infty$ $\boldsymbol{P}$-a.s.

PROPOSITION 11. *Assume* **A**, **A1**$(0,2)$–**A3**$(0,2)$ *are satisfied and* $\boldsymbol{E}|\boldsymbol{u}_0|_2^2 < \infty$. *Let* $(\boldsymbol{u}(t),\zeta)$ *be a maximal* $\mathbb{H}^0_2 = \mathbb{L}_2$*-solution to* (3.15).

Then the stopping time $\zeta = \infty$ $\boldsymbol{P}$*-a.s. Moreover, for each* $T > 0$, *there is a constant* C *so that for all stopping times* $\tau \le T$,

$$\boldsymbol{E}\left[\sup_{r\le\tau}|\boldsymbol{u}(r)|_2^2 + \int_0^\tau |\nabla\boldsymbol{u}(r)|_2^2\,dr\right] \le CE\left[|\boldsymbol{u}_0|_2^2 + \int_0^\tau |\boldsymbol{D}(\boldsymbol{0},r)|^2_{-1,2} + ||\boldsymbol{G}(\boldsymbol{0},r)||_2^2\,dr\right].$$

Proof. Let $M > 1$, $\tau_M = \inf\{t : |\boldsymbol{u}(t)|_{s,2} \ge M), \boldsymbol{u}_M(t) = \boldsymbol{u}(t\wedge\tau_M)\}$. By the Itô formula (see [38]) we have

$$\begin{aligned}|\boldsymbol{u}_M(t)|_2^2 &= |\boldsymbol{u}(0)|_2^2 + 2\int_0^{t\wedge\tau_M}\langle\boldsymbol{u}(r),\boldsymbol{D}(\boldsymbol{u}(r),r)\rangle_1\,ds\\ &\quad - 2\int_0^{t\wedge\tau_M}\int a^{ij}(r)\partial_i u^l(r))\partial_j u^l(r)\,dx\,dr\\ &\quad + 2\int_0^{t\wedge\tau_M}\left(\int u^l(r)\tilde{b}^l(r)\,dx\right)\cdot dW_r + \int_0^{t\wedge\tau_M}\int\sum_i |b^i(r)|_Y^2\,dx\,dr,\end{aligned}$$

where $\tilde{b}^k(r) = \sigma^i(r)\partial_i u^k(r) + Q^k(\boldsymbol{u},r), b^k(r) = \sigma^i(r)\partial_i u^k(r) + Q^k(\boldsymbol{u},r) - \mathcal{G}(\sigma^i(r)\partial_i u^k(r) + Q^k(\boldsymbol{u},r))$. Therefore, for each T, there is a constant C, independent of M, so that for all stopping times $\tau \le T$,

$$\begin{aligned}E\sup_{r\le\tau}|\boldsymbol{u}_M(r)|_2^2 + \int_0^\tau |\nabla\boldsymbol{u}_M(r)|_2^2\,dr \le CE\Bigg[|\boldsymbol{u}_0|_2^2 + \int_0^\tau (|\boldsymbol{u}_M(r)|_2^2\\ + |\boldsymbol{D}(\boldsymbol{0},r)|^2_{-1,2} + ||\boldsymbol{G}(\boldsymbol{0},r)||_2^2\,dr\Bigg],\end{aligned}$$

and the statement follows. □

3.4. Approximating sequence. Given a scalar function v on $\boldsymbol{R}^d$, we define

$$\Psi_n(v)(x) = \Psi^{1/n}(v)(x) = \int v(x-y)\psi_{1/n}(y)\,dy,$$

where $\psi_\varepsilon(x) = \varepsilon^{-d}\psi(x/\varepsilon)$. Similarly, for a vector function $\boldsymbol{v}$,

$$\Psi_n(\boldsymbol{v})(x) = \Psi^{1/n}(\boldsymbol{v})(x).$$

We construct a sequence of approximations to (3.7) by solving for $\boldsymbol{u} = (u^l)_{1\le l\le d} = \boldsymbol{u}_n = \left(u_n^l\right)_{1\le l\le d}$ the equation

$$\text{(3.23)}\qquad \begin{cases}\partial_t\boldsymbol{u}(t) = \mathcal{S}[\partial_i\left(a^{ij}(t)\,\partial_j\boldsymbol{u}(t)\right) - \Psi_n(u^i(t))\partial_i\boldsymbol{u}(t)\\ +b^i(t)\partial_i\boldsymbol{u}(t) + \boldsymbol{F}(\boldsymbol{u}(t),t)] + \mathcal{S}[\sigma^i(t)\partial_i\boldsymbol{u}(t) + \boldsymbol{G}(\boldsymbol{u}(t),t)]\,\dot{W}_t,\\ \boldsymbol{u}(0) = \boldsymbol{u}_{0,n},\end{cases}$$

where $\boldsymbol{u}_{0,n} = \boldsymbol{u}_0 * \psi_{1/n}$, $\Psi_n(v)(x) = \Psi^{1/n}(v)(x) = \int v(x-y)\psi_{1/n}(y)\,dy$. Alternatively, we may write it as

$$
\begin{aligned}
&\partial_t \boldsymbol{u}\,(t) = \partial_i \left(a^{\,ij}\,(t)\,\partial_j \boldsymbol{u}\,(t)\right) - \Psi_n(u^{\,i}\,(t))\partial_i \boldsymbol{u}(t) - \boldsymbol{\nabla} P\,(t) \\
&+ b^i(t)\partial_i \boldsymbol{u}(t) + \boldsymbol{F}\,(\boldsymbol{u}\,(t)\,,t) + [\sigma^i(t)\partial_i \boldsymbol{u}\,(t) + \boldsymbol{G}\,(\boldsymbol{u}\,(t)\,,t) - \boldsymbol{\nabla}\tilde{P}\,(t)]\,\dot{W}_t, \\
&\boldsymbol{u}\,(0) = \boldsymbol{u}_{0,n}, \operatorname{div} \boldsymbol{u} = 0.
\end{aligned} \tag{3.24}
$$

PROPOSITION 12. (a) *Let* **B1**, **B2**(p) *be satisfied* $(p \geq 2)$, $\boldsymbol{E}(|\boldsymbol{u}_0|^p_{1,p}) < \infty$. *Then for each* $n > 1$ *there is a unique maximal* $\mathbb{H}^1_p$*-solution* $(\boldsymbol{u}, \boldsymbol{\zeta}) = (\boldsymbol{u}_n, \zeta_n)$ *of* (3.24).

(b) *Let* **B1**, **B2**(p) $(p > 2)$, *and* **B2**(2), **B2**$(2,p)$ *be satisfied, and* $\boldsymbol{E}(|\boldsymbol{u}_0|^p_{1,p} + |\boldsymbol{u}_0|^p_{1,2}) < \infty$. *Then for each* $n > 1$ *the unique maximal* $\mathbb{H}^1_p$*-solution* $(\boldsymbol{u}, \boldsymbol{\zeta}) = (\boldsymbol{u}_n, \zeta_n)$ *is also a unique maximal* $\mathbb{H}^1_2$*-solution of* (3.24). *Moreover,* $\zeta = \zeta_n = \infty$ $\boldsymbol{P}$*-a.s.*

Proof. Fix n. For each $p \geq 2$, the conditions **B1**, **B2**(p) imply the assumptions **A**, **A1**$(0,p)$–**A3**$(0,p)$, and **A1**$(1,p)$–**A2**$(1,p)$ with

$$\boldsymbol{D}(\boldsymbol{v}, t) = b^i \partial_i \boldsymbol{v} + \boldsymbol{F}(\boldsymbol{v}, t),$$

We apply Propositions 6 and 8 to (3.24) in order to obtain part (a) of the statement.

Part (b) follows by Corollary 7 and Propositions 6–11. □

Applying curl operator to both sides of (3.24), we obviously obtain the following statement.

REMARK 3. *Under the assumptions of Proposition* 12*, for each stopping time* S *such that* $[0, S] \subseteq [0, \zeta)$, η*=curl* $\boldsymbol{u}$ *(definition and properties of curl and crossproduct for* $d > 3$ *are given in the appendix, subsection* 5.5*) satisfies in* $[[0, S]]$ *the equation*

$$
\begin{aligned}
&\partial_t \boldsymbol{\eta}\,(t) = \partial_i \left(a^{\,ij}\,(t)\,\partial_j \boldsymbol{\eta}\,(t)\right) - \Psi_n(u^i(t))\partial_i \boldsymbol{\eta}(t) \\
&+ \boldsymbol{r}_n(\boldsymbol{u}(t)) + b^i(t)\partial_i \boldsymbol{\eta}(t) + curl\{\boldsymbol{F}\,(\boldsymbol{u}(t), t)\} + \boldsymbol{r}(\boldsymbol{u}(t), t) \\
&+ [\sigma^i(t)\partial_i \boldsymbol{\eta}\,(t) + \tilde{\boldsymbol{r}}(\boldsymbol{u}(t), t) + curl\{\boldsymbol{G}\,(\boldsymbol{u}(t), t)\}]\,\dot{W}_t, \\
&\boldsymbol{\eta}\,(0) = curl\,\boldsymbol{u}_{0,n},
\end{aligned} \tag{3.25}
$$

where

$$
\begin{aligned}
\boldsymbol{r}(\boldsymbol{v}, t) &= \partial_i(\boldsymbol{\nabla} a^{ij}(t) \times \partial_j \boldsymbol{v}) + (\boldsymbol{\nabla} b^i(t)) \times \partial_i \boldsymbol{v}, \\
\boldsymbol{r}_n(\boldsymbol{v}) &= -\boldsymbol{\nabla}\Psi_n(v^i) \times \partial_i \boldsymbol{v},\ \tilde{\boldsymbol{r}}(\boldsymbol{v}, t) = \boldsymbol{\nabla}\sigma^i(t) \times \partial_i \boldsymbol{v}, \boldsymbol{v} \in \mathbb{H}^1_p.
\end{aligned}
$$

The equation is linear in $\boldsymbol{\eta}$*. In fact, each component of* (3.25) *has a unique* L_p*-solution in* $[[0, S]]$ *of a corresponding linear equation. (It is also* L_2*-solution in* $[[0, S]]$ *in case* (b)*.)*

For $\boldsymbol{v} \in \mathbb{H}^1_p$ we set

$$
\begin{aligned}
\boldsymbol{L}_n(\boldsymbol{v}) &= \mathcal{G}[\Psi_n(v^i)\partial_i \boldsymbol{v}], \\
\tilde{\boldsymbol{G}}(\boldsymbol{v}, r) &= (\tilde{G}^l(\boldsymbol{v}, r))_{1 \leq l \leq d} = \mathcal{S}(\boldsymbol{G}(\boldsymbol{v}, r)) - \mathcal{G}(\sigma^i(r)\partial_i \boldsymbol{v}), \\
\tilde{\boldsymbol{F}}(\boldsymbol{v}, r) &= \mathcal{S}(\boldsymbol{F}(\boldsymbol{v}, r)) + \mathcal{S}(b^i(r)\partial_i \boldsymbol{v}) - \boldsymbol{\partial}_i \mathcal{G}(a^{ij}(r)\partial_j \boldsymbol{v}).
\end{aligned}
$$

Also, we define

$$\boldsymbol{H}(\boldsymbol{v}, t) = \operatorname{curl}\{\boldsymbol{F}\,(\boldsymbol{v}, t)\} + \boldsymbol{r}(\boldsymbol{v}, t), \boldsymbol{B}(\boldsymbol{v}, t) = \tilde{\boldsymbol{r}}(\boldsymbol{v}, t) + \operatorname{curl}\{\boldsymbol{G}\,(\boldsymbol{v}, t)\},$$

where

$$\boldsymbol{r}(\boldsymbol{v}, t) = \partial_i(\boldsymbol{\nabla} a^{ij}(t) \times \partial_j \boldsymbol{v}) + (\boldsymbol{\nabla} b^i(t)) \times \partial_i \boldsymbol{v}, \tilde{\boldsymbol{r}}(\boldsymbol{v}, t) = \boldsymbol{\nabla}\sigma^i(t) \times \partial_i \boldsymbol{v}.$$

Then we can rewrite (3.24) as

$$\begin{aligned}&\partial_t \boldsymbol{u}(t) = \partial_i\left(a^{ij}(t)\,\partial_j \boldsymbol{u}(t)\right) - \Psi_n(u^i(t))\partial_i \boldsymbol{u}(t)\\ &+\tilde{\boldsymbol{F}}(\boldsymbol{u}(t),t) + \boldsymbol{L}_n(\boldsymbol{u}(t) + [\sigma^i(t)\partial_i \boldsymbol{u}(t) + \tilde{\boldsymbol{G}}(\boldsymbol{u}(t),t)]\,\dot{W}_t,\\ &\boldsymbol{u}(0) = \boldsymbol{u}_{0,n}, \operatorname{div}\boldsymbol{u} = 0.\end{aligned} \tag{3.26}$$

Similarly, (3.25) can be written as

$$\begin{aligned}&\partial_t \boldsymbol{\eta}(t) = \partial_i\left(a^{ij}(t)\,\partial_j \boldsymbol{\eta}(t)\right) - \Psi_n(u^i(t))\partial_i \boldsymbol{\eta}(t)\\ &+\boldsymbol{r}_n(\boldsymbol{u}(t)) + b^i(t)\partial_i \boldsymbol{\eta}(t) + \boldsymbol{H}(\boldsymbol{u}(t),t) + [\sigma^i(t)\partial_i \boldsymbol{\eta}(t) + \boldsymbol{B}(\boldsymbol{u}(t),t)\}]\,\dot{W}_t,\\ &\boldsymbol{\eta}(0) = \operatorname{curl}\boldsymbol{u}_{0,n},\end{aligned} \tag{3.27}$$

where

$$\boldsymbol{r}_n(\boldsymbol{v}) = -\boldsymbol{\nabla}\Psi_n(v^i)\times\partial_i \boldsymbol{v}, \boldsymbol{v}\in\mathbb{H}^1_p.$$

For the estimate of L_p-norm of $\boldsymbol{u}$, we will need some simple estimates of $\tilde{\boldsymbol{F}}, \tilde{\boldsymbol{G}}, \boldsymbol{H}, \boldsymbol{B}$.

LEMMA 13. *Assume* $\mathbf{B2}(p)$ *holds. Then*

(a) *there is a constant* C *so that for all* $\boldsymbol{v}\in\mathbb{H}^1_p, t$,

$$\begin{aligned}|\tilde{\boldsymbol{F}}(\boldsymbol{v},t)|_{-1,p} &\le C(|\boldsymbol{F}(\boldsymbol{0},t)|_p + |\boldsymbol{v}|_p),\\ ||\tilde{\boldsymbol{G}}(\boldsymbol{v},t)||_p &\le C(||G(\boldsymbol{0},t)||_p + |\boldsymbol{v}|_p),\\ |\boldsymbol{H}(\boldsymbol{v},t)|_{-1,p} &\le C(|\boldsymbol{F}(\boldsymbol{0},t)|_p + |\boldsymbol{v}|_p + |\nabla\boldsymbol{v}|_p),\\ ||\boldsymbol{B}(\boldsymbol{v},t)||_p &\le C(||\boldsymbol{G}(\boldsymbol{0},t)||_{1,p} + |\boldsymbol{v}|_p + |\nabla\boldsymbol{v}|_p);\end{aligned}$$

(b) *there is a constant* C *so that for all* $\boldsymbol{v},\bar{\boldsymbol{v}}\in\mathbb{H}^1_p, t\ge 0$,

$$\begin{aligned}|\tilde{\boldsymbol{F}}(\boldsymbol{v},t) - \tilde{\boldsymbol{F}}(\bar{\boldsymbol{v}},t)|_{-1,p} &\le C|\boldsymbol{v}|_p,\\ ||\tilde{\boldsymbol{G}}(\boldsymbol{v},t) - \tilde{\boldsymbol{G}}(\bar{\boldsymbol{v}},t)||_p &\le C|\boldsymbol{v}|_p),\\ |\boldsymbol{H}(\boldsymbol{v},t) - \boldsymbol{H}(\bar{\boldsymbol{v}},t)|_{-1,p} &\le C(|\boldsymbol{v}-\bar{\boldsymbol{v}}|_p + |\nabla\boldsymbol{v} - \boldsymbol{\nabla}\bar{\boldsymbol{v}}|_p),\\ ||\boldsymbol{B}(\boldsymbol{v},t) - \boldsymbol{B}(\bar{\boldsymbol{v}},t)||_p &\le C(|\boldsymbol{v}-\bar{\boldsymbol{v}}|_p + |\nabla\boldsymbol{v} - \boldsymbol{\nabla}\bar{\boldsymbol{v}}|_p).\end{aligned}$$

Proof. By our assumption and Lemma 1, there is a constant C so that

$$\begin{aligned}|\mathcal{S}(\boldsymbol{F}(\boldsymbol{v},t))|_p &\le |\boldsymbol{F}(\boldsymbol{0},t)|_p + C|\boldsymbol{v}|_p),\\ ||\mathcal{S}(\boldsymbol{G}(\boldsymbol{v},t))||_p &\le ||\boldsymbol{G}(\boldsymbol{0},t)||_p + C|\boldsymbol{v}|_p).\end{aligned}$$

By Corollary 40 and Lemma 39 (see the appendix), there is a constant C so that

$$|\boldsymbol{\partial}_i\mathcal{G}(a^{ij}(r)\partial_j\boldsymbol{v})|_{-1,p}\le C|\boldsymbol{v}|_p,\ |\mathcal{G}(\sigma^i(r)\partial_i\boldsymbol{v})|_p\le C|\boldsymbol{v}|_p.$$

Also,

$$|\mathcal{S}(b^i(r)\partial_i\boldsymbol{v})|_{-1,p} = |\partial_i\mathcal{S}(b^i(r)\boldsymbol{v}) - \mathcal{S}(\partial_i b^i(r)\boldsymbol{v})|_{-1,p}\le C|\boldsymbol{v}|_p,$$

and the statement obviously follows. ∎

The following standard estimate will be needed later as well.

LEMMA 14. *Let* $p\ge 2$.

(a) *There is a constant* C *such that for all* $\boldsymbol{v}\in H^1_p(\boldsymbol{R}^d,\boldsymbol{R}^m)$,

$$|\bar{\boldsymbol{v}}|_{1,p'}\le C\left[\left(\int|\boldsymbol{v}|^{p-2}|\nabla\boldsymbol{v}|^2\,dx\right)^{1/2}|\boldsymbol{v}|_p^{(p-2)/2} + |\boldsymbol{v}|_p^{p-1}\right]$$

where $\bar{\boldsymbol{v}} = |\boldsymbol{v}|^{p-2}\boldsymbol{v}, (p')^{-1} + p^{-1} = 1$.

For each $\varepsilon > 0$ *there is a constant* C_ε *such that for all* $\boldsymbol{v} \in H^1_p(\boldsymbol{R}^d, \boldsymbol{R}^m)$,

$$|\bar{\boldsymbol{v}}|_{1,p'}|\boldsymbol{v}|_p \le \varepsilon \left(\int |\boldsymbol{v}|^{p-2}|\nabla \boldsymbol{v}|^2 \, dx \right) + C_\varepsilon |\boldsymbol{v}|_p^p.$$

(b) *For each* $\varepsilon > 0$ *there is a constant* C_ε *such that for all* $\boldsymbol{v} \in H^1_p(\boldsymbol{R}^d, \boldsymbol{R}^m), \boldsymbol{h} \in L_p(\boldsymbol{R}^d, \boldsymbol{R}^m)$,

$$\int |\boldsymbol{v}|^{p-2}|\nabla \boldsymbol{v}||\boldsymbol{h}| \, dx \le \boldsymbol{\varepsilon} \left(\int |\boldsymbol{v}|^{p-2}|\nabla \boldsymbol{v}|^2 \, dx \right) + C_\varepsilon (|\boldsymbol{v}|_p^p + |\boldsymbol{h}|_p^p).$$

(c) *If for all* t, x, y

$$|\boldsymbol{\sigma}(t,x) - \boldsymbol{\sigma}(t,y)|_Y \le K|x-y|,$$

then there is a constant C *such that for all* $\boldsymbol{v} \in H^1_p(\boldsymbol{R}^d, \boldsymbol{R}^m), \boldsymbol{h} \in L_p(\boldsymbol{R}^d, \boldsymbol{R}^m)$,

$$\left| \int (|\boldsymbol{v}|^{p-2}\boldsymbol{v}, \sigma^k \partial_k \boldsymbol{v} + \boldsymbol{h}) \, dx \right| \le C(|\boldsymbol{v}|_p^p + |\boldsymbol{v}|_p^{p-1}|\boldsymbol{h}|_p).$$

Proof. We have

$$|\bar{\boldsymbol{v}}|_{1,p'} \le C \left(|\bar{\boldsymbol{v}}|_{p'} + \sum_k |\partial_k \bar{\boldsymbol{v}}|_{p'} \right),$$

and, obviously, $|\bar{\boldsymbol{v}}|_{p'} = |\boldsymbol{v}|_p^{p-1}$. By the Hölder inequality,

$$\begin{aligned}
|\partial_k \bar{\boldsymbol{v}}|_{p'} &\le C| \, |\boldsymbol{v}|^{p-2}|\nabla \boldsymbol{v}| \, |_{p'} = C \left(\int |\boldsymbol{v}|^{p'(p-2)}|\nabla \boldsymbol{v}|^{p'} \right)^{1/p'} \\
&= C \left(\int |\boldsymbol{v}|^{p'(p-2)/2}(|\nabla \boldsymbol{v}|^{p'}|\boldsymbol{v}|^{p'(p-2)/2}) \right)^{1/p'} \\
&\le C \left(\int (|\nabla \boldsymbol{v}|^{p'}|\boldsymbol{v}|^{p'(p-2)/2})^{2/p'} \right)^{1/2} \left(\int (|\boldsymbol{v}|^{p'(p-2)/2})^{2/(2-p')} \right)^{(2-p')/2} \\
&= C \left(\int |\boldsymbol{v}|^{p-2}|\nabla \boldsymbol{v}|^2 \, dx \right)^{1/2} |\boldsymbol{v}|_p^{(p-2)/2}.
\end{aligned}$$

Therefore

$$|\bar{\boldsymbol{v}}|_{1,p'}|\boldsymbol{v}|_p \le C \left(\int |\boldsymbol{v}|^{p-2}|\nabla \boldsymbol{v}|^2 \, dx \right)^{1/2} |\boldsymbol{v}|_p^{p/2} + |\boldsymbol{v}|_p^p),$$

and part (a) follows.

For each $\varepsilon > 0$ there is a constant C_ε such that

$$\int |\boldsymbol{v}|^{p-2}|\nabla \boldsymbol{v}||\boldsymbol{h}| \, dx \le \boldsymbol{\varepsilon} \left(\int |\boldsymbol{v}|^{p-2}|\nabla \boldsymbol{v}|^2 \, dx \right) + C_\varepsilon \int |\boldsymbol{v}|^{p-2}|\boldsymbol{h}|^2 \, dx,$$

and part (b) follows by the Hölder inequality.

Integrating by parts, we easily obtain (c). □

For the estimates of L_p-norms, we will need the following important quantity. For $\boldsymbol{v} \in H_p^1(\boldsymbol{R}^d, \boldsymbol{R}^m)$, we define

$$
\begin{aligned}
N_p(\boldsymbol{v},t) &= -\int \{[|\boldsymbol{v}|^{p-2} v^l \partial_i (a^{ij}(t)\partial_j v^l) + 2^{-1}[(p-2)|\boldsymbol{v}|^{p-4} v^i(s) v^j(s) \\
&\quad + |\boldsymbol{v}|^{p-2}\delta_{ij}]\sigma^k(t)\cdot\sigma^m(t)\partial_k v^i \partial_m v^j]\}\,dx \qquad (3.28)\\
&= \int |\boldsymbol{v}|^{p-2}\partial_i v^l A^{ij}(t)\partial_j v^l\,dx \\
&\quad + (p-2)\int |\boldsymbol{v}|^{p-4}\, v^m \partial_i v^m A^{ij}(t) v^l \partial_j v^l\,dx,
\end{aligned}
$$

where

$$
A^{ij}(t) = a^{ij}(t) - \frac{1}{2}\sigma^i(t)\cdot\sigma^j(t).
$$

Notice that

$$
(p-2)\int |\boldsymbol{v}|^{p-4}\, v^m \partial_i v^m A^{ij} v^l \partial_j v^l\,dx = [4(p-2)/p^2] a^{ij}\partial_i(|\boldsymbol{v}|^{p/2})\partial_j(|\boldsymbol{v}|^{p/2}).
$$

3.5. Estimates of approximations.

3.5.1. Estimate of $\mathbb{L}_p$-norm of $\boldsymbol{u}$. For estimating $\mathbb{L}_p$-norms of the approximations, we need some auxiliary statements. We start with some interpolation inequalities.

LEMMA 15 (see [12]). (a) *Given* $\boldsymbol{v} \in \mathbb{H}_p^1, p > d > 2$,

$$
\int |\boldsymbol{v}|^{p+2}\,dx \le C|\boldsymbol{v}|_p^{p+2-d} H(\boldsymbol{v})^{d/p},
$$

where $H(\boldsymbol{v}) = |\boldsymbol{\nabla}(|\boldsymbol{v}|^{p/2})|_2^2$;

(b) *Given* $\boldsymbol{v} \in \mathbb{H}_p^1, p > d = 2$,

$$
\left(\int |\boldsymbol{v}|^{2p}\,dx\right)^{2/p} \le 2^{2/p}|\boldsymbol{v}|_p^2 H(\boldsymbol{v})^{2/p}.
$$

Proof. (a) is proved in [12]: one applies the inequality

$$
|\phi|_{2(p+2)/p} \le c|\phi|_2^{1-d/(p+2)}|\boldsymbol{\nabla}\phi|_2^{d/(p+2)}
$$

for the scalar function $\phi = |\boldsymbol{v}|^{p/2}$.

In case (b), we apply the inequality

$$
\int \phi^4\,dx \le 2\int \phi^2\,dx \int |\boldsymbol{\nabla}\phi|^2\,dx
$$

for the scalar function $\phi = |\boldsymbol{v}|^{p/2}$. We have

$$
\int |\boldsymbol{v}|^{2p}\,dx \le 2\int |\boldsymbol{v}|^p\,dx \int |\boldsymbol{\nabla}(|\boldsymbol{v}|^{p/2})|^2\,dx = 2|\boldsymbol{v}|_p^p H(\boldsymbol{v}). \qquad \square
$$

Notice $H(\boldsymbol{v}) \le C \int |\boldsymbol{v}|^{p-2} |\boldsymbol{\nabla} \boldsymbol{v}|^2 \, dx \le C |\boldsymbol{v}|^p_{1,p}$. We have also the following obvious statement.

COROLLARY 16. (a) *Let $p > d > 2$. For each ε there is a constant C_ε such that for all $\boldsymbol{v} \in \mathbb{H}^1_p$,*

$$\int |\boldsymbol{v}|^{p+2} \, dx \le \varepsilon H(\boldsymbol{v}) + C_\varepsilon (|\boldsymbol{v}|^p_p)^{1+\mu},$$

where $\mu = 2/(p-d)$.

(b) *Let $p > d = 2$. For each ε there is a constant C_ε such that for all $\boldsymbol{v} \in \mathbb{H}^1_p$,*

$$|\boldsymbol{v}|^{p-2}_p \left(\int |\boldsymbol{v}|^{2p} \, dx \right)^{2/p} \le \varepsilon H(\boldsymbol{v}) + C_\varepsilon (|\boldsymbol{v}|^p_p)^{1+\mu},$$

where $\mu = 2/(p-d)$.

LEMMA 17. *For each ε there is a constant C_ε independent of n such that for all $\boldsymbol{v} \in \mathbb{H}^1_p$,*

$$\left| \int |\boldsymbol{v}|^{p-2} (\boldsymbol{v}, \boldsymbol{L}_n(\boldsymbol{v})) \, dx \right| \le \varepsilon \int |\boldsymbol{v}|^{p-2} |\boldsymbol{\nabla} \boldsymbol{v}|^2 \, dx + C_\varepsilon (|\boldsymbol{v}|^p_p)^{1+\mu}, \tag{3.29}$$

where $\mu = 2/(p-d)$.

Proof. Denote D^{il} the lth component of $\boldsymbol{D}^i = \mathcal{G}(\Psi_n(v^i)\boldsymbol{v})$. Since

$$\mathcal{G}(\Psi_n(v^i)\partial_i \boldsymbol{v}) = \partial_i \mathcal{G}(\Psi_n(v^i)\boldsymbol{v}),$$

integrating by parts, we get that for each ε there is a constant C_ε independent of n such that

$$\begin{aligned} \left| \int |\boldsymbol{v}|^{p-2} (\boldsymbol{v}, \boldsymbol{L}_n(\boldsymbol{v})) \, dx \right| &= \left| \int \partial_i (|\boldsymbol{v}|^{p-2} v^l) D^{il} \, dx \right| \le C \int |\boldsymbol{v}|^{p-2} |\nabla \boldsymbol{u}| |\boldsymbol{D}^i| \, dx \\ &\le \varepsilon \int |\boldsymbol{v}|^{p-2} |\nabla \boldsymbol{v}|^2 \, dx + C_\varepsilon \int |\boldsymbol{v}|^{p-2} |\boldsymbol{D}^i|^2 \, dx. \end{aligned}$$

We need to estimate the term $B = \int |\boldsymbol{v}|^{p-2} |\boldsymbol{D}^i|^2 \, dx = \int |\boldsymbol{v}|^{p-2} |\boldsymbol{D}^i|^2$. By the Hölder inequality and Lemma 1,

$$\begin{aligned} B &\le \left(\int |\boldsymbol{v}|^{p+2} \right)^{\frac{p-2}{p+2}} \left(\int |\boldsymbol{D}^i|^{\frac{p+2}{2}} \right)^{\frac{4}{p+2}} \le \left(\int |\boldsymbol{v}|^{p+2} \right)^{\frac{p-2}{p+2}} \left(\int \{ |\Psi_n(v^i)| |\boldsymbol{v}| \}^{\frac{p+2}{2}} \right)^{\frac{4}{p+2}} \\ &\le \left(\int |\boldsymbol{v}|^{p+2} \right)^{\frac{p-2}{p+2}} \left(\int |\Psi_\kappa(v^i)|^{p+2} \right)^{\frac{2}{p+2}} \left(\int |\boldsymbol{v}|^{p+2} \right)^{\frac{2}{p+2}} \le \int |\boldsymbol{v}|^{p+2}. \end{aligned} \tag{3.30}$$

On the other hand,

$$\begin{aligned} B &\le \left(\int |\boldsymbol{v}|^p \right)^{\frac{p-2}{p}} \left(\int |\boldsymbol{D}^i|^p \right)^{\frac{2}{p}} \le C |\boldsymbol{v}|^{p-2}_p \left(\int \{ |\Psi_\kappa(v^i)| |\boldsymbol{v}| \}^p \right)^{\frac{2}{p}} \\ &\le C |\boldsymbol{v}|^{p-2}_p \left(\int |\Psi_n(v^i)|^{2p} \right)^{\frac{1}{p}} \left(\int |\boldsymbol{v}|^{2p} \right)^{\frac{1}{p}} \le C |\boldsymbol{v}|^{p-2}_p \left(\int |\boldsymbol{v}|^{2p} \right)^{\frac{2}{p}}. \end{aligned} \tag{3.31}$$

By Corollary 16 (using (3.30) for $d > 2$, and (3.31) for $d = 2$), for each ε there is a constant C_ε such that

$$B \le \varepsilon H(\boldsymbol{v}) + C_\varepsilon(|\boldsymbol{v}|_p^p)^{1+\mu},$$

where $\mu = 2/(p-d)$ and

$$H(\boldsymbol{v}) = |\nabla(|\boldsymbol{v}|^{p/2})|_2^2 \le C\int |\boldsymbol{v}|^{p-2}|\boldsymbol{\nabla v}|^2\,dx \le C(|\boldsymbol{v}|_p^p + |\nabla \boldsymbol{v}|_p^p. \qquad \square$$

Using the Itô formula, we estimate the $\mathbb{L}_p$-norm of the solution.

PROPOSITION 18. (a) *Let* **B1**, **B2**(p) *be satisfied,* $p > d$, $E|\boldsymbol{u}_0|_{1,p}^p < \infty$. *Then for some* $\mathbb{F}$*-adapted functions* $a(s), b(s)$ *(in which* $a(s)$ *is real valued and* $b(s)$ *is* Y*-valued)* $\boldsymbol{P}$*-a.s. in* $[0, \zeta_n)$,

$$|\boldsymbol{u}(t)|_p^p = |\boldsymbol{u}_{0,n}|_p^p + \int_0^t a(r)\,ds + \int_0^t \gamma(r)\cdot \dot{W}_s ds. \tag{3.32}$$

Moreover, there is a constant C *independent of* n *such that*

$$\begin{aligned} a(r) &\le C[|\boldsymbol{u}(r)|_p^p + (|\boldsymbol{u}(r)|_p^p)^{1+\mu} + |\boldsymbol{G}(\boldsymbol{0}, r)|_p^p + |\boldsymbol{F}(\boldsymbol{0}, r)|_p^p], \\ |\gamma(r)|_Y &\le C[|\boldsymbol{u}(r)|_p^p + |\boldsymbol{u}(r)|_p^{p-1}|\boldsymbol{G}(\boldsymbol{0}, r)|_p), \end{aligned} \tag{3.33}$$

where $\mu = 2/(p-d)$.

(b) *Let* **B1**, **B2**(p), *and* **B2**(2) *be satisfied,* $p > d$, *and*

$$E(|\boldsymbol{u}_0|_{1,p}^p + |\boldsymbol{u}_0|_{1,2}^p) < \infty,$$

$$\int_0^t (||\boldsymbol{G}(\boldsymbol{0}, r)||_{1,2}^p + |\boldsymbol{F}(\boldsymbol{0}, r)|_{1,2}^p)\,dr < \infty$$

$\boldsymbol{P}$*-a.s. for all* t. *Then* (3.32), (3.33) *hold and for some* $\mathbb{F}$*-adapted functions* $\tilde{a}(s), \tilde{\gamma}(s)$ *and all* t,

$$|\boldsymbol{u}(t)|_2^p = |\boldsymbol{u}_{0,n}|_2^p + \int_0^t \tilde{a}(r)\,dr + \int_0^t \tilde{\gamma}(r)\cdot dW_r$$

$\boldsymbol{P}$*-a.s. Moreover, there is a constant* C *independent of* n *such that*

$$\begin{aligned} \tilde{a}(r) &\le C[|\boldsymbol{u}(r)|_2^p + |\boldsymbol{G}(\boldsymbol{0}, r)|_2^p + |\boldsymbol{F}(\boldsymbol{0}, r)|_2^p], \\ |\tilde{\gamma}(r)|_Y &\le C[|\boldsymbol{u}(r)|_2^p + |\boldsymbol{u}(r)|_2^{p-1}|\boldsymbol{G}(\boldsymbol{0}, r)|_2). \end{aligned}$$

Proof. According to Proposition 12, there is a solution $\boldsymbol{u} = \boldsymbol{u}_n$ to (3.24) such that $\boldsymbol{P}$-a.s. for all T

$$\sup_{t\le T} |\boldsymbol{u}(t)|_{1,p}^p + \int_0^T |\partial^2 \boldsymbol{u}(t)|_p^p\,dt < \infty.$$

Denoting $\boldsymbol{c}(r) = (c^i(r))_{1\le i\le d} = \sigma^k \partial_k \boldsymbol{u}(r) + \tilde{\boldsymbol{G}}(\boldsymbol{u}(r), r)$, and applying the Itô formula to $\boldsymbol{u}$ satisfying (3.26) (see [38]), we find that

$$\begin{aligned} |\boldsymbol{u}(t)|_p^p = |\boldsymbol{u}_0|_p^p &- p\int_0^t N_p(\boldsymbol{u}(r), r)\,dr + p\int_0^t \int |\boldsymbol{u}(r)|^{p-2}(\boldsymbol{u}(r), \boldsymbol{L}_n(\boldsymbol{u}(r)))\,dx\,dr \\ &+ p\int_0^t \int |\boldsymbol{u}(r)|^{p-2}(\boldsymbol{u}(r), \tilde{\boldsymbol{F}}(\boldsymbol{u}(r), r))\,dx + \int_0^t p\int |\boldsymbol{u}(r)|^{p-2}u^l(r)c^l(r)\,dx\dot{W}\,dr \\ &+ \frac{p}{2}\int_0^t \left(\int [(p-2)|\boldsymbol{u}(r)|^{p-4}u^i(r)u^j(r) + |\boldsymbol{u}(r)|^{p-2}\delta_{ij}]\bar{b}^{ij}(r)\,dx\right) dr, \end{aligned}$$

where

$$\bar{b}^{ij}(r) = \sigma^k(r)\partial_k u^i(r) \cdot d^j(r) + \sigma^k(r)\partial_k u^j(r) \cdot d^i(r) + d^i(r) \cdot d^j(r),$$

and $d^i(r) = \tilde{G}^i(\boldsymbol{u}(r), r)$. By Lemmas 13 and 14, for each $\varepsilon > 0$, there is a constant C_ε such that

(3.34)
$$\left| \int |\boldsymbol{u}(r)|^{p-2} (\boldsymbol{u}(r), \tilde{\boldsymbol{F}}(\boldsymbol{u}(r), r))\, dx \right| \le \varepsilon \int |\boldsymbol{u}(r)|^{p-2} |\nabla \boldsymbol{u}(r)|^2\, dx + C_\varepsilon (|\boldsymbol{u}(r)|_p^p + |\boldsymbol{F}(\boldsymbol{0}, r)|_p^p),$$

(3.35)
$$\left| \int [(p-2)|\boldsymbol{u}(r)|^{p-4} u^i(r) u^j(r) + |\boldsymbol{u}(r)|^{p-2} \delta_{ij}] \bar{b}^{ij}(r)\, dx \right| \le \varepsilon \int |\boldsymbol{u}(r)|^{p-2} |\nabla \boldsymbol{u}(r)|^2\, dx + C_\varepsilon (|\boldsymbol{u}(r)|_p^p + |\boldsymbol{G}(\boldsymbol{0}, r)|_p^p).$$

By Lemma 14

(3.36)
$$\left| \int |\boldsymbol{u}(r)|^{p-2} u^l(r) c^l(r)\, dx \right| \le C(|\boldsymbol{u}(r)|_p^p + |\boldsymbol{u}(r)|_p^{p-1} |\boldsymbol{G}(\boldsymbol{0}, r)|_p).$$

So, (3.33) follows by Lemma 17.

In the case (b), applying the Itô formula (see [38]), we obtain

$$\begin{aligned} |\boldsymbol{u}(t)|_2^p &= |\boldsymbol{u}(0)|_2^p - p \int_0^t |\boldsymbol{u}(r)|_2^{p-2} N_2(\boldsymbol{u}(r), r)\, dr \\ &+ p \int_0^t |\boldsymbol{u}(r)|_2^{p-2} \int (\boldsymbol{u}(r), \tilde{\boldsymbol{F}}(\boldsymbol{u}(r), r))\, dx\, dr + p \int_0^t |\boldsymbol{u}(r)|_2^{p-2} \left(\int u^l(r) c^l(r)\, dx \right) dW_r \\ &+ p/2 \int_0^t |\boldsymbol{u}(r)|_2^{p-2} \left(\int \bar{b}^{ii}(r)\, dx \right) dr + \frac{p}{2}(p-2) \int_0^t |\boldsymbol{u}(r)|_2^{p-4} \left| \int u^l(r) c^l(r)\, dx \right|_Y^2 dr. \end{aligned}$$

Since (3.34)–(3.36) holds for $p = 2$ as well, the assertion of part (b) follows. □

REMARK 4. *There is a constant $C = C(K, d, p)$ independent of δ such that*

$$\begin{aligned} a(s) &\le C[|\boldsymbol{u}(s)|_p^p + |\nabla \boldsymbol{u}(r)|_p^p + (|\boldsymbol{u}(s)|_p^p)^{1+\mu} + |\boldsymbol{G}(\boldsymbol{0}, s)|_p^p + |\boldsymbol{F}(\boldsymbol{0}, s)|_p^p], \\ |\gamma(s)|_Y &\le C[|\boldsymbol{u}(s)|_p^p + |\boldsymbol{u}(s)|_p^{p-1} |\boldsymbol{G}(\boldsymbol{0}, s)|_p). \end{aligned}$$

3.5.2. Estimate of $\mathbb{L}_p$-norm of $\nabla \boldsymbol{u}$. Since by the Biot–Savaret law (see Proposition 42 in the appendix), for each $p > 1$,

$$|\boldsymbol{\nabla u}|_p \le C|\boldsymbol{\eta}|_p, (\boldsymbol{\eta} = \operatorname{curl} \boldsymbol{u}),$$

we need to estimate $|\boldsymbol{\eta}|_p$. According to Remark 3, $\boldsymbol{\eta}$ satisfies linear equation (3.25) or (3.27).

PROPOSITION 19. (a) *Let* **B1**, **B2**(p) *be satisfied* $(p > d)$, $E(|\boldsymbol{u}_0|_{1,p}^p) < \infty$, *and let* $\boldsymbol{u}$ *be the solution of* (3.26). *Then for some* $\mathbb{F}$*-adapted functions* $h(t), \kappa(t)$ *($h(t)$ is real valued and $\kappa(t)$ is Y-valued)* $\boldsymbol{P}$*-a.s. in* $[0, \zeta_n)$,

$$|\boldsymbol{\eta}(t)|_p^p = |curl\, \boldsymbol{u}_{0,n}|_p^p + \int_0^t h(r)\ ds + \int_0^t \kappa(r) \cdot dW_s.$$

Moreover, there is a constant C independent of n such that

$$h(r) \le C[|\boldsymbol{\eta}(r)|_p^p + (|\boldsymbol{\eta}(r)|_p^p)^{1+\mu} + |\boldsymbol{u}(r)|_p^p + |\boldsymbol{F}(\mathbf{0},r)|_{1,p}^p + ||\boldsymbol{G}(\mathbf{0},r)||_{1,p}^p],$$
$$|\kappa(r)|_Y \le C[|\boldsymbol{\eta}(r)|_p^p + |\boldsymbol{u}(r)|_p^p + |\boldsymbol{\eta}(r)|_p^{p-1}|\boldsymbol{G}(\mathbf{0},r)|_{1,p}),$$

where $\mu = 2/(p-d)$.

(b) *Let* **B1**, **B2**(p), *and* **B2**(2) *be satisfied* $(p > d = 2)$, *and*

$$E(|\boldsymbol{u}_0|_{1,p}^p + |\boldsymbol{u}_0|_{1,2}^p) < \infty,$$
$$\int_0^t (||\boldsymbol{G}(\mathbf{0},r)||_{1,2}^p + |\boldsymbol{F}(\mathbf{0},r)|_{1,2}^p)\,dr < \infty$$

$\boldsymbol{P}$*-a.s. for all* t. *Let* $\boldsymbol{u}$ *be the solution of* (3.26). *Then for some* $\mathbb{F}$*-adapted functions* $\tilde{a}(s), \tilde{b}(s)$ *(*$\tilde{a}(s)$ *is real valued and* $\tilde{b}(s)$ *is* Y*-valued)* $\boldsymbol{P}$*-a.s. for all* t,

$$|\boldsymbol{\eta}(t)|_2^p = |curl\,\boldsymbol{u}_{0,n}|_2^p + \int_0^t \tilde{a}(r)\ ds + \int_0^t \tilde{b}(r)\cdot dW_s.$$

Moreover, there is a constant C independent of n such that

$$\tilde{a}(r) \le C[|\boldsymbol{\eta}(r)|_2^p + (|\boldsymbol{\eta}(r)|_2^p)^{1+2/p} + |\boldsymbol{u}(r)|_2^p + |\boldsymbol{F}(\mathbf{0},r)|_{1,2}^p + ||\boldsymbol{G}(\mathbf{0},r)||_{1,2}^p],$$
$$|\tilde{b}(r)|_Y \le C[|\boldsymbol{\eta}(r)|_2^p + |\boldsymbol{\eta}(r)|_2^{p-1}(|\boldsymbol{u}(r)|_2 + ||\boldsymbol{G}(\mathbf{0},r)||_{1,2})].$$

Proof. By applying the Itô formula to function $\boldsymbol{\eta}(t)$, which verifies (3.27), we find that

$$\begin{aligned}|\boldsymbol{\eta}(t)|_p^p = |\boldsymbol{\eta}(0)|_p^p &- p\int_0^t N_p(\boldsymbol{\eta}(r),r)\,dr - p\int_0^t \int |\boldsymbol{\eta}(r)|^{p-2}(\boldsymbol{\eta}(r),\boldsymbol{r}_n(\boldsymbol{u}(r)))\,dx\,dr\\ &+ p\int_0^t \left\langle |\boldsymbol{\eta}(r)|^{p-2}\boldsymbol{\eta}(r),\boldsymbol{H}(r)\right\rangle_{1,p} dr + \int_0^t p\int |\boldsymbol{\eta}(r)|^{p-2}\eta^l(r)c^l(r)\,dx\dot{W}_r dr\\ &+ \frac{p}{2}\int_0^t \left(\int [(p-2)|\boldsymbol{\eta}(r)|^{p-4}\eta^i(r)\eta^j(r) + |\boldsymbol{\eta}(r)|^{p-2}\delta_{ij}]\bar{b}^{ij}(r)\,dx\right) dr,\end{aligned}$$

where $\boldsymbol{c}(r) = (c^i(r))_i = \sigma^k(r)\partial_k\boldsymbol{\eta}(r) + \boldsymbol{B}(\boldsymbol{u}(r),r)$,

$$\boldsymbol{H}(r) = (H^l(r))_l = b^i(r)\partial_i\boldsymbol{\eta}(r)) + \boldsymbol{H}(\boldsymbol{u}(r),r),$$

and

$$\bar{b}^{ij}(r) = \sigma^k\partial_k\eta^i d^j(r) + \sigma^k\partial_k\eta^j d^i(r) + d^i(r)d^j(r),$$

and $\boldsymbol{d}(r) = \boldsymbol{B}(\boldsymbol{u}(r),r)$.

According to Lemmas 13 and 14, for every $\varepsilon > 0$, there is a constant C_ε so that

$$\begin{aligned}&\left|\int [(p-2)|\boldsymbol{\eta}(r)|^{p-4}\eta^i(r)\eta^j(r) + |\boldsymbol{\eta}(r)|^{p-2}\delta_{ij}]\bar{b}^{ij}(r)\,dx\right|\\ &\qquad \le \varepsilon\int |\boldsymbol{\eta}|^{p-2}|\nabla\boldsymbol{\eta}|^2\,dx + C_\varepsilon(|\boldsymbol{G}(\mathbf{0},r)|_{1,p}^p + |\boldsymbol{u}(r)|_{1,p}^p),\end{aligned} \tag{3.37}$$

and

$$|\left\langle |\boldsymbol{\eta}(r)|^{p-2}\boldsymbol{\eta}(r),\boldsymbol{H}(r)\right\rangle_{1,p}| \le \varepsilon\int |\boldsymbol{\eta}|^{p-2}|\nabla\boldsymbol{\eta}|^2\,dx + C_\varepsilon(|\boldsymbol{F}(\mathbf{0},r)|_{1,p}^p + |\boldsymbol{u}(r)|_{1,p}^p). \tag{3.38}$$

Also,

$$\left|\int |\boldsymbol{\eta}(r)|^{p-2}\eta^l(r)c^l(r)\,dx\right|_Y \le C(|\boldsymbol{\eta}(r)|_p^p + |\boldsymbol{u}(r)|_{1,p}^p + |\boldsymbol{G}(\mathbf{0},r)|_{1,p}|\boldsymbol{\eta}(r)|_p^{p-1}). \tag{3.39}$$

It remains to estimate the term

$$\begin{aligned} A &= \int |\boldsymbol{\eta}(r)|^{p-2}\boldsymbol{\eta}(r), \boldsymbol{r}_n(\boldsymbol{u}(r))\,dx \\ &= \int |\boldsymbol{\eta}(r)|^{p-2}(\boldsymbol{\eta}(r), \boldsymbol{\nabla}(\Psi_n(u^i(r))) \times \partial_i \boldsymbol{u}(r))\,dx\,. \end{aligned}$$

We have

$$|A| \le C\left(\int |\boldsymbol{\eta}|^p + \int |\boldsymbol{\eta}|^{p-2}|\boldsymbol{\nabla}\Psi_\kappa(\boldsymbol{u})|^2|\boldsymbol{\nabla}\boldsymbol{u}|^2\right),$$

and by the Hölder inequality

$$\int |\boldsymbol{\eta}|^{p-2}|\boldsymbol{\nabla}\Psi_\kappa(\boldsymbol{u})|^2|\boldsymbol{\nabla}\boldsymbol{u}|^2 \le |\boldsymbol{\eta}|_{p+2}^{p-2}|\boldsymbol{\nabla}\Psi_n(\boldsymbol{u})|_{p+2}^2|\boldsymbol{\nabla}\boldsymbol{u}|_{p+2}^2 \le \int |\boldsymbol{\eta}|^{p+2}$$

or

$$\int |\boldsymbol{\eta}|^{p-2}|\boldsymbol{\nabla}\Psi_\kappa(\boldsymbol{u})|^2|\boldsymbol{\nabla}\boldsymbol{u}|^2 \le C|\boldsymbol{\eta}|_p^{p-2}\left(\int |\boldsymbol{\eta}|^{2p}\right)^{2/p}.$$

So, by Corollary 16, for each ε there is a constant C_ε independent of n such that

$$|A| \le \varepsilon \int |\boldsymbol{\eta}|^{p-2}|\boldsymbol{\nabla}\boldsymbol{\eta}|^2\,dx + C_\varepsilon(|\boldsymbol{\eta}|_p^p)^{1+\mu}, \tag{3.40}$$

where $\mu = 2/(p-d)$. Part (a) of the statement obviously follows by summarizing all the estimates.

In case (b), we apply the Itô formula to $|\boldsymbol{\eta}(t)|_2^p$:

$$\begin{aligned} |\boldsymbol{\eta}(t)|_2^p = |\boldsymbol{\eta}(0)|_2^p &- p\int_0^t |\boldsymbol{\eta}(r)|_2^{p-2}N_2(\boldsymbol{\eta}(r),r)\,dr - p\int_0^t |\boldsymbol{\eta}(r)|_2^{p-2}\int(\boldsymbol{\eta}(r), \boldsymbol{r}_n(\boldsymbol{u}(r)))\,dx\,dr \\ &+ p\int_0^t |\boldsymbol{\eta}(r)|_2^{p-2}\int(\boldsymbol{\eta}(r), \boldsymbol{H}(r))\,dx\,dr + \int_0^t p|\boldsymbol{\eta}(r)|_2^{p-2}\left(\int \eta^l(r)c^l(r)\,dx\right)\,dW_r \\ &+ \frac{p}{2}\int_0^t |\boldsymbol{\eta}(r)|_2^{p-2}\left(\int \bar{b}^{ii}(r)\,dx\right)\,dr + \frac{p}{2}(p-2)\int_0^t |\boldsymbol{\eta}(r)|_2^{p-4}\left|\int \eta^l(r)c^l(r)\,dx\right|_Y^2\,dr. \end{aligned}$$

Since (3.37)–(3.39) hold for $p = 2$ as well, it remains to estimate $A = \int(\boldsymbol{\eta}(r), \boldsymbol{\nabla}(\Psi_n(\boldsymbol{u})^i) \times \partial_i \boldsymbol{u}(r))\,dx$:

$$|A| \le C|\boldsymbol{\eta}(r)|_2|\boldsymbol{\eta}(r)|_4^2 \le C|\boldsymbol{\eta}(r)|_2^2|\boldsymbol{\nabla}\boldsymbol{\eta}(r)|_2 \text{ if } d = 2.$$

So,

$$|\boldsymbol{\eta}(r)|_2^{p-2}|A| \le \varepsilon|\boldsymbol{\eta}(r)|_2^{p-2}|\boldsymbol{\nabla}\boldsymbol{\eta}(r)|_2^2 + C_\varepsilon(|\boldsymbol{\eta}(r)|_2^p)^{1+2/p}.$$

Now, part (b) follows. □

REMARK 5. (a) *Consider a scalar process* $y_t = |\boldsymbol{u}(t)|_p^p + |\boldsymbol{\eta}(t)|_p^p$. *Then, according to parts* (a) *of Propositions* 18 *and* 19, *for some adapted functions* $h(t)$ *and* $\kappa(t)$ (κ *is* Y*-valued) in* $[0, \zeta_n)$,

$$y_t = y_0 + \int_0^t h_r\, dr + \int_0^t \kappa_r \cdot \dot{W}_r dr, \tag{3.41}$$

and there is a constant $C = C(\delta, K, d, p)$ *such that*

$$\begin{aligned} h_r &\le C(y_r + y_r^{1+\mu} + z_r), \\ |\kappa_r|_Y &\le C(y_r + y_r^{1-1/p} \tilde{z}_r^{1/p}), \end{aligned} \tag{3.42}$$

where $z_r = |\boldsymbol{F}(\boldsymbol{0}, r)|_{1,p}^p + ||\boldsymbol{G}(\boldsymbol{0}, r)||_{1,p}^p$, $\tilde{z}_s = ||\boldsymbol{G}(\boldsymbol{0}, r)||_{1,p}^p$.

(b) *Consider a scalar process* $\tilde{y}_t = |\boldsymbol{u}(t)|_2^p + |\boldsymbol{\eta}(t)|_2^p$. *Then, according to parts* (b) *of Proposition* 18 *and* 19, *for some adapted functions* $\tilde{h}(t)$ *and* $\tilde{\kappa}(t)$ ($\tilde{\kappa}$ *is* Y*-valued),*

$$\tilde{y}_t = \tilde{y}_0 + \int_0^t \tilde{h}_r\, dr + \int_0^t \tilde{\kappa}_r \cdot \dot{W}_r dr,$$

and there is a constant $C = C(\delta, K, d, p)$ *such that*

$$\begin{aligned} \tilde{h}_r &\le C(\tilde{y}_r + \tilde{y}_r^{1+2/p} + \tilde{z}_r), \\ |\tilde{\kappa}_r|_Y &\le C(\tilde{y}_r + \tilde{y}_r^{1-1/p} (\tilde{z}_r')^{1/p}), \end{aligned}$$

where $\tilde{z}_r = |\boldsymbol{F}(\boldsymbol{0}, r)|_{1,2}^p + ||\boldsymbol{G}(\boldsymbol{0}, r)||_{1,2}^p$, $\tilde{z}_r' = ||\boldsymbol{G}(\boldsymbol{0}, r)||_{1,2}^p$.

We introduce the smooth scalar function

$$G(y) = \int_0^y (1 + x + x^{1+\mu})^{-1}\, dx.$$

Notice that $G'(y) = (1 + y + y^{1+\mu})^{-1} > 0$, $G''(y) \le 0$.

REMARK 6. *In the context of the previous remark, we obtain by the Itô formula*

$$G(y_t) = G(y_0) + \int_0^t \bar{r}_s ds + \int_0^t \bar{b}_s \cdot \dot{W}_s\, ds,$$

where $r_s = G'(y_s) r_s + 2^{-1} G''(y_s) |\bar{b}_s|_Y^2 \le C(1 + z_s)$, $|\bar{b}_s|_Y = G'(y_s) |b_s|_Y \le C(1 + \tilde{z}_s^{1/p})$.

A similar observation holds for $\tilde{y}_r$.

3.5.3. Convergence of approximations. The following two auxiliary statements will be needed later.

LEMMA 20. (a) *Let* $\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{f} \in \mathbb{H}_p^1$. *For each* $\varepsilon > 0$ *there is a constant* C_ε *such that*

$$\left| \int (\mathcal{S}[(\Psi_n(v^k) - \Psi_{n'}(\bar{v}^k)) \partial_k \boldsymbol{g}], \boldsymbol{f} |\boldsymbol{f}|^{p-2})\, dx \right| \le \varepsilon \int |\nabla \boldsymbol{f}|^2 |\boldsymbol{f}|^{p-2}\, dx + C_\varepsilon (|\boldsymbol{f}|_p^p + |\boldsymbol{g} A|_p^p),$$

$$\left| \int (\mathcal{S}[\Psi_{n'}(\bar{v}^k)) \partial_k \boldsymbol{g}], \boldsymbol{f} |\boldsymbol{f}|^{p-2})\, dx \right| \le \varepsilon \int |\nabla \boldsymbol{f}|^2 |\boldsymbol{f}|^{p-2}\, dx + C_\varepsilon (|\boldsymbol{f}|_p^p + |\boldsymbol{g} B|_p^p),$$

where $A = |\Psi_n(v^k) - \Psi_{n'}(\bar{v}^k)|, B = |\Psi_{n'}(\bar{v}^k)|$;

(b) *Let* $\boldsymbol{v} \in \mathbb{H}_p^1$, $\boldsymbol{f} = (f^l), \boldsymbol{g} = (g^l) \in H_p^1(\boldsymbol{R}^d, \boldsymbol{R}^{d(d-1)/2})$. *For each* $\varepsilon > 0$ *there is a constant* C_ε *such that*

$$\left| \int (\Psi_n(v^k) - \Psi_{n'}(\bar{v}^k)) \partial_k g^l f^l |\boldsymbol{f}|^{p-2}\, dx \right| \le \varepsilon \int |\nabla \boldsymbol{f}|^2 |\boldsymbol{f}|^{p-2}\, dx + C_\varepsilon (|A \boldsymbol{g}|_p^p + |\boldsymbol{f}|_p^p),$$

where $\boldsymbol{g} = (g^l), \boldsymbol{f} = (f^l), A = |\Psi_n(v^k) - \Psi_{n'}(\bar{v}^k)|$.

Proof. (a) Indeed, we have

$$\begin{aligned}&\left|\int (\mathcal{S}[(\Psi_n(v^k) - \Psi_{n'}(\bar{v}^k))\partial_k \boldsymbol{g}], \boldsymbol{f}|\boldsymbol{f}|^{p-2})\,dx\right| \\ &= \left|\int (\mathcal{S}[(\Psi_n(v^k) - \Psi_{n'}(\bar{v}^k))\boldsymbol{g}], \partial_k(f^l|\boldsymbol{f}|^{p-2}))\,dx\right| \\ &\le \varepsilon \int |\nabla \boldsymbol{f}|^2 |\boldsymbol{f}|^{p-2}\,dx + C_\varepsilon(|\boldsymbol{f}|_p^p + |\boldsymbol{g}A|_p^p).\end{aligned}$$

Similarly, the second estimate follows.

(b) We have

$$\left|\int (\Psi_n(v^k) - \Psi_{n'}(\bar{v}^k))\partial_k g^l f^l |\boldsymbol{f}|^{p-2}\,dx\right| = \left|\int (\Psi_n(v^k) - \Psi_{n'}(\bar{v}^k)) g^l \partial_k (f^l|\boldsymbol{f}|^{p-2})\,dx\right|$$

$$\le \varepsilon \int |\nabla \boldsymbol{f}|^2 |\boldsymbol{f}|^{p-2}\,dx + C_\varepsilon \int A^2 |\boldsymbol{g}|^2 |\boldsymbol{f}|^{p-2}\,dx \le \varepsilon \int |\nabla \boldsymbol{f}|^2 |\boldsymbol{f}|^{p-2}\,dx + C_\varepsilon |A\boldsymbol{g}|_p^2 |\boldsymbol{f}|_p^{p-2}$$

$$\le \varepsilon \int |\nabla \boldsymbol{f}|^2 |\boldsymbol{f}|^{p-2}\,dx + C_\varepsilon(|A\boldsymbol{g}|_p^p + |\boldsymbol{f}|_p^p),$$

and the statement follows. ∎

LEMMA 21. (a) *There is a constant* C *so that for all* $\boldsymbol{v}, \bar{\boldsymbol{v}} \in \mathbb{H}_p^1, n' \ge n > 1$,

$$|\Psi_n(v^k) - \Psi_{n'}(\bar{v}^k)|_p \le C(|\boldsymbol{v} - \bar{\boldsymbol{v}}|_{1,p} + n^{-1}(|\nabla \boldsymbol{v}|_p + |\nabla \bar{\boldsymbol{v}}|_p)].$$

(b) *Let* $p > d$. *Then there is a constant* C *so that for all* $\boldsymbol{v}, \bar{\boldsymbol{v}} \in \mathbb{H}_p^1, n' \ge n > 1$,

$$|\Psi_n(v^k) - \Psi_{n'}(\bar{v}^k)|_\infty \le C[|\boldsymbol{v} - \bar{\boldsymbol{v}}|_{1,p} + n^{-\nu}|\boldsymbol{v}|_{1,p}],$$

where $\nu = 1 - d/p$.

Proof. By Sobolev's embedding theorem there is a constant C so that for all $\boldsymbol{v} \in \mathbb{H}_p^1$

$$\sup_x |\boldsymbol{v}(x)| + \sup_{x,y} |\boldsymbol{v}(x) - \boldsymbol{v}(y)||x - y|^\nu \le C|\boldsymbol{v}|_{1,p},$$

where $\nu = 1 - d/p$. Therefore,

$$\begin{aligned}&\sup_x |\Psi_n(v^k) - \Psi_{n'}(\bar{v}^k)| \le \sup_x |\Psi_n(v^k) - \Psi_n(\bar{v}^k)| + \sup_x |\Psi_n(\bar{v}^k) - \bar{v}^k| \\ &\quad + \sup_x |\bar{v}^k - \Psi_{n'}(\bar{v}^k)| \le C[\sup_x |\boldsymbol{v}(x) - \bar{\boldsymbol{v}}(x)| + ((1/n)^\nu + (1/n')^\nu)|\bar{\boldsymbol{v}}|_{1,p}],\end{aligned}$$

and the statement follows. ∎

We will need the following equalities and estimates later.

LEMMA 22. *Let* $\boldsymbol{v}, \boldsymbol{d} \in \mathbb{H}_p^1, \eta = (\eta^{jl})_{j<l} \in H_p^1(\boldsymbol{R}^d, \boldsymbol{R}^{d(d-1)/2}), \bar{\eta} = \eta|\eta|^{p-2}, p > d$.

Then

$$\int \left(\nabla d^k \times \partial_k \boldsymbol{v}, \bar{\eta}\right)\,dx = \int \left(d^k \partial_k \boldsymbol{v} \times \boldsymbol{\nabla}, \bar{\eta}\right)\,dx + \int (\boldsymbol{\nabla} \times \boldsymbol{v}, d^k \partial_k \bar{\eta})\,dx.$$

Also, for each ε *there is a constant* C_ε *such that for all* $\boldsymbol{v}, \boldsymbol{d} \in \mathbb{H}_p^1, \eta = (\eta^{jl})_{j<l} \in H_p^1(\boldsymbol{R}^d, \boldsymbol{R}^{d(d-1)/2})$,

$$\int |\left(\nabla d^k \times \partial_k \boldsymbol{v}, \bar{\eta}\right)\,dx| \le \varepsilon \int |\eta|^{p-2} |\nabla \eta|^2\,dx + C_\varepsilon(|\eta|_p^p + |\nabla \boldsymbol{v}|_p^p |\boldsymbol{d}|_\infty^p).$$

Proof. It is enough to prove the statement for $\boldsymbol{v}, \boldsymbol{d} \in \mathbb{C}_0^\infty, \eta^{jl} \in C_0^\infty$. Integrating by parts, we have

$$\begin{aligned}\int \left(\nabla d^k \times \partial_k \boldsymbol{v}, \bar{\eta}\right) dx &= \int \varepsilon_{jk}(\partial_j d^k \partial_k v^l - \partial_l d^k \partial_k v^j)\bar{\eta}^{jl}\, dx \\ &= -\int \varepsilon_{jk} d^k(\partial_k \partial_j v^l - \partial_k \partial_l v^j)\bar{\eta}^{jl}\, dx - \int \varepsilon_{jk} d^k(\partial_k v^l \partial_j \bar{\eta}^{jl} - \partial_k v^j \partial_l\, \bar{\eta}^{jl})\, dx \\ &= \int \varepsilon_{jk}(d^k \partial_j v^l - d^k \partial_l v^j)\partial_k \bar{\eta}^{jl}\, dx + \int \varepsilon_{jk}(d^k \partial_k v^j \partial_l\, \bar{\eta}^{jl} - d^k \partial_k v^l \partial_j \bar{\eta}^{jl})\, dx,\end{aligned}$$

where $\varepsilon_{jk} = (-1)^{j+k-1}$. Therefore, for each ε there is a constant C_ε such that

$$\left|\int \left(\nabla d^k \times \partial_k \boldsymbol{v}, \bar{\eta}\right) dx\right| \le \varepsilon \int |\eta|^{p-2}|\nabla \eta|^2\, dx + C_\varepsilon \int |\eta|^{p-2}|\boldsymbol{d}|^2 |\nabla \boldsymbol{v}|^2\, dx,$$

and the statement follows by the Hölder inequality. □

REMARK 7. *If $d = 2$, then for all $\boldsymbol{v} \in \mathbb{H}_p^1$*

$$\nabla v^k \times \partial_k \boldsymbol{v} = 0.$$

Let $\boldsymbol{u} = \boldsymbol{u}_n = (u_n^l) = (u^l)$ be a maximal $\mathbb{H}_p^1$-solution to (3.24). Let $\boldsymbol{\eta} = \boldsymbol{\eta}_n = (\eta_n^l) = \operatorname{curl} \boldsymbol{u}_n$. Fix a large number $M > 0$ and $T > 0$. Given a positive integer n, let $\mathcal{T}_n = \mathcal{T}_n^{M,T}$ be the set of all stopping times $\tau \le T \wedge \zeta_n$ such that $\boldsymbol{P}$-a.s.

$$\sup_{s \le \tau}(|\boldsymbol{u}_n(s)|_p + |\boldsymbol{\eta}_n(s)|_p) \le M.$$

In the case $d = 2$, we also introduce the set $\tilde{\mathcal{T}}_n = \tilde{\mathcal{T}}_n^{M,T}$ of all stopping times $\tau \le T$ such that

$$\sup_{s \le \tau}(|\boldsymbol{u}_n(s)|_p + |\boldsymbol{\eta}_n(s)|_p + |\boldsymbol{u}_n(s)|_2 + |\boldsymbol{\eta}_n(s)|_2) \le M.$$

Let $\mathcal{T}_{n,n'} = \mathcal{T}_n \cap \mathcal{T}_{n'}, \tilde{\mathcal{T}}_{n,n'} = \tilde{\mathcal{T}}_n \cap \tilde{\mathcal{T}}_n$.

LEMMA 23. (a) *Let* **B1**, **B2**(p), **B3**(p) *be satisfied* $(p > d)$, $\boldsymbol{E}(|\boldsymbol{u}_0|_{1,p}^p) < \infty$. *Let* $\boldsymbol{u} = \boldsymbol{u}_n = (u_n^l) = (u^l)$ *be* $\mathbb{H}_p^1$*-solutions to* (3.24). *Then*

$$\lim_n \sup\{E \sup_{s \le \tau} |\boldsymbol{u}_{n'} - \boldsymbol{u}_n|_{1,p}^p :\ n' \ge n, \tau \in \mathcal{T}_{n',n}\} = 0,$$

where $\mathcal{T}_{n',n} = \mathcal{T}_n \cap \mathcal{T}_{n'}$.

(b) *Let* **B1**, **B2**(p), **B3**(p), **B2**(2), **B3**(2) *be satisfied* $(p > d = 2)$, *and*

$$E(|\boldsymbol{u}_0|_{1,p}^p + |\boldsymbol{u}_0|_{1,2}^p) < \infty,$$

$$\int_0^t (||\boldsymbol{G}(\boldsymbol{0}, r)||_{1,2}^p + |\boldsymbol{F}(\boldsymbol{0}, r)|_{1,2}^p)\, dr < \infty$$

$\boldsymbol{P}$*-a.s. for all* t. *Let* $\boldsymbol{u} = \boldsymbol{u}_n = (u_n^l) = (u^l)$ *be an* $\mathbb{H}_p^1 \cap \mathbb{H}_2^1$*-solution to* (3.24). *Then*

$$\lim_{n \to \infty} \sup\{E \sup_{s \le \tau}(|\boldsymbol{u}_{n'} - \boldsymbol{u}_n|_{1,p}^p + |\boldsymbol{u}_{n'} - \boldsymbol{u}_n|_{1,2}^p) :\ n' \ge n, \tau \in \tilde{\mathcal{T}}_{n',n}\} = 0,$$

where $\tilde{\mathcal{T}}_{n',n} = \tilde{\mathcal{T}}_n \cap \tilde{\mathcal{T}}_{n'}$.

Proof. Let $\tau \in \mathcal{T}_{n',n}, n' \geq n$. Consider

$$\boldsymbol{w}(t) = \boldsymbol{u}_{\boldsymbol{n'}}(t \wedge \tau) - \boldsymbol{u}_n(t \wedge \tau),\ \boldsymbol{\xi}(t) = \boldsymbol{\eta}_{n'}(t \wedge \tau) - \boldsymbol{\eta}_n(t \wedge \tau).$$

Denote $\bar{\boldsymbol{u}} = \boldsymbol{u}_{n'}$, $\boldsymbol{u} = \boldsymbol{u}_n$.

Applying the Itô formula (see [38]), we have

$$\begin{aligned}
|\boldsymbol{w}(t)|_p^p &= |\boldsymbol{w}(0)|_p^p - p\int_0^{t\wedge\tau} N_p(\boldsymbol{w}(r), r)\,dr + p\int_0^{t\wedge\tau}\int |\boldsymbol{w}(r)|^{p-2}(\boldsymbol{w}(r), \boldsymbol{a}(r)\,dx\,dr \\
&+ p\int_0^{t\wedge\tau}\int |\boldsymbol{w}(r)|^{p-2} w^l(r) c^l(r)\,dW_r \\
&+ \frac{p}{2}\int_0^{t\wedge\tau}\left(\int [(p-2)|\boldsymbol{w}(r)|^{p-4} w^i(r) w^j(r) + |\boldsymbol{w}(r)|^{p-2}\delta_{ij}]\bar{c}^{ij}(r)\,dx\right) dr \\
&- p\int_0^{t\wedge\tau}\left(\int |\boldsymbol{w}(r)|^{p-2}(\boldsymbol{w}(r), \mathcal{S}[\partial_k \boldsymbol{w}(r)\Psi_{n'}(\bar{u}^k) + (\Psi_{n'}(\bar{u}^k) - \Psi_n(u^k))\partial_k \boldsymbol{u}]\,dx\right) dr,
\end{aligned} \tag{3.43}$$

where $\boldsymbol{c}(r) = (c^i(r))_i = \sigma^k \partial_k \boldsymbol{w}(r) + \tilde{\boldsymbol{G}}(\bar{\boldsymbol{u}}, r) - \tilde{\boldsymbol{G}}(\boldsymbol{u}, r)$,

$$\boldsymbol{a}(r) = (a^l(r))_l = \tilde{F}^l(\boldsymbol{v}, r) - \tilde{F}^l(\boldsymbol{u}, r);$$

and

$$\bar{c}^{ij}(r) = \sigma^k(r)\partial_k w^i(r)\cdot d^j(r) + \sigma^k(r)\partial_k w^j(r)\cdot d^i(r) + d^i(r)\cdot d^j(r),$$

where $d^i(r) = \tilde{G}^i(\bar{\boldsymbol{u}}(r), r) - \tilde{G}^i(\boldsymbol{u}(r), r)$.

Also, by the Itô formula (see [38]),

$$\begin{aligned}
|\boldsymbol{\xi}(t)|_p^p &= |\boldsymbol{\xi}(0)|_p^p - p\int_0^{t\wedge\tau} N_p(\boldsymbol{\xi}(r), r)\,dr + p\int_0^{t\wedge\tau}\int |\boldsymbol{\xi}(r)|^{p-2}\xi^l(r) H^l(r)\,dx\,dr \\
&+ p\int_0^{t\wedge\tau}\int |\boldsymbol{\xi}(r)|^{p-2}(\boldsymbol{\xi}(r), \boldsymbol{r}_{n'}(\bar{\boldsymbol{u}}(r)) - \boldsymbol{r}_n(\boldsymbol{u}(r))\,dx\,dr \\
&+ p\int_0^{t\wedge\tau}\int |\boldsymbol{\xi}(r)|^{p-2}\xi^l(r)\kappa^l(r)\,dx\,dW_r \\
&+ \frac{p}{2}\int_0^{t\wedge\tau}\left(\int [(p-2)|\boldsymbol{\xi}(r)|^{p-4}\xi^i(r)\xi^j(r) + |\boldsymbol{\xi}(r)|^{p-2}\delta_{ij}]\bar{\kappa}^{ij}(r)\,dx\right) dr \\
&- p\int_0^{t\wedge\tau}\int |\boldsymbol{\xi}(s)|^{p-2}\xi^l(s)[\partial_k \xi^l(s)\Psi_{n'}(\bar{u}^k) + (\Psi_{n'}(\bar{u}^k) - \Psi_n(u^k))\partial_k \eta_n^l\,dx\,ds,
\end{aligned}$$

where $\boldsymbol{\kappa}(r) = (\kappa^i(r))_i = \sigma^k\partial_k\boldsymbol{\xi}(r) + \boldsymbol{B}(\bar{\boldsymbol{u}}, r) - \boldsymbol{B}(\boldsymbol{u}, r))$,

$$\boldsymbol{H}(r) = (H^l(r))_l = b^i\partial_i\boldsymbol{\xi}(r) + \boldsymbol{H}(\bar{\boldsymbol{u}}(r), r) - \boldsymbol{H}(\boldsymbol{u}(r), r));$$

and

$$\bar{\kappa}^{ij}(r) = \sigma^k(r)\partial_k\xi^i(r)\cdot D^j(r) + \sigma^k(r)\partial_k\xi^j(r)\cdot D^i(r) + D^i(r)\cdot D^j(r),$$

where $D^i(r) = B^i(\bar{\boldsymbol{u}}(r), r) - B^i(\boldsymbol{u}(r), r)$.

By Lemmas 20 (a) and 21 (b) and the Sobolev embedding theorem

$$\begin{aligned}\bar{L}_1 &= \left| \int |\boldsymbol{w}(r)|^{p-2}(\boldsymbol{w}(r), \mathcal{S}[\partial_k \boldsymbol{w}(r)\Psi_{n'}(\bar{u}^k) + (\Psi_{n'}(\bar{u}^k) - \Psi_n(u^k))\partial_k \boldsymbol{u}]\, dx \right| \\ &\le \varepsilon \int |\nabla \boldsymbol{w}(r)|^2 |\boldsymbol{w}(r)|^{p-2}\, dx + C_\varepsilon(|\boldsymbol{w}(r)|_p^p + |\boldsymbol{w}(r)A|_p^p + |\boldsymbol{w}(r)B|_p^p) \\ &\le \varepsilon \int |\nabla \boldsymbol{w}(r)|^2 |\boldsymbol{w}(r)|^{p-2}\, dx + C_\varepsilon(M)(|\boldsymbol{w}(r)|_p^p + |\boldsymbol{\xi}(r)|_p^p + n^{-\nu p}),\end{aligned} \tag{3.44}$$

where $A = |\Psi_{n'}(\bar{u}^k) - \Psi_n(u^k)|, B = |\Psi_{n'}(\bar{u}^k)|, \nu = 1 - d/p$.

Similarly, by Lemmas 20 (b) and 21 (b),

$$\begin{aligned}\bar{L}_2 &= \left| \int |\boldsymbol{\xi}(s)|^{p-2}\xi^l(s)(\Psi_{n'}(\bar{u}^k) - \Psi_n(u^k))\partial_k \eta_n^l\, dx \right| \\ &\le \varepsilon \int |\nabla \xi(r)|^2 |\boldsymbol{\xi}(r)|^{p-2}\, dx + C_\varepsilon(|A\boldsymbol{\eta}_n(r)|_p^p + |\boldsymbol{\xi}(r)|_p^p) \\ &\le \varepsilon \int |\nabla \xi(r)|^2 |\boldsymbol{\xi}(r)|^{p-2}\, dx + C_\varepsilon(M)(|\boldsymbol{w}(r)|_p^p + |\boldsymbol{\xi}(r)|_p^p + n^{-p\nu}).\end{aligned} \tag{3.45}$$

Obviously,

$$\begin{aligned}\bar{L}_3 &= \left| \int (|\boldsymbol{\xi}(r)|^{p-2}\boldsymbol{\xi}(r), \boldsymbol{\nabla}(\Psi_{n'}(\bar{u}^i)) \times \partial_i \bar{\boldsymbol{u}}(r) - \boldsymbol{\nabla}(\Psi_n(u^i)) \times \partial_i \boldsymbol{u}(r))\, dx \right| \\ &\le \left| \int (|\boldsymbol{\xi}(r)|^{p-2}\boldsymbol{\xi}(r), (\boldsymbol{\nabla}(\Psi_{n'}(\bar{u}^i)) - \boldsymbol{\nabla}(\Psi_n(u^i)) \times \partial_i \bar{\boldsymbol{u}}(r))\, dx \right| \\ &\quad + \left| \int (|\boldsymbol{\xi}(r)|^{p-2}\boldsymbol{\xi}(r), \boldsymbol{\nabla}(\Psi_{n'}(\bar{u}^i)) \times \partial_i \boldsymbol{w}(r))\, dx \right| = \bar{L}_{31} + \bar{L}_{32}.\end{aligned}$$

By Lemmas 22 and 21 (part (b)),

$$\begin{aligned}\bar{L}_{31} &\le \varepsilon \int |\boldsymbol{\xi}(r)|^{p-2} |\nabla \boldsymbol{\xi}(r)|^2\, dx + C_\varepsilon(|\boldsymbol{\xi}(r)|_p^p + |\nabla \bar{\boldsymbol{u}}(r)|_p^p |A|_\infty^p) \\ &\le \varepsilon \int |\boldsymbol{\xi}(r)|^{p-2} |\nabla \boldsymbol{\xi}(r)|^2\, dx + C_\varepsilon(M)(|\boldsymbol{\xi}(r)|_p^p + |\boldsymbol{w}(r)|_p^p + n^{=p\nu}),\end{aligned} \tag{3.46}$$

$$\begin{aligned}\bar{L}_{32} &\le \varepsilon \int |\boldsymbol{\xi}(r)|^{p-2} |\nabla \boldsymbol{\xi}(r)|^2\, dx + C_\varepsilon(|\boldsymbol{\xi}(r)|_p^p + |\nabla \boldsymbol{w}(r)|_p^p |B|_\infty^p) \\ &\le \varepsilon \int |\boldsymbol{\xi}(r)|^{p-2} |\nabla \boldsymbol{\xi}(r)|^2\, dx + C_\varepsilon(M)(|\boldsymbol{\xi}(r)|_p^p).\end{aligned}$$

Let $Z_t = |\boldsymbol{w}(t)|_p^p + |\boldsymbol{\xi}(t)|_p^p$. Using (3.44)–(3.46) and estimating the remaining terms by Lemmas 13 and 14, we obtain that for some adapted functions $a(t)$, $b(t)$ ($b(t)$ is Y-valued)

$$dZ_t = a(t)\, dt + b(t) \cdot dW_t, \tag{3.47}$$

and there is a constant C such that on $[0, \tau]$ for all $\tau \in \mathcal{T}_{n',n}$, $n' \ge n$,

$$a(t) \le C(Z_t + (1/n^{\nu p})),\ |b(t)|_Y \le C Z_t. \tag{3.48}$$

(We use **B3**(p) to estimate the term

$$\int [(p-2)|\boldsymbol{\xi}(r)|^{p-4}\xi^i(r)\xi^j(r) + |\boldsymbol{\xi}(r)|^{p-2}\delta_{ij}]\bar{\kappa}^{ij}(r)\, dx.)$$

Now part (a) of the statement follows by Lemma 36.

In case (b), we have ($\tau \in \tilde{\mathcal{T}}_{n',n}$)

$$
\begin{aligned}
&|\boldsymbol{w}(t)|_2^p = |\boldsymbol{w}(0)|_2^p - p\int_0^{t\wedge\tau} |\boldsymbol{w}(\boldsymbol{s})|_2^{p-2} N_2(\boldsymbol{w}(r),r)\,dr\\
&+ p\int_0^{t\wedge\tau} |\boldsymbol{w}(r)|_2^{p-2}\int (\boldsymbol{w}(r),\boldsymbol{a}(r))\,dx\,dr + p\int_0^{t\wedge\tau} |\boldsymbol{w}(r)|_2^{P-2}\left(\int w^l(r)c^l(r)\,dx\right)\,dW_r\\
&+ p/2\left(\int_0^{t\wedge\tau} |\boldsymbol{w}(r)|_2^{p-2}\left(\int \bar{c}^{ii}(r)\,dx\right)\,dr + (p-2)\int_0^{t\wedge\tau} |\boldsymbol{w}(r)|_2^{p-4}\Big|\int w^l(r)c^l(r)\,dx\Big|_Y^2\,dr\right)\\
&- \int_0^{t\wedge\tau} |\boldsymbol{w}(r)|_2^{p-2}\left(\int w^l(r)[\partial_k w^l(r)\Psi_{n'}(\bar{u}^k(r)) + (\Psi_{n'}(\bar{u}^k) - \Psi_\kappa(u^k))\partial_k u^l]\,dx\right)\,dr.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&|\boldsymbol{\xi}(t)|_2^p = |\boldsymbol{\xi}(0)|_2^p - p\int_0^{t\wedge\tau} |\boldsymbol{\xi}(r)|_2^{p-2} N_2(\boldsymbol{\xi}(r),r)\,ds\\
&- p\int_0^{t\wedge\tau} |\boldsymbol{\xi}(r)|_2^{p-2}\int (\boldsymbol{\xi}(r), \boldsymbol{\nabla}(\Psi_n(\bar{u}^i(r)))\times\partial_i\boldsymbol{w}(r)\\
&+ (\boldsymbol{\nabla}(\Psi_{n'}(u^i(r)) - \boldsymbol{\nabla}(\Psi_n(u^i(r)))\times\partial_i\boldsymbol{u}(r))\,dx\,dr\\
&p\int_0^{t\wedge\tau} |\boldsymbol{\xi}(r)|_2^{p-2}\left(\int (\boldsymbol{\xi}(r),\boldsymbol{H}(r)\right)\,dx\,dr + \int_0^{t\wedge\tau} p|\boldsymbol{\xi}(r)|_2^{p-2}\left(\int \xi^l(r)\kappa^l(r)\,dx\right)\,dW_r\\
&+ \frac{p}{2}\int_0^{t\wedge\tau} |\boldsymbol{\xi}(r)|_2^{p-2}\left(\int \bar{\kappa}^{ii}(r)\,dx\right)\,dr + \frac{p}{2}(p-2)\int_0^{t\wedge\tau} |\boldsymbol{\xi}(r)|_2^{p-4}\Big|\int \boldsymbol{\xi}^l(r)\kappa^l(r)\,dx\Big|_Y^2\,dr.
\end{aligned}
$$

Let $n' \geq n$. According to Lemmas 20 (part (a)) and 21 (part (a)) and the Sobolev embedding theorem, for each $\varepsilon > 0$, there is a constant C_ε such that $\boldsymbol{P}$-a.s. on $[0,\tau[$,

$$
\begin{aligned}
H &= \left|\int w^l(r)(\Psi_{n'}(\bar{u}^k(r)) - \Psi_\kappa(u^k(r)))\partial_k u^l(r)]\,dx\right|\\
&\leq \varepsilon|\nabla\boldsymbol{w}(r)|_2^2 + C_\varepsilon|\boldsymbol{u}(r)|_\infty^2|\Psi_{n'}(\bar{u}^k(r)) - \Psi_n(u^k(r)|_2^2\\
&\leq \varepsilon|\nabla\boldsymbol{w}(r)|_2^2 + C_\varepsilon M^2(|\boldsymbol{w}(r)|_{1,2}^2 + n^{-2}(|\nabla\bar{\boldsymbol{u}}(r)|_2^2 + |\nabla\bar{\boldsymbol{u}}(r)|_2^2)\\
&\leq \varepsilon|\nabla\boldsymbol{w}(r)|_2^2 + C_\varepsilon M^2(|\boldsymbol{w}(r)|_{1,2}^2 + n^{-2}),
\end{aligned}
$$

and

$$
|\boldsymbol{w}(r)|_2^{p-2}H \leq \varepsilon|\nabla\boldsymbol{w}(r)|_2^2|\boldsymbol{w}(r)|_2^{p-2} + C_\varepsilon(M)(|\boldsymbol{w}(r)|_2^p + |\boldsymbol{\xi}(r)|_2^p + n^{-p}). \tag{3.49}
$$

By Lemma 22 and the Sobolev imbedding theorem,

$$
\begin{aligned}
L_1 &= \left|\int (\boldsymbol{\xi}(r), \boldsymbol{\nabla}(\Psi_n(\bar{u}^i(r)))\times\partial_i\boldsymbol{w}(r)\,dx\right|\\
&\leq \varepsilon|\nabla\boldsymbol{\xi}(r)|_2^2 + C_\varepsilon(|\boldsymbol{\xi}(r)|_2^2 + |\Psi_n(\bar{u}^i(r))|_\infty^2|\nabla\boldsymbol{w}(r)|_2^2)\\
&\leq \varepsilon|\nabla\boldsymbol{\xi}(r)|_2^2 + C_\varepsilon(1+M^2)|\boldsymbol{\xi}(r)|_2^2.
\end{aligned} \tag{3.50}
$$

By Lemmas 22 and 21 (b),

$$
\begin{aligned}
L_2 &= \left| \int (\boldsymbol{\xi}(r), (\boldsymbol{\nabla}(\Psi_{n'}(\bar{u}^i(r)) - \boldsymbol{\nabla}(\Psi_n(u^i(r))) \times \partial_i \boldsymbol{u}(r))\, dx \right| \\
&\le \varepsilon |\nabla \boldsymbol{\xi}(r)|_2^2 + C_\varepsilon (|\boldsymbol{\xi}(r)|_2^2 + |\Psi_n(\bar{u}^i(r)) - \Psi_{n'}(\bar{u}^i(r))|_\infty^2 |\nabla \boldsymbol{u}(r)|_2^2) \\
&\le \varepsilon |\nabla \boldsymbol{\xi}(r)|_2^2 + C_\varepsilon [|\boldsymbol{\xi}(r)|_2^2 + M^2 (|\boldsymbol{w}(r)|_{1,p}^2 + n^{-2\nu} M^2)] \\
&\le \varepsilon |\nabla \boldsymbol{\xi}(r)|_2^2 + C_\varepsilon (M) (|\boldsymbol{\xi}(r)|_2^2 + |\boldsymbol{w}(r)|_{1,p}^2 + n^{-2\nu}),
\end{aligned} \tag{3.51}
$$

where $\nu = 1 - 2/p$. So,

$$
\begin{aligned}
&|\boldsymbol{\xi}(r)|_2^{p-2} L_2 \le \varepsilon |\nabla \boldsymbol{\xi}(r)|_2^2 |\boldsymbol{\xi}(r)|_2^{p-2} + C_\varepsilon (M) (|\boldsymbol{\xi}(r)|_2^p + |\boldsymbol{w}(r)|_{1,p}^p + n^{-2\nu}), \\
&|\boldsymbol{\xi}(r)|_2^{p-2} L_1 \le \varepsilon |\nabla \boldsymbol{\xi}(r)|_2^2 |\boldsymbol{\xi}(r)|_2^{p-2} + C_\varepsilon (M) |\boldsymbol{\xi}(r)|_2^p.
\end{aligned} \tag{3.52}
$$

Let $K_t = |\boldsymbol{w}(t)|_2^p + |\boldsymbol{\xi}(t)|_2^p$. Using (3.49)–(3.52) and estimating the remaining terms by Lemmas 13 and 14, we obtain that for some adapted functions $\bar{a}(t)$, $\bar{b}(t)$ ($\bar{b}(t)$ is Y-valued),

$$
dK_t = \bar{a}(t)\, dt + \bar{b}(t) \cdot dW_t, \tag{3.53}
$$

and there is a constant C such that on $[0, \tau]$ for all $\tau \in \mathcal{T}_{n',n}$, $n' \ge n$,

$$
\bar{a}(t) \le C(K_t + |\boldsymbol{w}(t)|_{1,p}^p + (1/n^p)),\ |\bar{b}(t)|_Y \le C K_t. \tag{3.54}
$$

(We use **B3**(2) to estimate $\int \bar{b}^{ii}(r)\, dx$.)

Combining (3.47), (3.53), (3.48), and (3.54), we find that for some adapted functions $\tilde{a}(t), \tilde{b}(t)$

$$
\begin{aligned}
R_t &= |\boldsymbol{w}(t)|_2^p + |\boldsymbol{\xi}(t)|_2^p + |\boldsymbol{w}(t)|_p^p + |\boldsymbol{\xi}(t)|_p^p \\
&= R(0) + \int_0^t \tilde{a}(r)\, dr + \int_0^t \tilde{b}(r)\ \cdot dW_r,
\end{aligned}
$$

and there is a constant C such that on any $[0, \tau], \tau \in \tilde{\mathcal{T}}_{n',n}$,

$$
\tilde{a}(t) \le C[R(t) + (1/n^p) + (1/n^{p\nu})],\ |\bar{b}(t)|_Y \le C R(t).
$$

Now part (b) of the statement follows by Lemma 36 (see the appendix). ∎

3.6. Local existence and uniqueness.

3.6.1. Uniqueness.

PROPOSITION 24. *Let τ be a bounded stopping time, $p > d$. Let* **B1**, **B2**(p) *be satisfied. Assume $\boldsymbol{u}(t)$ and $\bar{\boldsymbol{u}}(t)$ are $\mathbb{L}_p$-solutions of* (3.7) *in $[[0, \tau]]$ and also $\mathbb{H}_p^1$-valued and continuous.*

Then $\boldsymbol{P}$-a.s. $\boldsymbol{u}(t \wedge \tau) = \bar{\boldsymbol{u}}(t \wedge \tau)$ for all t.

Proof. For $\boldsymbol{v} \in \mathbb{H}_p^1$, we set

$$
\begin{aligned}
\tilde{\boldsymbol{G}}(\boldsymbol{v}, r) &= (\tilde{G}^l(\boldsymbol{v}, r))_{1 \le l \le d} = \mathcal{S}(\boldsymbol{G}(\boldsymbol{v}, r)) - \mathcal{G}(\sigma^i(r) \partial_i \boldsymbol{v}), \\
\tilde{\boldsymbol{F}}(\boldsymbol{v}, r) &= \mathcal{S}(\boldsymbol{F}(\boldsymbol{v}, r) + b^i(r) \partial_i \boldsymbol{v}) - \partial_i \mathcal{G}(a^{ij}(r) \partial_j \boldsymbol{w}).
\end{aligned}
$$

Then for all $\boldsymbol{v}, \bar{\boldsymbol{v}} \in \mathbb{H}_p^1$,

$$
\begin{aligned}
|\tilde{\boldsymbol{G}}(\boldsymbol{v}, r) - \tilde{\boldsymbol{G}}(\bar{\boldsymbol{v}}, r)|_p &\le C |\boldsymbol{v} - \bar{\boldsymbol{v}}|_p, \\
|\tilde{\boldsymbol{F}}(\boldsymbol{v}, r) - \tilde{\boldsymbol{F}}(\bar{\boldsymbol{v}}, r)|_{-1,p} &\le C |\boldsymbol{v} - \bar{\boldsymbol{v}}|_p.
\end{aligned} \tag{3.55}
$$

Let

$$N_p(\boldsymbol{v},r) = -\int \{[|\boldsymbol{v}|^{p-2}v^l\partial_i(a^{ij}(r)\partial_j v^l) + 2^{-1}[(p-2)|\boldsymbol{v}|^{p-4}v^iv^j + |\boldsymbol{v}|^{p-2}\delta_{ij}]\sigma^k(r)\cdot\sigma^m(r)\partial_k v^i\partial_m v^j]\}\,dx.$$

Obviously (see (3.28)),

$$N_p(\boldsymbol{v},r) \geq \delta\int |\boldsymbol{v}|^{p-2}|\nabla\boldsymbol{v}|^2\,dx.$$

Let $\boldsymbol{w}(t) = \boldsymbol{u}(t\wedge\tau) - \bar{\boldsymbol{u}}(t\wedge\tau)$. By the Itô formula (see [38]),

$$\begin{aligned}|\boldsymbol{w}(t)|_p^p &= -p\int_0^{t\wedge\tau} N_p(\boldsymbol{w}(r),r)\,dr + p\int_0^{t\wedge\tau}\int |\boldsymbol{w}(r)|^{p-2}w^l(r)a^l(r)\,dx\,dr\\ &\quad + p\int_0^{t\wedge\tau}\int |\boldsymbol{w}(r)|^{p-2}w^l(r)c^l(s)\,dW_s\\ &\quad + \frac{p}{2}\int_0^{t\wedge\tau}\left(\int [(p-2)|\boldsymbol{w}(r)|^{p-4}w^i(r)w^j(r) + |\boldsymbol{w}(r)|^{p-2}\delta_{ij}]\bar{c}^{ij}(r)\,dx\right)dr\\ &\quad - p\int_0^{t\wedge\tau}\left(\int |\boldsymbol{w}(r)|^{p-2}(\boldsymbol{w}(r),\mathcal{S}[\partial_k\boldsymbol{w}(r)\bar{u}^k(r) + w^k(r)\partial_k\boldsymbol{u}(r)]\,dx\right)dr,\end{aligned}$$

where $\boldsymbol{c}(r) = (c^i(r))_{1\leq i\leq d} = \sigma^k\partial_k\boldsymbol{w}(r) + \tilde{\boldsymbol{G}}(\bar{\boldsymbol{u}},r) - \tilde{\boldsymbol{G}}(\boldsymbol{u},r)$,

$$\boldsymbol{a}(r) = (a^l(r))_l = \tilde{F}^l(\bar{\boldsymbol{u}},r) - \tilde{F}^l(\boldsymbol{u},r),$$

and

$$\bar{c}^{ij}(r) = \sigma^k(r)\partial_k w^i(r)\cdot d^j(r) + \sigma^k(r)\partial_k w^j(r)\cdot d^i(r) + d^i(r)\cdot d^j(r),$$

where $d^i(r) = \tilde{G}^i(\bar{\boldsymbol{u}}(r),r) - \tilde{G}^i(\boldsymbol{u}(r),r)$.

By (3.55), for each $\varepsilon>0$, there is a constant C_ε such that

$$\left|\int |\boldsymbol{w}(r)|^{p-2}w^l(r)a^l(r)\,dx\right| \leq \varepsilon\int |\boldsymbol{w}(r)|^{p-2}|\nabla\boldsymbol{w}(r)|^2\,dx + C_\varepsilon|\boldsymbol{w}(r)|_p^p. \tag{3.56}$$

Integrating by parts, and using the Sobolev embedding theorem ($p>d$) and the Hölder inequality, we obtain that for each $\varepsilon>0$ there is a constant C_ε such that

$$\begin{aligned}&\left|\int |\boldsymbol{w}(r)|^{p-2}(\boldsymbol{w}(r),\mathcal{S}[\partial_k\boldsymbol{w}(r)\bar{u}^k(r) + w^k(r)\partial_k\boldsymbol{u}(r)])\,dx\right|\\ &= \left|\int |\boldsymbol{w}(r)|^{p-2}(\boldsymbol{w}(r),\partial_k\mathcal{S}[\boldsymbol{w}(r)\bar{u}^k(r) + w^k(r)\boldsymbol{u}(r)]\,dx\right|\\ &\leq \varepsilon\int |\boldsymbol{w}(r)|^{p-2}|\nabla\boldsymbol{w}(r)|^2\,dx + C_\varepsilon\int |\boldsymbol{w}(r)|^{p-2}|\mathcal{S}[\boldsymbol{w}(r)\bar{u}^k(r) + w^k(r)\boldsymbol{u}|^2\,dx\\ &\leq \varepsilon\int |\boldsymbol{w}(r)|^{p-2}|\nabla\boldsymbol{w}(r)|^2\,dx + C_\varepsilon|\boldsymbol{w}(r)|_p^p(|\bar{\boldsymbol{u}}(r)|_{1,p} + |\boldsymbol{u}(r)|_{1,p}).\end{aligned} \tag{3.57}$$

By (3.55), for each $\varepsilon>0$ there is a constant C_ε such that

$$\begin{aligned}&\left|\int [(p-2)|\boldsymbol{w}(r)|^{p-4}w^i(r)w^j(r) + |\boldsymbol{w}(r)|^{p-2}\delta_{ij}]\bar{c}^{ij}(r)\,dx\right|\\ &\quad\leq \varepsilon\int |\boldsymbol{w}(r)|^{p-2}|\nabla\boldsymbol{w}(r)|^2\,dx + C_\varepsilon|\boldsymbol{w}(r)|_p^p.\end{aligned} \tag{3.58}$$

Integrating by parts and by (3.55), we have

$$\left|\int |\boldsymbol{w}(r)|^{p-2} w^l(r) c^l(r)\, dx\right| \le C|\boldsymbol{w}(r)|_p^p. \tag{3.59}$$

Let $M > 1$ and $\tau_M = \inf(t : |\bar{\boldsymbol{u}}(t)|_{1,p} + |\boldsymbol{u}(t)|_{1,p} \ge M) \wedge \tau$. Since (3.56)–(3.59) hold, the assumptions of Lemma 36 (see the appendix) are satisfied with

$$Z_t = |\boldsymbol{w}(t)|_p^p, c_t = \int |\boldsymbol{w}(t)|^{p-2} |\nabla \boldsymbol{w}(t)|^2\, dx, f_t = g_t = 0, Z_0 = 0.$$

Therefore, $\boldsymbol{P}$-a.s. $\boldsymbol{w}(t \wedge \tau_M) = \boldsymbol{0}$ for all t. Since M is arbitrary, pathwise uniqueness follows. □

3.6.2. Existence. Now we extract a converging subsequence.

LEMMA 25. (a) *Let* **B1**, **B2**(p), **B3**(p) *be satisfied* $(p > d)$, $\boldsymbol{E}(|\boldsymbol{u}_0|_{1,p}^p) < \infty$. *Then there is a bounded stopping time* τ *such that* $\boldsymbol{P}(\tau > 0) = 1$ *and a unique* $\mathbb{L}_p$*-solution* $\boldsymbol{u}(t)$ *of* (3.7) *in* $[[0, \tau]]$ *which is also an* $\mathbb{H}_p^1$*-valued continuous process such that*

$$E \sup_{t \le \tau} |\boldsymbol{u}(t)|_{1,p}^p < \infty.$$

(b) *Let* **B1**, **B2**(p), **B2**(2), **B3**(p), **B3**(2) *be satisfied* $(p > d = 2)$, *and*

$$E(|\boldsymbol{u}_0|_{1,p}^p + |\boldsymbol{u}_0|_{1,2}^p) < \infty,$$
$$\int_0^t (||\boldsymbol{G}(\boldsymbol{0}, r)||_{1,2}^p + |\boldsymbol{F}(\boldsymbol{0}, r)|_{1,2}^p)\, dr < \infty$$

$\boldsymbol{P}$*-a.s. for all* t. *Then there is a stopping time* τ *such that* $\boldsymbol{P}(\tau > 0) = 1$ *and a unique* $\mathbb{L}_p \cap \mathbb{L}_2$*-solution* $\boldsymbol{u}(t)$ *of* (3.7) *in* $[[0, \tau]]$, *which is also an* $\mathbb{H}_p^1 \cap \mathbb{H}_2^1$*-valued continuous process such that*

$$E \sup_{t \le \tau} (|\boldsymbol{u}(t)|_{1,p}^p + |\boldsymbol{u}(t)|_{1,2}^p) < \infty.$$

Proof. We apply Lemma 37 (see the appendix) to extract a converging subsequence. We choose the Banach space

$$B = \begin{cases} \mathbb{H}_p^1 & \text{in case (a)}, \\ \mathbb{H}_p^1 \cap \mathbb{H}_2^1 & \text{in case (b)}. \end{cases}$$

In $\mathbb{H}_p^1 \cap \mathbb{H}_2^1$ we use the norm

$$|\boldsymbol{v}|_B = (|\boldsymbol{v}|_p^p + |\operatorname{curl} \boldsymbol{v}|_p^p + |\boldsymbol{v}|_2^p + |\operatorname{curl} \boldsymbol{v}|_2^p)^{1/p}.$$

In $\mathbb{H}_p^1$ the norm $|\boldsymbol{v}|_B = (|\boldsymbol{v}|_p^p + |\operatorname{curl} \boldsymbol{v}|_p^p)^{1/p}$ is used.

Fix arbitrary $T_0 > 0, M_0 > 1$. Since Lemma 23 holds, according to Lemma 37, it is enough to prove that

$$\lim_{T \to 0} \sup_{n, \tau \in \mathcal{T}_n^{M_0, T_0}} \boldsymbol{P}(\sup_{s \le \tau \wedge T} |\boldsymbol{u}_n(s)|_B > |\boldsymbol{u}_n(0)|_B + M_0 - 1) = 0, \tag{3.60}$$

where $\mathcal{T}_n^{M_0,T_0}$ is the set of all stopping times $\tau \le T_0$ such that $\sup_{s\le\tau}|\boldsymbol{u}_n(s)|_B \le M_0 + |\boldsymbol{u}_n(0)|_B$. Let $T < T_0$;

$$S_n = \inf(t : |\boldsymbol{u}_n(t)|_B > |\boldsymbol{u}_n(0)|_B + M_0 - 1).$$

Let

$$K_t = \begin{cases} \int_0^t (||\boldsymbol{G}(\mathbf{0},r)||_{1,p}^p + |\boldsymbol{F}(\mathbf{0},r)|_{1,p}^p)\,dr & \text{in case (a)},\\ \int_0^t (||\boldsymbol{G}(\mathbf{0},r)||_{1,p}^p + |\boldsymbol{F}(\mathbf{0},r)|_{1,p}^p + ||\boldsymbol{G}(\mathbf{0},r)||_{1,2}^p + |\boldsymbol{F}(\mathbf{0},r)|_{1,2}^p)\,dr & \text{in case (b)}. \end{cases}$$

Define $\tau^M = \inf(t : K_t \ge M) \wedge T_0$. By Propositions 18 and 19, for $\tau \in \mathcal{T}_n^{M_0,T_0}$, we have for each M

$$\begin{aligned} &\boldsymbol{P}\left(\sup_{t\le\tau\wedge T}|\boldsymbol{u}_n(t)|_B > |\boldsymbol{u}_n(0)|_B + M_0 - 1\right)\\ &\le \boldsymbol{P}(|\boldsymbol{u}_n(S_n\wedge T)|_B > |\boldsymbol{u}_n(0)|_B + M_0 - 1)\\ &\le \boldsymbol{P}(|\boldsymbol{u}_n(S_n\wedge T)|_B^p > |\boldsymbol{u}_n(0)|_B^p + (M_0-1)^p)\\ &\le \boldsymbol{P}(|\boldsymbol{u}_n(S_n\wedge\tau^M\wedge T)|_B^p > |\boldsymbol{u}_n(0)|_B^p + (M_0-1)^p)\\ &\quad + \boldsymbol{P}\left(\tau^M < T_0\right) \le C(M_0)[T + EK_{\tau^M\wedge T}] + \boldsymbol{P}\left(\tau^M < T_0\right). \end{aligned}$$

Therefore, for each M,

$$\limsup_{T\to 0}\sup_{n,\tau\in\mathcal{T}_n^{M_0,T_0}} \boldsymbol{P}\left(\sup_{s\le\tau\wedge T}|\boldsymbol{u}_n(s)|_B > |\boldsymbol{u}_n(0)|_B + M_0 - 1\right) \le \boldsymbol{P}\left(\tau^M < T_0\right),$$

and (3.60) follows. By Lemma 37, there is a stopping time τ such that $\boldsymbol{P}(\tau > 0) = 1$, a B-valued stochastic process $\boldsymbol{u}$ on the interval $[0,\tau]$ and a subsequence $\boldsymbol{u}_{n_k}$ converging uniformly on $[0,\tau]$ to $\boldsymbol{u}$**.** Obviously, $\boldsymbol{u}(t)$ is an $\mathbb{L}_p$-solution (respectively, $\mathbb{L}_p \cap \mathbb{L}_2$-solution) of (3.7) in $[[0,\tau]]$, which is also $\mathbb{H}_p^1$ (respectively, $\mathbb{H}_p^1 \cap \mathbb{H}_2^1$) valued and continuous. Uniqueness follows by Proposition 24. □

The following almost obvious statement is a straightforward generalization of Lemma 25.

LEMMA 26. (a) *Let* **B1**, **B2**(p), **B3**(p) *be satisfied* $(p > d)$ *and* $\boldsymbol{E}(|\boldsymbol{u}_0|_{1,p}^p) < \infty$. *Assume that* $\boldsymbol{u}(t)$ *is an* $\mathbb{H}_p^1$*-valued continuous* $\mathbb{L}_p$*-solution of* (3.7) *on* $[0,S]$*, where* S *is a finite stopping time and*

$$E\sup_{t\le S}|\boldsymbol{u}(t)|_{1,p}^p < \infty.$$

Then there exist a finite stopping time τ *and an* $\mathbb{H}_p^1$*-valued continuous* $\mathbb{L}_p$*-solution* $\boldsymbol{v}(t)$ *to* (3.7) *in* $[[0,\tau]]$ *such that* $\boldsymbol{P}(\tau > S) = 1$ *and* $\boldsymbol{v}$ *coincides with* $\boldsymbol{u}$ *on* $[0,S]$ *and*

$$E\sup_{t\le\tau}|\boldsymbol{v}(t)|_{1,p}^p < \infty.$$

(b) *Let* **B1**, **B2**(p), **B3**(p), **B2**(2), **B3**(2) *be satisfied, and*

$$E(|\boldsymbol{u}_0|_{1,p}^p + |\boldsymbol{u}_0|_{1,2}^p) < \infty,$$

$$\int_0^t (||\boldsymbol{G}(\mathbf{0},r)||_{1,2}^p + |\boldsymbol{F}(\mathbf{0},r)|_{1,2}^p)\,dr < \infty$$

$\boldsymbol{P}$-a.s. for all t. Assume that $\boldsymbol{u}(t)$ is an $\mathbb{H}^1_p \cap \mathbb{H}^1_2$-valued continuous $\mathbb{L}_p \cap \mathbb{L}_2$-solution of (3.7) *in* $[[0, S]]$*, where S is a finite stopping time and*

$$E \sup_{t \le S} (|\boldsymbol{u}(t)|^p_{1,p} + |\boldsymbol{u}(t)|^p_{1,2}) < \infty.$$

Then there exist a finite stopping time τ and an $\mathbb{H}^1_p \cap \mathbb{H}^1_2$-valued continuous $\mathbb{L}_p \cap \mathbb{L}_2$-solution $\boldsymbol{v}(t)$ of (3.7) *in* $[[0, \tau]]$ *such that $\boldsymbol{P}(\tau > S) = 1$, and $\boldsymbol{v}$ coincides with $\boldsymbol{u}$ on* $[0, S]$ *and*

$$E \sup_{t \le \tau} (|\boldsymbol{v}(t)|^p_{1,p} + |\boldsymbol{v}(t)|^p_{1,2}) < \infty.$$

Now we can prove the main result.

3.6.3. Proof of Theorem 3. We follow here with the proof of Theorem 14.21 in [22]. Consider the set $\mathcal{S}$ of all finite stopping times S such that an $\mathbb{H}^1_p$ (respectively, $\mathbb{H}^1_p \cap \mathbb{H}^1_2$)-valued continuous $\mathbb{L}_p$ (respectively, $\mathbb{L}_p \cap \mathbb{L}_2$)-solution $\boldsymbol{u}(t)$ of (3.7) exists in $[[0, S]]$ and

$$E \sup_{t \le S} |\boldsymbol{u}(t)|^p_{1,p} < \infty \text{ (respectively, } E \sup_{t \le S} (|\boldsymbol{u}(t)|^p_{1,p} + |\boldsymbol{u}(t)|^p_{1,2}) < \infty).$$

By Lemma 25, $\mathcal{S}$ is not empty. It is closed with respect to the finite minimum and finite maximum operations. Let ζ be the essential upper bound of the set $\mathcal{S}$. So, there is a sequence $T_n \in \mathcal{S}$ increasing to ζ. Let $\boldsymbol{U}_n$ be a corresponding sequence of solutions on $[0, T_n]$. Since Proposition 24 holds, the sequence $\boldsymbol{U}_n$ defines a continuous process $\boldsymbol{u}$ on $\cup_n [0, T_n]$.

Let $y_t = |\boldsymbol{U}(t)|_{1,p}$ (respectively, $y_t = |\boldsymbol{u}(t)|_{1,p} + |\boldsymbol{u}(t)|_{1,2}$). Let $R_m = \zeta \wedge \inf(t : y_t \ge m)$. Then $T_q \wedge R_m \in \mathcal{S}$ and $\boldsymbol{u}(\cdot \wedge T_q \wedge R_m)$ is a solution in $[[0, T_q \wedge R_m]]$. Passing to a limit as $q \to \infty$, we obtain that $R_m \in \mathcal{S}$ and $\boldsymbol{u}(\cdot \wedge R_m)$ is a solution in $[[0, R_m]]$. If $\boldsymbol{P}(R_m = \zeta < \infty) > 0$, Lemma 26 would imply that there is a stopping time $S \in \mathcal{S}$ such that $S \ge R_m$ and $\boldsymbol{P}(R_m = \zeta < S) > 0$. This would contradict the definition of ζ. Thus $\boldsymbol{P}$-a.s. $R_m < \zeta$ on $\{\zeta < \infty\}$, and, obviously, $\limsup_{t \uparrow \zeta} y_t = \infty$ on $\{\zeta < \infty\}$. So, the sequence (R_m) "announces" ζ and ζ is a predictable stopping time. Obviously, $[0, S] \subseteq [0, \zeta)$ for all $S \in \mathcal{S}$. Let S be a stopping time such that $\boldsymbol{P}$-a.s. $S < \zeta$. Then $T_q \wedge S \in \mathcal{S}$ and $\boldsymbol{u}(\cdot \wedge T_q \wedge S)$ is a solution in $[[0, T_q \wedge S]]$. Passing to the limit as $q \to \infty$, we obtain that $\boldsymbol{u}(\cdot \wedge S)$ is a solution in $[[0, S]]$.

Let, in addition, $\boldsymbol{E}(|\boldsymbol{u}_0|^p_{2-2/p,p}) < \infty$. Let S be a stopping time such that $S < \zeta$ $\boldsymbol{P}$-a.s. Consider a linear equation in $[[0, S]]$ for $\boldsymbol{v}(t)$

(3.61)
$$\begin{cases} \partial_t \boldsymbol{v}(t) = \mathcal{S}[\partial_i \left(a^{ij}(t)\, \partial_j \boldsymbol{v}(t)\right) - u^k(t)\, \partial_k \boldsymbol{u}(t) + b^i(t) \partial_i \boldsymbol{u}(t) + \boldsymbol{F}(\boldsymbol{u}(t), t)] \\ \quad + \mathcal{S}[\sigma^i(t) \partial_i \boldsymbol{v}(t) + \boldsymbol{G}(\boldsymbol{u}(t), t)]\, \dot{W}_t, \\ \boldsymbol{v}(0) = \boldsymbol{u}_0. \end{cases}$$

By Theorem 3.3 in [37], in case (a) there is a unique $\mathbb{H}^1_p$-solution of (3.61) which is also a unique $\mathbb{L}_p$-solution. So, $\boldsymbol{u}(t) = \boldsymbol{v}(t)$ on $[0, S]$ and $\boldsymbol{u}(t)$ is an $\mathbb{H}^1_p$-solution to (3.7) in $[[0, S]]$. In case (b) we do the same using Corollary 3.7 in [37] and obtain that $\boldsymbol{u}(t)$ is an $\mathbb{H}^1_p \cap \mathbb{H}^1_2$-solution in $[[0, S]]$.

It remains to prove that, in case (a), $\lim_{t \uparrow \zeta} |\boldsymbol{u}(t)|_{1,p} = \infty$ on $\{\zeta < \infty\}$ if $\boldsymbol{E}(|\boldsymbol{u}_0|^p_{2-2/p,p}) < \infty$.

Fix an arbitrary $m > 1$. Let

$$\tau_{m+1} = \inf(t : y_t \geq m + 1).$$

Define a sequence of stopping times

$$S_1 = \inf(t > \tau_{m+1} : y_t \leq m) \wedge \zeta,\ S_{2n} = \inf(t > S_{2n-1} : y_t \geq m + 1) \wedge \zeta,$$
$$S_{2n+1} = \inf(t > S_{2n} : y_t \leq m) \wedge \zeta.$$

Let $S = \lim_n S_n$. Applying the Itô formula (see the proof of Proposition 18), we find that for some adapted $a(r), b(r)$

$$y(t \wedge S) = |\boldsymbol{u}(t \wedge S)|_p^p + |\operatorname{curl} \boldsymbol{u}(t \wedge S)|_p^p = y(0) + \int_0^{t\wedge S} a(r)\, dr + \int_0^{t\wedge S} b(r) \cdot dW_r, \tag{3.62}$$

and

$$a_r \leq C(y_r + y_r^{1+\mu} + z_r),\ |b_r|_Y \leq C(y_r + y_r^{1-1/p} \tilde{z}_r^{1/p}), \tag{3.63}$$

where $z_r = |\boldsymbol{F}(\boldsymbol{0}, r)|_{1,p}^p + ||\boldsymbol{G}(\boldsymbol{0}, r)||_{1,p}^p$, $\tilde{z}_r = ||\boldsymbol{G}(\boldsymbol{0}, r)||_{1,p}^p, \mu > 0$.

We will prove that $\boldsymbol{P}(S = \zeta < \infty) = 0$. Since m is arbitrary, this will imply that $\boldsymbol{P}$-a.s. $\lim_{t\uparrow\zeta} |\boldsymbol{u}(t)|_{1,p} = \infty$ on $\{\zeta < \infty\}$.

For $M > 1$ we set

$$\tau^M = \inf\left(t : \int_0^t (|\boldsymbol{F}(\boldsymbol{0}, r)|_{1,p}^p + ||\boldsymbol{G}(\boldsymbol{0}, r)||_{1,p}^p)\, dr \geq M\right).$$

It is enough to prove that for all q, M,

$$\boldsymbol{P}(S = \zeta < q \wedge \tau^M) = 0.$$

If $\boldsymbol{P}(S = \zeta < q \wedge \tau^M) > 0$ for some M, q, then by (3.62), (3.63),

$$\infty = E \sum_{k\geq 1} [y(S_{2k}) - y(S(2k-1)] \leq C(m, M) < \infty.$$

The statement follows.

3.7. Stochastic Navier–Stokes equation in two dimensions.

LEMMA 27. *Let* **B1**, **B2**(p), **B3**(p), **B2**(2), **B3**(2) *be satisfied,* $p > d = 2$, *and*

$$E(|\boldsymbol{u}_0|_{2-2/p,p}^p + |\boldsymbol{u}_0|_{1,2}^p) < \infty,$$
$$\int_0^t (||\boldsymbol{G}(\boldsymbol{0}, r)||_{1,2}^p + |\boldsymbol{F}(\boldsymbol{0}, r)|_{1,2}^p)\, dr < \infty$$

$\boldsymbol{P}$*-a.s. for all* t.

Then there is a maximal unique $\mathbb{H}_p^1 \cap \mathbb{H}_2^1$*-solution* $\boldsymbol{u}(t)$ *of* (3.7), *and for some* $\mathbb{F}$*-adapted functions* $a_l(t), b_l(t)$ *(*$a_l(t)$ *is real valued and* $b_l(t)$ *is* Y*-valued),* $l = p, 2$, $\boldsymbol{P}$*-a.s. on* $[0, S] \subseteq [0, \zeta)$

$$|\boldsymbol{\eta}(t)|_l^p = |\boldsymbol{\eta}(0)|_l^p + \int_0^t a_l(r)\ ds + \int_0^t b_l(r) \cdot dW_s,$$
$$|\boldsymbol{u}(t)|_2^p = |\boldsymbol{u}(0)|_2^p + \int_0^t a(r)\ ds + \int_0^t b(r) \cdot dW_s,$$

$l = p, 2$, $\boldsymbol{\eta}(t) = curl\, \boldsymbol{u}(t)$. *Moreover, there is a constant C independent of S such that*

$$\begin{aligned}
|a_l(r)| &\le C[|\boldsymbol{\eta}(r)|_2^p + |\boldsymbol{u}(r)|_l^p + ||\boldsymbol{G}(\mathbf{0}, r)||_{1,l}^p + |\boldsymbol{F}(\mathbf{0}, r)|_{1,l}^p],\\
|b_l(r)|_Y &\le C[|\boldsymbol{\eta}(r)|_l^p + |\boldsymbol{u}(r)|_l^p + ||\boldsymbol{\eta}(r)|_l^{p-1}||\boldsymbol{G}(\mathbf{0}, r)|_{1,l}),\\
|a(r)| &\le C[|\boldsymbol{u}(r)|_2^p + ||\boldsymbol{G}(\mathbf{0}, r)||_{1,2}^p + |\boldsymbol{F}(\mathbf{0}, r)|_{1,2}^p],\\
|b(r)|_Y &\le C[|\boldsymbol{u}(r)|_2^p + |\boldsymbol{u}(r)|_2^{p-1}||\boldsymbol{G}(\mathbf{0}, r)||_2),
\end{aligned}$$

$l = p, 2$.

Proof. The existence of a unique maximal $\mathbb{H}_p^1 \cap \mathbb{H}_2^1$-solution $(\boldsymbol{u}(t), \zeta)$ is guaranteed by Theorem 3. Since in two dimensions $\nabla v^i \times \partial_i \boldsymbol{v} = \mathbf{0}$, the following "regular growth" equation holds for $\boldsymbol{\eta}(t)$ in any $[[0, S]] \subseteq [0, \zeta)$ (cf. (3.27)):

$$\begin{aligned}
\partial_t \boldsymbol{\eta}\,(t) = \partial_i \left(a^{\,ij}\,(t)\,\partial_j \boldsymbol{\eta}\,(t)\right) - u^i\,(t)\,\partial_i \boldsymbol{\eta}(t) + b^i(t)\partial_i \boldsymbol{\eta}(t) + \boldsymbol{H}\,(\boldsymbol{u}(t), t)\\
+ \left[\sigma^i(t)\partial_i \boldsymbol{\eta}\,(t) + \boldsymbol{B}\,(\boldsymbol{u}(t), t))\right] dW_t, \quad \boldsymbol{\eta}\,(0) = \operatorname{curl} \boldsymbol{u}_0.
\end{aligned}$$

By the Itô formula (see [38]),

$$\begin{aligned}
|\boldsymbol{u}(t \wedge S)|_2^p = |\boldsymbol{u}(0)|_2^p - p \textstyle\int_0^{t\wedge S} |\boldsymbol{u}(r)|_2^{p-2} N_2(\boldsymbol{u}(r))\, ds\\
+ p \textstyle\int_0^{t\wedge S} |\boldsymbol{u}(r)|_2^{p-2} \int (\boldsymbol{u}(r), \bar{\boldsymbol{a}}(r))\, dx\, dr + p \int_0^{t\wedge S} |\boldsymbol{u}(r)|_2^{p-2} (\int u^l(r) c^l(r)\, dx)\, dW_r\\
+ p/2 \textstyle\int_0^{t\wedge S} |\boldsymbol{u}(r)|_2^{p-2} (\int \bar{b}^{ii}(r)\, dx)\, dr + \frac{p}{2}(p-2) \int_0^{t\wedge S} |\boldsymbol{u}(r)|_2^{q-4} |\int u^l(r) c^l(r)\, dx|_Y^2\, dr,
\end{aligned}$$

where

$$\begin{aligned}
\bar{\boldsymbol{a}}(r) &= b^i(r)\partial_i \boldsymbol{u}(r) + \boldsymbol{F}(\boldsymbol{u}(r), r),\\
\bar{b}^{ij}(r) &= \sigma^k(r)\partial_k u^i(r) \cdot d^j(r) + \sigma^k(r)\partial_k u^j(r) \cdot d^i(r) + d^i(r) \cdot d^j(r),
\end{aligned}$$

and $\boldsymbol{d}(r) = (d^i(r)) = \tilde{\boldsymbol{G}}(\boldsymbol{u}(r), r))$, $\boldsymbol{c}(r) = (c^l(r)) = \sigma^k(r)\partial_k \boldsymbol{u}(r) + \boldsymbol{d}(r)$.

Similarly, for each stopping time S so that $[0, S] \subseteq [0, \zeta)$,

$$\begin{aligned}
|\boldsymbol{\eta}(t \wedge S)|_2^p = |\boldsymbol{\eta}(0)|_2^p - p \int_0^{t\wedge S} |\boldsymbol{\eta}(r)|_2^{p-2} N_2(\boldsymbol{\eta}(r), r)\, dr\\
+ p \int_0^{t\wedge S} |\boldsymbol{\eta}(r)|_2^{p-2} \int (\boldsymbol{\eta}(r), \boldsymbol{H}(r)\, dx\, dr\\
+ \int_0^{t\wedge S} p|\boldsymbol{\eta}(r)|_2^{p-2} \int \eta(r)\gamma(r)\, dx)\, dW_r + \frac{p}{2} \int_0^{t\wedge S} |\boldsymbol{\eta}(r)|_2^{p-2} \left(\int \bar{\gamma}(r)\, dx\right) dr\\
+ \frac{p}{2}(p-2) \int_0^{t\wedge S} |\boldsymbol{\eta}(r)|_2^{p-4} \left|\int \eta(r)\gamma(r)\, dx\right|_Y^2 dr,
\end{aligned}$$

where

$$\boldsymbol{H}(r) = b^i(r)\partial_i \boldsymbol{\eta}(r) + \boldsymbol{H}(\boldsymbol{u}(r), r)),\ \bar{\gamma}(r) = 2\sigma^k(r)\partial_k \eta \cdot \bar{d}(r) + |\bar{d}(r)|_Y^2,$$

and $\bar{d}(r) = \boldsymbol{\nabla}\sigma^i \times \partial_i \boldsymbol{u} + \operatorname{curl}(\boldsymbol{G}\,(\boldsymbol{u}, t))$, $\gamma(r) = \sigma_k(r)\partial_k \boldsymbol{\eta}(r) + \bar{d}(r)$. By Lemmas 13 and 14,

$$\begin{aligned}
|\boldsymbol{\eta}(t \wedge S)|_2^p = |\boldsymbol{\eta}(0)|_2^p + \int_0^{t\wedge S} a_2(r)\ ds + \int_0^{t\wedge S} b_2(r) \cdot dW_s,\\
|\boldsymbol{u}(t \wedge S)|_2^p = |\boldsymbol{u}(0)|_2^p + \int_0^{t\wedge S} a(r)\ ds + \int_0^{t\wedge S} b(r) \cdot dW_s,
\end{aligned}$$

and there is a constant C independent of S such that

$$\begin{aligned} a_2(r) &\le C[|\boldsymbol{\eta}(r)|_2^p + |\boldsymbol{u}(r)|_2^p + ||\boldsymbol{G}(\mathbf{0},r)||_{1,2}^p + |\boldsymbol{F}(\mathbf{0},r)|_{1,2}^p],\\ |b_2(r)|_Y &\le C[|\boldsymbol{\eta}(r)|_2^p + |\boldsymbol{u}(r)|_2^p + ||\boldsymbol{\eta}(r)|_2^{p-1}||\boldsymbol{G}(\mathbf{0},r)|_{1,2}),\\ a(r) &\le C[|\boldsymbol{u}(r)|_2^p + ||\boldsymbol{G}(\mathbf{0},r)||_{1,2}^p + |\boldsymbol{F}(\mathbf{0},r)|_{1,2}^p],\\ |b(r)|_Y &\le C[|\boldsymbol{u}(r)|_2^p + |\boldsymbol{u}(r)|_2^{p-1}||\boldsymbol{G}(\mathbf{0},r)||_2^p). \end{aligned}$$

Also by the Itô formula

$$\begin{aligned} |\boldsymbol{\eta}(t)|_p^p = |\boldsymbol{\eta}(0)|_p^p - p\int_0^t N_p(\boldsymbol{\eta}(r),r)\,ds + p\int_0^t \left\langle |\boldsymbol{\eta}(r)|^{p-2}\boldsymbol{\eta}(r), \bar{\boldsymbol{a}}(r)\right\rangle_{1,p} dr\\ + \int_0^t p\int |\boldsymbol{\eta}(r)|^{p-2}\eta(r)c(r)\,dx d\dot{W}_r + p\int_0^t \left(\int (p-2)|\boldsymbol{\eta}(r)|^{p-2}\bar{b}(r)\,dx\right) dr, \end{aligned}$$

where

$$\begin{aligned} \bar{\boldsymbol{a}}(r) &= b^i\partial_i\boldsymbol{\eta} + \operatorname{curl}(\boldsymbol{F}(\boldsymbol{u},r)) + \partial_i(\boldsymbol{\nabla}a^{ij}\times\partial_j\boldsymbol{u}) + (\boldsymbol{\nabla}b^i)\times\partial_i\boldsymbol{u}\\ &= b^i\partial_i\boldsymbol{\eta} + \operatorname{curl}(\boldsymbol{F}(\boldsymbol{u},r)) + \boldsymbol{r}(\boldsymbol{u}(r),r),\\ \bar{b}(r) &= 2\sigma^k\partial_k\eta\cdot d(r) + |\bar{d}(r)|_Y^2, \end{aligned}$$

and $\bar{\boldsymbol{d}} = \boldsymbol{\nabla}\sigma^i\times\partial_i\boldsymbol{u} + \operatorname{curl}(\boldsymbol{G}(\boldsymbol{u},t))$, $c(r) = \sigma^k(r)\partial_k\eta(r) + \bar{\boldsymbol{d}}(r)$. Using Lemmas 13 and 14, we obtain

$$|\boldsymbol{\eta}(t)|_p^p = |\boldsymbol{\eta}(0)|_p^p + \int_0^t a_p(r)\ ds + \int_0^t b_p(r)\cdot dW_s;$$

and there is a constant C independent of S such that

$$\begin{aligned} |a_p(r)| &\le C[|\boldsymbol{\eta}(r)|_p^p + |\boldsymbol{u}(r)|_p^p + ||\boldsymbol{G}(\mathbf{0},r)||_{1,p}^p + |\boldsymbol{F}(\mathbf{0},r)|_{1,p}^p],\\ |b_p(r)|_Y &\le C[|\boldsymbol{\eta}(r)|_p^p + |\boldsymbol{u}(r)|_p^p + ||\boldsymbol{\eta}(r)|_p^{p-1}||\boldsymbol{G}(\mathbf{0},r)|_{1,p}). \end{aligned}$$

So, the statement follows. □

Now we can complete the proof of Theorem 4.

3.7.1. Proof of Theorem 4. We have immediately the existence of a maximal solution by Theorem 3. It remains to prove that $\boldsymbol{P}(\zeta=\infty)=1$ and the estimate. Let

$$y_t = |\boldsymbol{u}(t)|_2^p + |\operatorname{curl}\boldsymbol{u}(t)|_2^p + |\operatorname{curl}\boldsymbol{u}(t)|_p^p,$$

$R_m = \inf(t : y_t \ge m)\wedge\zeta$. Since in two dimensions $\mathbb{L}_p$ is continuously embedded into $\mathbb{H}_2^1$ ($\mathbb{L}_p\subseteq\mathbb{H}_2^1$), we obtain by Lemma 27 that for some adapted functions $h(t),\kappa(t)$,

$$y_{t\wedge R_m} = y_0 + \int_0^{t\wedge R_m} h(r)\,dr + \int_0^{t\wedge R_m}\kappa(r)\cdot dW_r,$$

and there is a constant C independent of m such that

$$h(r)\le C(y_r+z_r),\ \ \kappa(r)\le C(y_r + y_r^{1-1/p}\tilde{z}_r^{1/p}),$$

where

$$z_r = ||\boldsymbol{G}(\mathbf{0},r)||^p_{1,p} + |\boldsymbol{F}(\mathbf{0},r)|^p_{1,p} + ||\boldsymbol{G}(\mathbf{0},r)||^p_{1,2} + |\boldsymbol{F}(\mathbf{0},r)|^p_{1,2},$$
$$\tilde{z}_r = ||\boldsymbol{G}(\mathbf{0},r)||^p_{1,p} + ||\boldsymbol{G}(\mathbf{0},r)||^p_{1,2}.$$

By Lemma 36 (see the appendix), for each T there is a constant C (independent of m) so that for all stopping times $\tau \le T$

$$E\sup_{t\le\tau} y_t \le CE\Big[y_0 + \int_0^\tau (||\boldsymbol{G}(\mathbf{0},r)||^p_{1,p} + |\boldsymbol{F}(\mathbf{0},r)|^p_{1,p} + ||\boldsymbol{G}(\mathbf{0},r)||^p_{1,2} + |\boldsymbol{F}(\mathbf{0},r)|^p_{1,2})\,dr\Big]. \tag{3.64}$$

Let

$$K_t = \int_0^t (||\boldsymbol{G}(\mathbf{0},r)||^p_{1,p} + |\boldsymbol{F}(\mathbf{0},r)|^p_{1,p} + ||\boldsymbol{G}(\mathbf{0},r)||^p_{1,2} + |\boldsymbol{F}(\mathbf{0},r)|^p_{1,2})\,dr.$$

For $M > 1$, set $\tau^M = \inf(t : K_t \ge M)$. Since the sequence R_m "announces" the predictable stopping time ζ, for each $T > 0$ and $M > 1$, we have

$$\begin{aligned}\boldsymbol{P}(R_m < T) &\le \boldsymbol{P}(y_{Rm\wedge T\wedge\tau^M} \ge m) + \boldsymbol{P}(\tau^M < T)\\ &\le m^{-1}Ey_{R_m\wedge T\wedge\tau^M} + \boldsymbol{P}(\tau^M < T).\end{aligned}$$

So, by (3.64), for each $M > 1$,

$$\limsup_m \boldsymbol{P}(R_m < T) \le \boldsymbol{P}(\tau^M < T).$$

Therefore $\lim_m \boldsymbol{P}(R_m < T) = 0$, and $\boldsymbol{P}(\zeta = \infty) = 1$. The statement follows.

4. Wiener chaos and moment theory.

4.1. Preliminaries. In this section we continue the study of global solutions of stochastic Navier–Stokes equations. We will deal with the equation

$$\begin{aligned}&\partial_t \boldsymbol{u} = \partial_i\left(a^{ij}\partial_j\boldsymbol{u}\right) + b^i\partial_i\boldsymbol{u} - u^k\partial_k\boldsymbol{u} + \boldsymbol{h}^i\cdot\mathcal{G}^i\left(\sigma^{ik}\partial_i\boldsymbol{u} + \boldsymbol{g}\right)\\ &-\boldsymbol{\nabla}P + \boldsymbol{f} + [\sigma^{ik}\partial_i\boldsymbol{u} + \boldsymbol{g} - \boldsymbol{\nabla}\tilde{P}]\cdot\dot{W}_t,\ \operatorname{div}\boldsymbol{u} = 0,\\ &\boldsymbol{u}(0,x) = \boldsymbol{u}_0(x)\end{aligned} \tag{4.1}$$

with the free forces $\boldsymbol{f} = \boldsymbol{f}(t,x)$ and $\boldsymbol{g} = \boldsymbol{g}(t,x)$ that do not include a solution as an independent variable. Since the existence of global solutions is proved only for $d = 2$, we restrict ourselves to this case.

Our goal now is to investigate how the SNS equation (4.1) propagates the chaos generated by the driving Brownian motion and randomness in the initial conditions. We are particularly interested in deriving formulas for the statistical moments of a solution to (4.1).

Let $(\Omega,\mathcal{F},\boldsymbol{P})$ be a complete probability space. Let $W(t)$ and ξ^0 be a cylindrical Brownian motion and a cylindrical Gaussian random variable in Y. We assume that $W(t)$ and ξ^0 are defined on $(\Omega,\mathcal{F},\boldsymbol{P})$ and independent.

Let us fix a positive number $T < \infty$. Let $\mathcal{F}_T$ be a $\boldsymbol{P}$-completion of $\boldsymbol{\sigma}\{\xi^0, W(r) : r \le t\}$ and $\mathcal{F}_t$ be the a $\boldsymbol{\sigma}$-algebra generated by $\cap_{t<s\le T}\boldsymbol{\sigma}\{\xi^0, W(r), r \le s\}$ and all

the negligible sets of $\mathcal{F}$. The filtration of right continuous $\boldsymbol{\sigma}$-algebras $(\mathcal{F}_t)_{t\le T}$ will be denoted by $\mathbb{F}$.

We will assume in the future that the initial value $\boldsymbol{u}_0(x)$ is random, but its randomness is due solely to its dependence on ξ^0.

To begin with we shall introduce additional notation and recall some basic facts about Wiener chaos theory (see, e.g., [19], [20], [34], [36], etc.).

Let us fix a positive number $T<\infty$. Let $\{m_k, k\ge 1\}$ be an orthonormal basis in $L_2(0,T)$ and $\{\ell_k, k\ge 1\}$ an orthonormal basis in Y. Write $\xi_i^k=\int_0^T m_i(s)\,dw^k(t)$, where $w^k(t)=(W(t),\ell_k)_Y$.

Without any loss of generality, we assume that $\mathcal{F}_0$ is generated by the sequence of independent standard (i.e., $\mathcal{N}(0,1)$) Gaussian random variables $\left\{\xi_i^0, i\ge 1\right\}$.

Let $\alpha=\left\{\alpha_i^k,\ k\ge 0; i\ge 1\right\}$ be a multiindex, i.e., for every (i,k), $\alpha_i^k\in\mathbb{N}=\{0,1,2,\dots\}$. We shall consider only such α that $|\alpha|=\sum_{k,i}\alpha_i^k<\infty$, i.e., only a finite number of α_i^k are nonzero, and we denote by $\mathcal{J}$ the set of all such multiindices. Obviously, if $\alpha\in\mathcal{J}$, the number $\alpha!=\Pi_{k,i}\alpha_i^k!$ is well defined. We also write $\alpha!=\prod_{k,l}(\alpha_k^l!)$.

For $\alpha\in\mathcal{J}$, we write

$$\zeta_\alpha:=\prod_{i=1}^{\infty}\prod_{k=0}^{\infty}H_{\alpha_i^k}(\xi_i^k),$$

where H_n is the nth Hermite polynomial $H_n(x)=(-1)^n(\frac{d^n}{dx^n}e^{-\frac{x^2}{2}})e^{\frac{x^2}{2}}$. The random variable ζ_α is often referred to as (unnormalized) αth Wick polynomial.

Let $\mathcal{J}_0$ be a subset of $\mathcal{J}$ consisting of all multiindices of the form

$$\alpha=\left\{\alpha_i^k, k\ge 0,\ i\ge 1:\alpha_i^k=0 \text{ if } k\ne 0\right\}.$$

We will often denote a multiindex from $\mathcal{J}_0$ by α^0. Obviously, for $\alpha^0\in\mathcal{J}$,

$$\zeta_{\alpha^0}:=\prod_{i=1}^{\infty}H_{\alpha_i^0}(\xi_i^0).$$

It is a standard fact that

$$E\zeta_\alpha\zeta_\beta=\begin{cases}0 \text{ if } \alpha\ne\beta\\ \alpha! \text{ if } \alpha=\beta\end{cases}. \tag{4.2}$$

The most important feature of the Wick polynomials ζ_α is that the set $\{\zeta_\alpha/\sqrt{\alpha!},\ \alpha\in\mathcal{J}\}$ is an orthonormal basis in $L_2(\Omega,\mathcal{F}_T,\boldsymbol{P})$ (see, e.g., [6], [36]) . This result is often referred to as the Cameron–Martin theorem. The following lemma is an obvious extension of the Cameron–Martin theorem to the vector case.

Let $\mathcal{H}$ be a separable Hilbert space and $\{h_i, i\ge 1\}$ be an orthonormal basis in $\mathcal{H}$.

LEMMA 28. *Let $\eta:\Omega\longrightarrow\mathcal{H}$ be an $\mathcal{F}$-measurable random variable so that $E\|\eta\|_{\mathcal{H}}^2<\infty$. Then, η admits the Wiener chaos expansion in $L^2(\Omega;\mathcal{H})$,*

$$\eta=\sum_{\boldsymbol{\alpha}\in\mathcal{J}}\frac{\hat\eta_\alpha}{\alpha!}\zeta_{\boldsymbol{\alpha}}, \tag{4.3}$$

where $\hat\eta_\alpha=E[\eta\zeta_{\boldsymbol{\alpha}}]:=\sum_{i=1}^{\infty}E[(\eta,h_i)\zeta_{\boldsymbol{\alpha}}]h_i$.

Moreover,

$$E\left\|\eta\right\|_{\mathcal{H}}^{2}=\sum_{\boldsymbol{\alpha}\in\mathcal{J}}\frac{\left|\hat{\eta}_{\alpha}\right|^{2}}{\alpha!}=\sum_{\boldsymbol{\alpha}\in\mathcal{J}}\sum_{i=1}^{\infty}\frac{1}{\alpha!}E[(\eta,h_{i})_{\mathcal{H}}\zeta_{\boldsymbol{\alpha}}]^{2}. \tag{4.4}$$

In the future, we will refer to the functions $\hat{\eta}_{\alpha}$ as unnormalized Hermite–Fourier coefficients, or simply Hermite–Fourier coefficients, of η.

4.2. Propagator. Suppose that the assumptions of Theorem 4 hold. Then, (4.1) has a unique $\mathbb{F}$-adapted global solution in $\mathbb{H}_{p}^{1}(\boldsymbol{R}^{2})\bigcap\mathbb{H}_{2}^{1}(\boldsymbol{R}^{2})$ and

$$E\sup_{s\leq T}\left(\left|\boldsymbol{u}(t)\right|_{1,p}^{2}+\left|\boldsymbol{u}(t)\right|_{1,2}^{p}\right)<\infty.$$

By (4.3), a solution of (4.1) allows the Wiener chaos expansion $\boldsymbol{u}\left(t,x\right)=\sum_{\boldsymbol{\alpha}\in\mathcal{J}}\frac{\hat{\boldsymbol{u}}_{\alpha}(t,x)}{\alpha!}\zeta_{\boldsymbol{\alpha}}$.

This equality holds for all t in $L_{2}\left(\Omega;\mathbb{H}_{2}^{1}\left(R^{2}\right)\right)$ as well as for all t,x in $L_{2}\left(\Omega;R^{2}\right)$. The latter is due to the well-known imbedding $H_{1,p}\subset C^{1-2/p}$.

Of course, the main problem of interest is how to characterize the Hermite–Fourier coefficients $\boldsymbol{u}_{\alpha}\left(t,x\right)$. It will be shown below that these coefficients verify a certain nonlinear system of equations. This system describes the pattern of deterministic propagation of randomness in (4.1).

In this section we shall make the following additional assumptions:

(C1) *The initial value u_{0} is a measurable $\mathcal{F}_{0}$-adapted function.*

(C2) *The coefficients α^{ij} and b^{i} are measurable functions on $[0,T)\times\boldsymbol{R}^{2}$; f^{l} are predictable functions on $[0,T)\times\boldsymbol{R}^{2}\times\Omega$; $\sigma^{l},h^{l,i}$ are Y-valued measurable functions on $[0,T)\times\boldsymbol{R}^{2}$; g^{l} are Y-valued predictable functions on $[0,T)\times\boldsymbol{R}^{2}\times\Omega$; and for all t,x $\sum_{i=0}^{1}\left(\left|\partial_{x}^{k}h^{l,i}\left(t,x\right)\right|_{Y}+\left|\partial_{x}^{k}\boldsymbol{\sigma}\left(t,x\right)\right|\right)\leq C$.*

(C3) *For $p>2$* and $l=2,p$,

$$\int_{0}^{T}E\left(\left|\boldsymbol{g}\left(t\right)\right|_{1,l}^{p}+\sum_{i}\left|\boldsymbol{f}\left(t\right)\right|_{1,l}^{p}\right)dt<\infty.$$

Note that, in contrast to the previous sections, we postulate that a^{ij}, b^{i}, and $h^{l,i}$ are nonrandom.

Now we introduce some additional notation.

Write

$$\widehat{u^{i}\partial_{i}\boldsymbol{u}}_{\alpha}\left(t\right)=\sum_{p\in\mathcal{J}}\sum_{0\leq\beta\leq\alpha}\frac{1}{p!}\binom{\alpha}{\beta}\hat{u}_{p+\beta}^{i}\left(t\right)\partial_{i}\hat{\boldsymbol{u}}_{p+\alpha-\beta}\left(t\right), \tag{4.5}$$

$$\mathcal{M}\left(\hat{\boldsymbol{u}}_{\alpha},t\right)=\sigma^{j}(t)\partial_{j}\hat{\boldsymbol{u}}_{\alpha}\left(t\right)+\hat{\boldsymbol{g}}_{\alpha}\left(t\right).$$

For $j\neq0$, we define multiindex $\alpha\left(i,j\right)\in\mathcal{J}$ by the formula

$$\alpha\left(i,j\right)_{l}^{k}=\begin{cases}\alpha_{l}^{k}\text{ if }(k,l)\neq(j,i)\text{ or }k=0,\\ \left(\alpha_{l}^{k}-1\right)\wedge0\text{ if }(k,l)=(j,i);\end{cases} \tag{4.6}$$

i.e., the multiindex $\alpha\left(i,j\right)$ might differ from α only by its (i,j) entry, which is equal to $\left(\alpha_{j}^{i}-1\right)\vee0$.

Set

$$D\mathcal{M}\left(\hat{\boldsymbol{u}}_\alpha, t\right) = \begin{cases} \sum_{k\neq 0}\left(\hat{\boldsymbol{g}}_{\alpha(i,k)}\left(t\right) + \sigma^j\left(t\right)\partial_j\hat{\boldsymbol{u}}_{\alpha(i,k)}\left(t\right)\right)\alpha_i^k m_i\left(t\right) & \text{if } \alpha\notin\mathcal{J}_0, \\ 0 & \text{otherwise,}\end{cases}$$

and

$$\mathcal{L}_0\left(\hat{\boldsymbol{u}}_\alpha, t\right) = \partial_i\left(a^{ij}\left(t\right)\partial_j\hat{\boldsymbol{u}}_\alpha\left(t\right)\right) + b^i(t)\partial_i\hat{\boldsymbol{u}}_\alpha(t) + \hat{\boldsymbol{f}}_\alpha\left(t\right) + \boldsymbol{h}^i(t)\mathcal{G}^i(\mathcal{M}\left(\hat{\boldsymbol{u}}_\alpha, t\right).$$

THEOREM 29. *Assume that $C1-C3$, as well as the assumptions of Theorem 4, hold.*

Then the Fourier–Hermite coefficients $\hat{\boldsymbol{u}}_\alpha$ of the global solution of SNS (4.1) are continuous $\mathbb{H}^1_p(\boldsymbol{R}^2)\bigcap\mathbb{H}^1_2(\boldsymbol{R}^2)$-valued functions on $[0,T]$. Moreover, the set of functions $\{\hat{\boldsymbol{u}}_\alpha\left(t,x\right),\alpha\in\mathcal{J}\}$ verifies the following system of equations:

$$\begin{cases} \left(\hat{\boldsymbol{u}}_\alpha\left(t\right),\boldsymbol{\varphi}\right)_2 = \left(\hat{\boldsymbol{u}}_\alpha\left(0\right),\boldsymbol{\varphi}\right)_2 + \int_0^t\{\langle\mathcal{L}_0\left(\hat{\boldsymbol{u}}_\alpha, s\right),\boldsymbol{\varphi}\rangle \\ + \left(\widehat{-u^i\partial_i\boldsymbol{u}_\alpha}\left(s\right) + D\mathcal{M}\left(\hat{\boldsymbol{u}}_\alpha, s\right),\boldsymbol{\varphi}\right)_2\}ds;\ \mathrm{div}\hat{\boldsymbol{u}}_\alpha\left(t\right) = 0 \\ \text{for all } t\le T \text{ and } \boldsymbol{\varphi}\in\left(C_0^\infty\left(\boldsymbol{R}^2\right)\right)^2 \text{ so that } \mathrm{div}\boldsymbol{\varphi} = 0.\end{cases} \tag{4.7}$$

4.2.1. Proof of Theorem 29. To begin with, we remark that, for $\alpha = 0$, $\hat{\boldsymbol{u}}_\alpha\left(0\right) = E\boldsymbol{u}_0$, and if α has at least one positive entry α_i^k with $k\neq 0, \hat{\boldsymbol{u}}_\alpha\left(0\right) = 0$.

By Theorem 4 we have that $\boldsymbol{u}\left(t\right)$, a solution of (4.1), is a continuous $\mathbb{H}^1_p\left(\boldsymbol{R}^2\right)\bigcap\mathbb{H}^1_2\left(\boldsymbol{R}^2\right)$-valued function and

$$E\sup_{t\le T}(|\boldsymbol{u}(t)|^p_{1,p} + |\boldsymbol{u}(t)|^p_{1,2}) < \infty. \tag{4.8}$$

Owing to (4.8), we have by the Hölder inequality that for $l = 2, p$

$$\sup_{t\le T}\left(|\hat{\boldsymbol{u}}_\alpha(t)|_l + |(\partial_i\boldsymbol{u})_\alpha(t)|_l\right) < \infty\ . \tag{4.9}$$

By the Fubini theorem, for all $\boldsymbol{\varphi}\in\left(C_0^\infty\left(\boldsymbol{R}^2\right)\right)^2$ and $t\le T$,

$$\left(\left(\partial_i\boldsymbol{u}(t)\right)_\alpha,\boldsymbol{\varphi}\right)_2 = -E\left[\left(\boldsymbol{u}\left(t\right),\partial_i\boldsymbol{\varphi}\right)_2\zeta_\alpha\right] = -\left(\hat{\boldsymbol{u}}_\alpha\left(t\right),\partial_i\boldsymbol{\varphi}\right)_2.$$

Thus,

$$\partial_i\hat{\boldsymbol{u}}_\alpha(t) = \widehat{\left(\partial_i\boldsymbol{u}\left(t\right)\right)}_\alpha. \tag{4.10}$$

Now, by (4.9) and (4.10) we have that for all t, $\hat{\boldsymbol{u}}_\alpha(t)\in\mathbb{H}^1_p(\boldsymbol{R}^2)\bigcap\mathbb{H}^1_2(\boldsymbol{R}^2)$. Since for integer n and $q\ge 1$, the norm $|\cdot|_{n,q}$ is equivalent to the norm of the Sobolev space $W^{n,q}$, by (4.10) and the Hölder inequality, we have that for $l = 2, p$

$$\begin{aligned} &|\hat{\boldsymbol{u}}_\alpha(t) - \hat{\boldsymbol{u}}_\alpha(s)|_{1,l} \\ &\le C\left(|E\left(\boldsymbol{u}(t) - \boldsymbol{u}(s)\right)\zeta_\alpha|_l + \sum_i|E\left(\partial_i\boldsymbol{u}(t) - \partial_i\boldsymbol{u}(s)\right)\zeta_\alpha|_l\right) \\ &\le C'E|\boldsymbol{u}(t) - \boldsymbol{u}(s)|_{1,l}.\end{aligned}$$

Thus, by the dominated convergence theorem we have that the Fourier–Hermite coefficients $\hat{\boldsymbol{u}}_\alpha\left(t\right)$ are continuous in $\mathbb{H}^1_p(\boldsymbol{R}^2)\bigcap\mathbb{H}^1_2(\boldsymbol{R}^2)$.

Owing to (4.10), we also have that, for every $\alpha \in \mathcal{J}$, $\operatorname{div}\hat{\boldsymbol{u}}_\alpha(t) = 0$.

We continue with two simple but useful lemmas.

LEMMA 30. *Let η and ψ be $\mathcal{F}$-measurable $\mathcal{H}$-valued random variables, and $E[\|\eta\|_{\mathcal{H}}^2 + \|\psi\|_{\mathcal{H}}^2] < \infty$.*

Then,

$$\begin{aligned}(\psi,\eta)_{\mathcal{H}} &= \sum_{\boldsymbol{\gamma},\boldsymbol{\beta}\in\mathcal{J}} \left(\hat{\psi}_\gamma, \hat{\eta}_\beta\right)_{\mathcal{H}} \sum_{p\le\gamma\wedge\beta} \left((\gamma-p)!\,(\beta-p)!p!\right)^{-1} \zeta_{\gamma+\beta-2p} \\ &= \sum_{\alpha,p\in\mathcal{J}} \sum_{0\le\beta\le\alpha} \binom{\alpha}{\beta} \frac{1}{\alpha!p!} \left(\hat{\psi}_{p+\alpha-\beta}, \hat{\eta}_{p+\beta}\right)_{\mathcal{H}} \zeta_\alpha. \qquad (4.11)\end{aligned}$$

Proof. It is a standard fact (see, e.g., [36]) that

$$\zeta_\gamma\zeta_\beta = \sum_{p\le\gamma\wedge\beta} \binom{\gamma}{p}\binom{\beta}{p} p!\zeta_{\gamma+\beta-2p}. \qquad (4.12)$$

By Lemma 28 and (4.12), we have

$$\begin{aligned}(\psi,\eta)_{\mathcal{H}} &= \sum_{\boldsymbol{\gamma},\boldsymbol{\beta}\in\mathcal{J}} \frac{1}{\gamma!\beta!} \left(\hat{\psi}_\gamma, \hat{\eta}_\beta\right)_{\mathcal{H}} \zeta_\gamma\zeta_\beta \\ &= \sum_{\boldsymbol{\gamma}',\boldsymbol{\beta}'\in\mathcal{J}} \left(\hat{\psi}_{\gamma'}, \hat{\eta}_{\beta'}\right)_{\mathcal{H}} \sum_{p\le\gamma'\wedge\beta'} \left((\gamma'-p)!\,(\beta'-p)!p!\right)^{-1} \zeta_{\gamma'+\beta'-2p}. \qquad (4.13)\end{aligned}$$

By making the change of variables $\alpha = \gamma' + \beta' - 2p, \beta = \beta' - p$ in (4.13) and observing that $\gamma' - p = \alpha - \beta$, we arrive at

$$(\psi,\eta)_{\mathcal{H}} = \sum_{\alpha,p\in\mathcal{J}} \sum_{0\le\beta\le\alpha} \binom{\alpha}{\beta} \frac{1}{\alpha!p!} \left(\hat{\psi}_{p+\alpha-\beta}, \hat{\eta}_{p+\beta}\right)_{\mathcal{H}} \zeta_\alpha. \qquad \square$$

LEMMA 31 (see [43]). *The process $\zeta_\alpha(t) = E[\zeta_\alpha|\mathcal{F}_t]$ verifies the following equation:*

$$d\zeta_\alpha(t) = m_i(t)\,\alpha_i^k \zeta_{\alpha(i,k)}(t) dw^k(t). \qquad (4.14)$$

REMARK 8. *Write $D\zeta_\alpha(t) = m_i(t)\,\alpha_i^k\zeta_{\alpha(i,k)}(t)\ell_k$, where as before, $(\ell_k)_{k\ge1}$ is an orthonormal basis in Y. It is readily checked (cf. [45]) that $D\zeta_\alpha(t)$ is the Malliavin derivative of $\zeta_\alpha(t)$. Now we can rewrite (4.14) in the following more compact and maybe more insightful form:*

$$d\zeta_\alpha(t) = D\zeta_\alpha(t)\cdot dW(t).$$

Note that since $(\zeta_\alpha(t), \mathcal{F}_t)$ is a uniformly integrable martingale, we can sharpen (4.3) as follows.

COROLLARY 32. *If $\boldsymbol{v} \in L^2(\Omega, \mathcal{F}_s, \boldsymbol{P}; \mathbb{L}_2)$ for some $s \in [0,T]$, then $\hat{\boldsymbol{v}}_\alpha = E[\boldsymbol{v}\zeta_\alpha(s)]$, and*

$$\boldsymbol{v} = \sum_{\alpha\in\mathcal{J}} \frac{\hat{\boldsymbol{v}}_\alpha}{\alpha!}\zeta_\alpha(s)$$

in $L^2(\Omega, \mathcal{F}_s, \boldsymbol{P}; \mathbb{L}_2)$.

Write $\mathcal{M}^k(\boldsymbol{u},t)=\sigma^{jk}(t)\partial_j\boldsymbol{u}(t)+\boldsymbol{g}^k(t)$ and $\mathcal{M}^k(\hat{\boldsymbol{u}}_\alpha,t)=\sigma^{jk}(t)\partial_j\hat{\boldsymbol{u}}_\alpha(t)+\hat{\boldsymbol{g}}^k_\alpha(t)$, where $\sigma^{jk}=\left(\sigma^i,\ell_k\right)_Y$, $\boldsymbol{g}^k=(\boldsymbol{g},\ell_k)_Y$, and $(\ell_k,k\geq 1)$ is an orthonormal basis in Y. By the Itô formula, Lemma 31, and (4.1), we have

$$\begin{cases} d\left((\boldsymbol{u}(t),\boldsymbol{\varphi})_2\zeta_\alpha(t)\right)=[\zeta_\alpha(t)\left\langle\mathcal{L}(\boldsymbol{u},t),\boldsymbol{\varphi}\right\rangle \\ \qquad\qquad +I_{\{\alpha\notin\mathcal{J}\}}\sum_{k\neq 0}m_i(t)\,\alpha_i^k\zeta_{\alpha(i,k)}(t)\left(\mathcal{M}^k(\boldsymbol{u},t),\boldsymbol{\varphi}\right)_2]dt \\ +\left[\zeta_\alpha(t)\left(\mathcal{M}^k(\boldsymbol{u},t),\boldsymbol{\varphi}\right)_2+I_{\{\alpha\neq\alpha^0\}}(\boldsymbol{u}(t),\boldsymbol{\varphi})_2 m_i(t)\,\alpha_i^k\zeta_{\alpha(i,k)}(t)\right]dw^k(t), \end{cases}$$

where

$$\left\langle\mathcal{L}(\boldsymbol{u}),\boldsymbol{\varphi}\right\rangle:=[-\left(a^{ij}\partial_j\boldsymbol{u},\partial_i\boldsymbol{\varphi}\right)_2+(b^i\partial_i\boldsymbol{u}-u^k\partial_k\boldsymbol{u}+\boldsymbol{f}+\left(\boldsymbol{h}^i\cdot\mathcal{G}^i(\mathcal{M}(\boldsymbol{u})),\boldsymbol{\varphi}\right)_2].$$

Taking the expectations of both sides of the equation and using Corollary 32, we arrive at

(4.15)
$$\partial_t\hat{\boldsymbol{u}}_\alpha(t)=E\left[\zeta_\alpha(t)\left\langle\mathcal{L}(\boldsymbol{u},t),\boldsymbol{\varphi}\right\rangle\right]+I_{\{\alpha\notin\mathcal{J}_0\}}\sum_{k\neq 0}m_i(t)\,\alpha_i^k E\left[\zeta_{\alpha(i,k)}(t)\left(\mathcal{M}^k(\boldsymbol{u},t),\boldsymbol{\varphi}\right)_2\right].$$

Now we shall express $E\left[\zeta_\alpha(t)\left\langle\mathcal{L}(\boldsymbol{u},t),\boldsymbol{\varphi}\right\rangle\right]$ and $E\left[\zeta_{\alpha(i,k)}(t)\left(\mathcal{M}^k(\boldsymbol{u},t),\boldsymbol{\varphi}\right)_2\right]$ in terms of Hermite–Fourier coefficients of $\boldsymbol{u}$, $\boldsymbol{f}$, and $\boldsymbol{g}$.

Write $\langle\tilde{\mathcal{L}}_0(\boldsymbol{u},t),\boldsymbol{\varphi}\rangle=-(a^{ij}(t)\partial_j\boldsymbol{u}(t),\partial_i\boldsymbol{\varphi})_2+(b^i(t)\partial_i\boldsymbol{u}(t)+\boldsymbol{f}(t),\boldsymbol{\varphi})_2$. Obviously, $\langle\mathcal{L}_0(\boldsymbol{u},t),\boldsymbol{\varphi}\rangle=\langle\tilde{\mathcal{L}}_0(\boldsymbol{u},t),\boldsymbol{\varphi}\rangle+R(t,\boldsymbol{\varphi})$, where

$$R(t,\boldsymbol{\varphi})=\left(\boldsymbol{h}^i(t)\cdot\mathcal{G}(\sigma^j(t)\partial_j\boldsymbol{u}(t)+\boldsymbol{g}(t)),\boldsymbol{\varphi}\right)_2. \tag{4.16}$$

It is easily seen that

$$E\left[\zeta_\alpha(t)\left(\tilde{\mathcal{L}}_0(\boldsymbol{u},t),\boldsymbol{\varphi}\right)_2\right]=\left(\tilde{\mathcal{L}}_0(\hat{\boldsymbol{u}}_\alpha,t),\boldsymbol{\varphi}\right)_2, \tag{4.17}$$

and for $\alpha\notin\mathcal{J}_0$,

(4.18)
$$\sum_{k\neq 0}m_i(t)\,E\left[\zeta_{\alpha(i,k)}(t)\alpha_i^k\left(\mathcal{M}^k(\boldsymbol{u},t),\boldsymbol{\varphi}\right)_2\right]=\sum_{k\neq 0}m_i(t)\,\alpha_i^k\left(\mathcal{M}^k\left(\hat{\boldsymbol{u}}_{\alpha(i,k)},t\right),\boldsymbol{\varphi}\right)_2.$$

Let us consider now the term $\left(u^i(t)\,\partial_i\boldsymbol{u}(t),\boldsymbol{\varphi}\right)_2$. By the Schwartz inequality,

$$\begin{aligned} &E\int_0^t\int_{\boldsymbol{R}^2}\left|\zeta_\alpha u^i(s,x)\,\partial_i u^j(s,x)\,\boldsymbol{\varphi}^j(x)\right|dsdx \\ &\leq\left(\int_0^t\int_{\boldsymbol{R}^2}E\left|\zeta_\alpha u^i(s,x)\right|^2dsdx\right)^{1/2}\left(\int_0^t\int_{\boldsymbol{R}^2}E\left|\partial_i u^j(s,x)\,\boldsymbol{\varphi}^j(x)\right|^2dsdx\right)^{1/2}<\infty. \end{aligned}$$

Thus, by the Fubini theorem and Lemma 30, we have that

$$\begin{aligned} &E\left[\zeta_\alpha(t)\int_0^t\left(u^i(s)\,\partial_i\boldsymbol{u}(s),\boldsymbol{\varphi}\right)_2ds\right] \\ &=\sum_{\alpha,p\in\mathcal{J}}\sum_{0\leq\beta\leq\alpha}\frac{1}{p!}\binom{\alpha}{\beta}\left(\hat{u}^i_{p+\beta}(t),\partial_i\hat{\boldsymbol{u}}_{p+\alpha-\beta}(t),\boldsymbol{\varphi}\right)_2. \end{aligned} \tag{4.19}$$

It remains to evaluate

$$R(t) = E\left[\boldsymbol{\zeta}_\alpha(t)\left(\boldsymbol{h}^i(t)\cdot\mathcal{G}(\sigma^j(t)\partial_j\boldsymbol{u}(t)+\boldsymbol{g}(t)),\boldsymbol{\varphi}\right)_{\mathbb{L}_2}\right].$$

To this end, we need the following simple result.

LEMMA 33. *If $\boldsymbol{v}\in L^2(\Omega,\mathcal{F}_s,\boldsymbol{P};\mathbb{L}_2)$, then for all $\alpha\in\mathcal{J}$, $\widehat{(\mathcal{G}\boldsymbol{v})}_\alpha=\mathcal{G}\hat{\boldsymbol{v}}_\alpha$ and $\widehat{(\mathcal{S}\boldsymbol{v})}_\alpha=\mathcal{S}\hat{\boldsymbol{v}}_\alpha$.*

Proof. Since $\mathcal{S}(\boldsymbol{v})=\boldsymbol{v}-\mathcal{G}(\boldsymbol{v})$, it is sufficient to prove only the first equality. By Stein's theorem, $\mathcal{G}(\boldsymbol{v})$ is $\mathcal{F}_s$-measurable and $|\mathcal{G}\boldsymbol{v}|_{\mathbb{L}_2}\le C|\boldsymbol{v}|_{\mathbb{L}_2}$, which yields that $\mathcal{G}\boldsymbol{v}\in L^2(\Omega,\mathcal{F}_s,\boldsymbol{P};\mathbb{L}_2)$. Thus, by the Fubini theorem, we have

$$\widehat{(\mathcal{G}\boldsymbol{v})}_\alpha=E\left[\zeta_\alpha\nabla\int\Gamma_{x_i}(x-y)v^i(y)\,dy\right]=\nabla\int\Gamma_{x_i}(x-y)v^i_\alpha(y)\,dy=\mathcal{G}\hat{\boldsymbol{v}}_\alpha. \qquad \square$$

It follows immediately from the lemma that

$$R(t,\boldsymbol{\varphi})=\left(\boldsymbol{h}^i(t)\cdot\mathcal{G}(\sigma^j(t)\partial_j\hat{\boldsymbol{u}}_\alpha(t)+\boldsymbol{g}_\alpha(t)),\boldsymbol{\varphi}\right)_2.$$

This completes the proof of Theorem 29.

Now we shall derive another convenient representation for the term $\widehat{u^i\partial_i\boldsymbol{u}}_\alpha$.

For $\alpha,\beta\in\mathcal{J}$, define $|\alpha-\beta|=(|a_1-\beta_1|,|a_2-\beta_2|,\dots)$.

DEFINITION 34. *We say that a triple of multiindices (α,β,γ) is complete, written $(\alpha,\beta,\gamma)\in\mathcal{C}$, if all the entries of the multiindex $\alpha+\beta+\gamma$ are even numbers and $|\alpha-\beta|\le\gamma\le\alpha+\beta$.*

Obviously, if (α,β,γ) is complete, then we also have that $|\alpha-\gamma|\le\beta\le\alpha+\gamma, |\gamma-\beta|\le\alpha\le\gamma+\beta$, and $\alpha+\beta-\gamma,\alpha-\beta+\gamma$, and $\beta+\gamma-\alpha$ are even multiindices.

It is readily checked that the following criterion holds.

LEMMA 35. *A triple (α,β,γ) is complete if and only if $\alpha+\beta+\gamma=2p$ for some $p\in\mathcal{J}$ and $p\le\alpha\wedge\beta$.*

For $(\alpha,\beta,\gamma)\in\mathcal{C}$, we define

$$\Phi(\alpha,\beta,\gamma)=\left(\left(\frac{\alpha-\beta+\gamma}{2}\right)!\left(\frac{\beta-\alpha+\gamma}{2}\right)!\left(\frac{\alpha+\beta-\gamma}{2}\right)!\right)^{-1}.$$

Obviously $\Phi(\alpha,\beta,\gamma)$ is invariant with respect to permutations of the arguments.

For $\alpha\in\mathcal{J}$, write $U^\alpha=\{\gamma,\beta\in\mathcal{J}:(\alpha,\beta,\gamma)\in\mathcal{C}\}$.

By Lemma 30,

$$E u^i\partial_i u^j\zeta_\alpha=\sum_{\boldsymbol{\gamma},\boldsymbol{\beta}\in\mathcal{J}}\hat{u}^i_\gamma\partial_i\hat{u}^j_\beta\sum_{p\le\gamma\wedge\beta}\left((\gamma-p)!\,(\beta-p)!p!\right)^{-1}\alpha!I_{(\alpha=\gamma+\beta-2p)}. \tag{4.20}$$

Since $\gamma+\beta-\alpha=2p$, $\gamma+\beta+\alpha$ is also an even multi-index. Also, the inequality $p\le\gamma\wedge\beta$ implies $|\gamma-\beta|\le\alpha\le\gamma+\beta$. Thus (γ,β,α) is complete. Now, it follows from (4.20) that

$$\widehat{(u^i\partial_i\boldsymbol{u})}_\alpha=\sum_{\boldsymbol{\gamma},\boldsymbol{\beta}\in U^\alpha}\alpha!\hat{u}^i_\gamma\partial_i\hat{\boldsymbol{u}}_\beta\Phi(\alpha,\beta,\gamma). \tag{4.21}$$

The propagators for advection type equations were studied in [33], [42], [45] (see also the references therein). Applications of Wiener chaos expansions to fluid mechanics have been sporadically discussed in the literature since the 1960s. For example, the inertial range spectrum of low order Wiener chaos truncations of a (random) Burgers equation were discussed in [10], [9], [48], [46]. There also exists a body of engineering literature on numerical aspects of Wiener chaos approximations (see, e.g., [32], [23], and the references therein).

4.3. Moments. Making use of the Wiener chaos expansion [6] for a solution of the SNS (4.1), one can immediately compute the first two moments of the solution via the Hermite–Fourier coefficients given by (4.7) for the propagator. Indeed, let us assume that the assumptions of Theorem 29 hold. It was shown in the previous section that the Hermite–Fourier coefficients $\hat{\boldsymbol{u}}_\alpha(t)$ are $\mathbb{H}^1_p(\boldsymbol{R}^2)\bigcap\mathbb{H}^1_2(\boldsymbol{R}^2)$-valued uniformly continuous functions of t. Owing to the imbedding $H_{1,p}\left(\boldsymbol{R}^2\right)\subset C^{1-2/p}\left(\boldsymbol{R}^2\right)$, we can interpret the Hermite–Fourier coefficients $\hat{\boldsymbol{u}}_\alpha(t,x)$ as continuous real functions on $\boldsymbol{R}^2\times[0,T]$.

Since $E\zeta_\alpha=0$ for $\alpha\neq 0$ and $E\zeta_{\mathbf{0}}=1$, where $\mathbf{0}$ is the multiindex $\alpha\in\mathcal{J}$ such that $|\alpha|=0$, we have that for all t,x,

$$\boldsymbol{E}\boldsymbol{u}(t,\boldsymbol{x})=\hat{\boldsymbol{u}}_{\mathbf{0}}(t,\boldsymbol{x}).$$

By [6] and Parceval's identity, one has that for all $\boldsymbol{x},\boldsymbol{y}\in\boldsymbol{R}^d$ and $t,s\in[0,T]$,

$$E\left(\boldsymbol{u}(t,\boldsymbol{x}),\boldsymbol{u}(s,\boldsymbol{y})\right)=\sum_{\alpha\in\mathcal{J}}\frac{1}{\alpha!}\left(\hat{\boldsymbol{u}}_\alpha(t,\boldsymbol{x}),\hat{\boldsymbol{u}}_\alpha(s,\boldsymbol{y})\right).$$

Similarly, given the solution of (4.7), the higher order moments of the solution to SNS equation (4.1) can be obtained by computing the moments of the Wick polynomials ζ_α.

Below we will derive some convenient formulas for these moments.

Let us consider the triple product $\zeta_\alpha\zeta_\beta\zeta_\gamma$. By (4.12)

$$\zeta_\alpha\zeta_\beta\zeta_\gamma=\sum_{p\leq\alpha\wedge\beta}\binom{\alpha}{p}\binom{\beta}{p}p!\zeta_{\alpha+\beta-2p}\zeta_\gamma. \tag{4.22}$$

It is readily checked that if f is a function on $\mathcal{J}$, then for $\alpha,\beta\in\mathcal{J}$,

$$\sum_{p\leq\alpha\wedge\beta}f(\alpha+\beta-2p)\binom{\alpha}{p}\binom{\beta}{p}p!=\alpha!\beta!\sum_{r\in U^{\alpha,\beta}}f(r)\,\Phi(\alpha,\beta,r). \tag{4.23}$$

Therefore, from (4.22), (4.23), and (4.2), it follows that

$$\begin{aligned}E\zeta_\alpha\zeta_\beta\zeta_\gamma&=\alpha!\beta!\sum_{r\in U^{\alpha,\beta}}\Phi(\alpha,\beta,r)\,E\zeta_r\zeta_\gamma\\&=\alpha!\beta!\gamma!\sum_{r\in U^{\alpha,\beta}}\Phi(\alpha,\beta,r)\,I_{(r=\gamma)}=\alpha!\beta!\gamma!\Phi(\alpha,\beta,\gamma)\,I_{\{(\alpha,\beta,\gamma)\in\mathcal{C}\}}.\end{aligned} \tag{4.24}$$

By induction, it is easy to verify that

$$\begin{aligned}E\Pi_{i=1}^{m+1}\zeta_{\alpha^i}&=\Pi_{i=0}^{m-3}r^i!\alpha^{m-i}!\sum_{r^{i+1}\in U(\alpha^{m-i},r^i)}\Phi\left(\alpha^{m-i},r^i,r^{i+1}\right)\\&\quad\times r^{m-2}!\alpha^2!\alpha^1!\Phi\left(\alpha^2,r^{m-2},\alpha^1\right)I_{\{(\alpha^2,r^{m-2},\alpha^1)\in\mathcal{C}\}},\end{aligned} \tag{4.25}$$

where $r^0=\alpha^{m+1}$(cf. [34, Thm. 5.3]).

For example,

$$E\zeta_\alpha\zeta_\beta\zeta_\gamma\zeta_\kappa=\alpha!\beta!\gamma!\kappa!\sum_{r\in U(\alpha,\beta)}\Phi(\alpha,\beta,r)\,r!\Phi(r,\gamma,\kappa)\,I_{\{(r,\gamma,\kappa)\in\mathcal{C}\}}.$$

Formula (4.25) allows us to compute pseudomoments of orders higher then 2. Let v be an $\mathcal{F}_T$-measurable random variable and $Ev^3 < \infty$.

Obviously, the set $\{\zeta_\alpha, \alpha \in J\}$ is total in all $L_p(\Omega)$. Given $v \in L_p(\Omega)$, there is a sequence of finite linear combinations $v^m = \sum_\alpha c_\alpha^m \xi_\alpha$ so that $E|v - v^m|^p \to 0$ as $m \to \infty$. If $p = 3$, then, of course,

$$E(v^m)^3 = \sum_{(\alpha,\beta,\gamma)\in\mathcal{C}} \hat{v}_\alpha^m \hat{v}_\beta^m \hat{v}_\gamma^m \Phi(\alpha,\beta,\gamma) \to Ev^3.$$

It is readily checked that $\hat{v}_\alpha^m \to \hat{v}_\alpha$ for all α. Therefore, we may define the third pseudomoment $\langle v^3 \rangle$ by the formula

$$\langle v^3 \rangle = \sum_{(\alpha,\beta,\gamma)\in\mathcal{C}} \hat{v}_\alpha \hat{v}_\beta \hat{v}_\gamma \Phi(\alpha,\beta,\gamma).$$

Formula

$$\langle v^4 \rangle = \sum_{\alpha,\beta,\gamma,\kappa} \hat{v}_\alpha \hat{v}_\beta \hat{v}_\gamma \hat{v}_\kappa \sum_{r\in U(\alpha,\beta)} \Phi(\alpha,\beta,r)\, r!\Phi(r,\gamma,\kappa)\, I_{\{(r,\gamma,\kappa)\in\mathcal{C}\}} \tag{4.26}$$

as well as similar formulas for higher pseudomoments could be proved by similar arguments.

Of course, the pseudomoments $\langle v^p \rangle$ coincide with the respective moments for $p = 1, 2$. However, if $p > 2$, the relation between the moments and the related pseudomoments is an open problem.

5. Appendix.

5.1. Nonnegative semimartingales. We will need also some estimates of nonnegative semimartingales.

LEMMA 36. *Let Z_t be a nonnegative semimartingale such that*

$$Z_t = Z_0 + \int_0^t a_s\, ds + \int_0^t b_s \cdot dW_s.$$

Assume that there are nonnegative measurable functions c_s, f_s, g_s, and a number $\delta \geq 0$ such that for any ε,

$$a_s \leq (-\delta + \varepsilon\delta)c_s + C_\varepsilon(Z_s + f_s),\ |b_s|_Y \leq \delta\varepsilon(c_s Z_s)^{1/2} + C_\varepsilon(Z_s + Z_s^{1-1/p} g_s^{1/p}),$$

where C_ε is a constant depending on ε.

Then for every $T > 0$, there is a constant $C = C(T)$ such that for all stopping times $\tau \leq T$

$$E\left[\sup_{s\leq\tau} |Z_s| + (\delta/2)\int_0^\tau c_s\, ds\right] \leq CE\left[Z_0 + \int_0^\tau (f_s + g_s)\, ds\right].$$

Proof. Let $s < t \leq T, t - s \leq 1$, and $\tilde{\tau}$ be a stopping time such that $\sup_{s\leq\tilde{\tau}} Z_s$ is bounded and

$$E\int_0^{\tilde{\tau}} (f_s + g_s)\, ds < \infty.$$

Fix an arbitrary stopping time τ. Let $\bar\tau = \tilde\tau \wedge \tau$. Then by Burkholder's inequality

$$E\left[\sup_{s\le r\le t} Z_{r\wedge\bar\tau} + \delta\int_{s\wedge\bar\tau}^{t\wedge\bar\tau} c_r\,dr\right] \le EZ_{s\wedge\bar\tau} + E\Bigg[\delta\varepsilon\int_{s\wedge\bar\tau}^{t\wedge\bar\tau} c_r\,dr + C_{\varepsilon,\delta}\sup_{s\le r\le t} Z_{r\wedge\bar\tau}(t-s)$$
$$+ C_{\varepsilon,\delta}\int_{s\wedge\bar\tau}^{t\wedge\bar\tau} f_r\,dr + N\left(\int_{s\wedge\bar\tau}^{t\wedge\bar\tau}(\varepsilon^2\delta^2 Z_r c_r + C^2_{\varepsilon,\delta}Z_r^2 + C^2_{\varepsilon,\delta}Z_r^{2(1-1/p)}g_r^{2/p})\,dr\right)^{1/2}\Bigg].$$

Obviously, for every $\varepsilon > 0$, there is a constant C_ε independent of T such that

$$\left(\int_{s\wedge\bar\tau}^{t\wedge\bar\tau} Z_r^{2(1-1/p)}g_s^{2/p}\,ds\right)^{1/2} \le \sup_{s\le r\le t} Z_{r\wedge\bar\tau}^{1-1/p}\left(\int_{s\wedge\bar\tau}^{t\wedge\bar\tau} g_s^{2/p}\,ds\right)^{1/2}$$
$$\le \varepsilon\sup_{s\le r\le t} Z_{r\wedge\bar\tau} + C_\varepsilon\left(\int_{s\wedge\bar\tau}^{t\wedge\bar\tau} g_s^{2/p}\,ds\right)^{p/2}.$$

Hence,

$$E\left[\sup_{s\le r\le t} Z_{r\wedge\bar\tau} + \delta\int_{s\wedge\bar\tau}^{t\wedge\bar\tau} c_r\,dr\right] \le EZ_{s\wedge\bar\tau} + E\Bigg\{\delta\varepsilon\int_{s\wedge\bar\tau}^{t\wedge\bar\tau} c_r\,dr$$
$$+ (N+1)\,C_{\varepsilon,\delta}\sup_{s\le r\le t} Z_{r\wedge\bar\tau}(t-s)^{1/2} + \tilde C_{\varepsilon,\delta}\left[\int_{s\wedge\bar\tau}^{t\wedge\bar\tau} f_r\,dr + \left(\int_{s\wedge\bar\tau}^{t\wedge\bar\tau} g_r^{2/p}\,dr\right)^{p/2}\right]$$
$$+ \varepsilon\sup_{s\le r\le t} Z_{r\wedge\bar\tau} + 2^{-1}N\varepsilon\delta\sup_{s\le r\le t} Z_{r\wedge\bar\tau} + 2^{-1}N\varepsilon\delta\int_{s\wedge\bar\tau}^{t\wedge\bar\tau} c_r\,dr\Bigg\}.$$

Let us take ε so that

$$\left(1-\varepsilon\left(1+2^{-1}N\delta\right)\right)/\left(1-\varepsilon\left(1+2^{-1}N\delta\right)\right) = 1/4.$$

Then, there is a constant $C = C(T)$ such that

$$E\sup_{s\le r\le t} Z_{r\wedge\bar\tau} + (1/4)\delta\int_s^t c_r\,dr \le CE\Bigg[Z_{s\wedge\bar\tau} + \sup_{s\le r\le t} Z_{r\wedge\bar\tau}(t-s)^{1/2})$$
$$+ \int_{s\wedge\bar\tau}^{t\wedge\bar\tau} f_r\,dr + \left(\int_{s\wedge\bar\tau}^{t\wedge\bar\tau} g_r^{2/p}\,dr\right)^{p/2}\Bigg].$$

By choosing $\kappa = t - s$ small enough, we obtain

$$E\sup_{s\le r\le t} Z_{r\wedge\bar\tau} + (1/2)\delta\int_{s\wedge\bar\tau}^{t\wedge\bar\tau} c_r\,dr \le CE\left[Z_{s\wedge\bar\tau} + \int_{s\wedge\bar\tau}^{t\wedge\bar\tau}(f_s+g_s)\,ds\right]. \tag{5.1}$$

To prove a similar estimate for $s = 0$ and an arbitrary $t \le T$, we apply estimate (5.1) successively on the intervals, $[0,\kappa], [\kappa, 2\kappa], \ldots$. Now, the statement follows. □

5.2. Convergence lemma. Let B be a Banach space with a norm $|\cdot|_B$. Let X_n be a sequence of B-valued continuous processes defined on $[0,\zeta_n)$, where $\zeta_n = \zeta_n(X_n)$ is such that $\boldsymbol{P}$-a.s. $\zeta_n > 0$ and

$$\limsup_{t\uparrow\zeta_n} |X_n(t)|_B = \infty$$

on $\{\zeta_n < \infty\}$. For $M > 0, T > 0, n, n'$, let $\mathcal{T}_n^{M,T}$ be the set of all stopping times $\tau \le T$ such that $\sup_{s\le\tau} |X_n(s)|_B \le M + |X_n(0)|_B$, $\mathcal{T}_{n,n'}^{M,T} = \mathcal{T}_n^{M,T} \cap \mathcal{T}_{n'}^{M,T}$.

LEMMA 37. (a) *Let $\boldsymbol{P}$-a.s. $\zeta_n = \infty$ for all n. Assume that for each M, T*

$$\lim_{n\to\infty} \sup_{n'\ge n, \tau\in\mathcal{T}_{n,n'}^{M,T}} E \sup_{s\le\tau} |X_n(s) - X_{n'}(s)|_B = 0,$$

$$\sup_n E \sup_{s\le T} |X_n(s)|_B < \infty.$$

Then there is a B-valued continuous process X and a subsequence n_k such that $\boldsymbol{P}$-a.s. for each T

$$\sup_{s\le T} |X_{n_k}(s) - X(s)|_B \to 0.$$

(b) *Assume that for some $M_0 > 1, T_0 > 0$*

$$\lim_{n\to\infty} \sup_{n'\ge n, \tau\in\mathcal{T}_{n,n'}^{M_0,T_0}} E \sup_{s\le\tau} |X_n(s) - X_{n'}(s)|_B = 0,$$

and

$$\lim_{T\to 0} \sup_{n, \tau\in\mathcal{T}_n^{M_0,T_0}} \boldsymbol{P}\Big(\sup_{s\le\tau\wedge T} |X_n(s)|_B > |X_n(0)|_B + M_0 - 1\Big) = 0.$$

Then there is a bounded stopping time τ such that $\boldsymbol{P}(\tau > 0) = 1$ and a B-valued continuous process X on $[0, \tau]$ and a subsequence n_k such that $\boldsymbol{P}$-a.s.

$$\sup_{s\le\tau} |X_{n_k}(s) - X(s)|_B \to 0.$$

Moreover, if $\sup_n E|X_n(0)|_B^p < \infty, p \ge 1$, then $E \sup_{t\le\tau} |X(t)|_B^p < \infty$.

Proof. (a) Obviously, there are $n_k \uparrow \infty, T_k \uparrow \infty, M_k \uparrow \infty$ such that

$$\sup_{n'\ge n_k} E \sup_{s\le\tau\in\mathcal{T}_{n,n'}^k} |X_{n_k}(s) - X_{n'}(s)|_B \le 2^{-2k},$$

where $\mathcal{T}_{n,n'}^k = \mathcal{T}_{n,n'}^{M_k,T_k}$. Fix $T > 0$ and $M > 0$. Let $\tau_k = \inf(t : |X_{n_k}(t)|_B > |X_{n_k}(0)|_B + M + 2^{-k}) \wedge T$. Then, for k so that $M_k > M + 2^{-k}$, $T_k > T$, we have

$$\boldsymbol{P}\Big(\sup_{s\le\tau_k\wedge\tau_{k+1}} |X_{n_k}(s) - X_{n_{k+1}}(s)|_B > 2^{-k}/4\Big)$$

$$\le 4\cdot 2^k E \sup_{s\le\tau_k\wedge\tau_{k+1}} |X_{n_k}(s) - X_{n_{k+1}}(s)|_B \le 4\cdot 2^{-k}.$$

By the Borelli–Cantelli lemma,

$$\sup_{s\le\tau_k\wedge\tau_{k+1}} |X_{n_k}(s) - X_{n_{k+1}}(s)|_B \le 2^{-k}/4$$

or $\tau_{k+1} \le \tau_k$ for sufficiently large k. Let $\tau = \lim_k \tau_k$. Then $X_{n_k}(t)$ converges to some process $X(t)$ on $[0, \tau]$. Also,

$$\boldsymbol{P}(\tau < T) = \lim \boldsymbol{P}(\tau_k < T) \le M^{-1} E \sup_{s\le T} |X_{n_k}(s)|_B.$$

(b) Obviously, there is $n_k \uparrow \infty$ such that

$$\sup_{n' \geq n_k} E \sup_{s \leq \tau \in \mathcal{T}_{n,n'}} |X_{n_k}(s) - X_{n'}(s)|_B \leq 2^{-2k},$$

where $\mathcal{T}_{n,n'} = \mathcal{T}_{n,n'}^{M_0,T_0}$. Let $\tau_k = \inf(t : |X_{n_k}(t)|_B > |X_{n_k}(0)|_B + M_0 - 1 + 2^{-k}) \wedge T_0$. Then

$$\boldsymbol{P}(\sup_{s \leq \tau_k \wedge \tau_{k+1}} |X_{n_k}(s) - X_{n_{k+1}}(s)|_B > 2^{-k}/4)$$
$$\leq 4 \cdot 2^k E \sup_{s \leq \tau_k \wedge \tau_{k+1}} |X_{n_k}(s) - X_{n_{k+1}}(s)|_B \leq 4 \cdot 2^{-k}.$$

By the Borelli–Cantelli lemma,

$$\sup_{s \leq \tau_k \wedge \tau_{k+1}} |X_{n_k}(s) - X_{n_{k+1}}(s)|_B \leq 2^{-k}/4$$

or $\tau_{k+1} \leq \tau_k$ for sufficiently large k. Let $\tau = \lim_k \tau_k$. Then $X_{n_k}(s)$ converges to some process $X(s)$ on $[0, \tau]$. Also,

$$\boldsymbol{P}(\tau < \varepsilon) = \lim \boldsymbol{P}(\tau_k < \varepsilon) \leq \limsup_k \boldsymbol{P}(\sup_{s \leq \tau_k \wedge \varepsilon} |X_{n_k}(s)|_B > |X_{n_k}(0)|_B + M_0 - 1) \to 0,$$

as $\varepsilon \to 0$, i.e., $\boldsymbol{P}(\tau = 0) = 0$.

Since

$$\sup_k E \sup_{t \leq \tau} |X_{n_k}(t)|_B^p < \infty,$$

if $\sup_n \boldsymbol{E}|X_n(0)|_B^p < \infty$, the last assertion follows by the Fatou lemma. □

5.3. Estimates of gradient projection. In this section, for the sake of convenience, we summarize some basic estimates for gradient projections proved in [37].

LEMMA 38 (see Lemma 2.13 in [37]). *Assume $\boldsymbol{v} \in \mathbb{H}_p^{s+1}(Y), p \in (1, \infty)$. Then*

$$\mathcal{G}(\partial_l \boldsymbol{v}) = \partial_l \mathcal{G}(\boldsymbol{v}) = -(1-\Delta)^{-s/2} \boldsymbol{R} R_l ((1-\Delta)^{s/2} \operatorname{div} \boldsymbol{v}). \tag{5.2}$$

There is a constant C such that for all $\boldsymbol{v} \in \mathbb{H}_p^{s+1}(Y)$,

$$||\partial \mathcal{G}(\boldsymbol{v})||_{s,p} \leq C ||\operatorname{div} \boldsymbol{v}||_{s,p},$$

and for all $\boldsymbol{v} \in \mathbb{H}_p^s(Y)$,

$$||\mathcal{G}(\boldsymbol{v})||_{s,p} \leq C ||\operatorname{div} \boldsymbol{v}||_{s-1,p} + ||\boldsymbol{v}||_{s-1,p}.$$

We need L_p-estimates of the function $\mathcal{G}(\boldsymbol{h})$, where $\boldsymbol{h} = c^j \partial_j \boldsymbol{v}$.

LEMMA 39 (see Lemma 2.14 in [37]). *Let $\boldsymbol{h} = c^j(x) \partial_j \boldsymbol{v}(x)$, where $c = (c^j)$ is a measurable d-vector of Hilbert space Y-valued functions, $\boldsymbol{v} \in \mathbb{H}_p^{s+1}$, $\operatorname{div} \boldsymbol{v} = 0$, $\varepsilon \in (0,1)$. Assume*

$$||c||_{B^{|s|}} < \infty \text{ if } s \geq 1,$$
$$||c||_{B^1} < \infty \text{ if } s \in (-1, 1),$$
$$||c||_{B^{-s+\varepsilon}} < \infty \text{ if } s \leq -1.$$

Then

$$||\mathcal{G}(\boldsymbol{h})||_{s,p} \le \begin{cases} C(||\partial_l c^j \boldsymbol{\partial}_j \boldsymbol{v}||_{s-1,p} + ||c^j \boldsymbol{\partial}_j \boldsymbol{v}||_{s-1,p}) & \text{if } s > 0, \\ C(||\partial_l c^j v^l||_{s,p} + ||(\operatorname{div} c)\, \boldsymbol{v}||_{s,p}] & \text{if } s \le 0. \end{cases}$$

Also, we need L_p-estimates of the function $\mathcal{G}(\boldsymbol{h})$, where $\boldsymbol{h} = \boldsymbol{\partial}_i(c^{ij}(x)\partial_j \boldsymbol{v})$.

COROLLARY 40 (see Corollary 2.15 in [37]). *Let* $\boldsymbol{h} = \boldsymbol{\partial}_i(c^{ij}(x)\partial_j \boldsymbol{v})$, *where* $c = (c^{ij})$ *is a measurable function,* $\boldsymbol{v} \in \mathbb{H}_p^{s+1}$, $\operatorname{div} \boldsymbol{v} = 0$, $\varepsilon \in (0,1)$. *Assume*

$$\begin{aligned} |c|_{B^{|s|}} &< \infty \text{ if } s \ge 1, \\ |c|_{B^1} &< \infty \text{ if } s \in (-1,1), \\ |c|_{B^{-s+\varepsilon}} &< \infty \text{ if } s \le -1. \end{aligned}$$

Then

$$|\mathcal{G}(\boldsymbol{h})|_{s-1,p} \le \begin{cases} C(|\partial_l c^{ij} \boldsymbol{\partial}_j v^l|_{s-1,p} + |c^{ij} \boldsymbol{\partial}_j v^l|_{s-1,p}) & \text{if } s > 0, \\ C(|\partial_l c^{ij} v^j|_{s,p} + |\partial_j c^{ij}\, \boldsymbol{v}|_{s,p}] & \text{if } s \le 0. \end{cases}$$

5.4. Biot–Savart law in $\mathbf{R}^d$. The Biot–Savart law is usually discussed only in dimensions $d = 2, 3$. In this subsection we introduce a slightly more general construction for any d.

DEFINITION 41. (a) *Given two vectors* $\boldsymbol{a} = (a^1, \dots, a^d), \boldsymbol{b} = (b^1, \dots, b^d)$ *in* $\boldsymbol{R}^d$, *we define their product*

$$\boldsymbol{a} \times \boldsymbol{b} = \left(\varepsilon_{lk}(a^k b^l - a^l b^k)\right)_{1 \le k < l \le d} \in \boldsymbol{R}^{d(d-1)/2},$$

where $\varepsilon_{lk} = (-1)^{l+k-1}$.

(In standard notation, we could also use a matrix $\sqrt{2}\boldsymbol{a} \wedge \boldsymbol{b} = \left(a^k b^l - a^l b^k\right)_{1 \le k,l \le d}$.*)*

(b) *Given a vector field* $\boldsymbol{v}(x) = (v^1(x), \dots, v^d(x))$, *we define a new vector field*

$$curl\, \boldsymbol{v} = \boldsymbol{\nabla} \times \boldsymbol{v} = \left(\varepsilon_{lk}(\partial_k v^l - \partial_l v^k)\right)_{1 \le k < l \le d} \in \boldsymbol{R}^{d(d-1)/2}.$$

REMARK 9. (a) *Given a scalar function* a, *we have*

$$curl\,(a\boldsymbol{v}) = a\, curl\,(\boldsymbol{v}) + (\nabla a) \times \boldsymbol{v}. \tag{5.3}$$

(b) *If* $\boldsymbol{v} = \boldsymbol{\nabla} p$ *(*p *is a scalar function), then* $curl\, \boldsymbol{v} = \boldsymbol{\nabla} \times \boldsymbol{v} = \boldsymbol{0}$.

PROPOSITION 42. *For each* $\boldsymbol{v} \in \mathbb{H}_p^1$

$$\partial_m \mathcal{S}(\boldsymbol{v}) = -\sum_j R_m R_j (\partial_j \boldsymbol{v} - \boldsymbol{\nabla} v^j).$$

There is a constant C *so that for all* $\boldsymbol{v} \in \mathbb{H}_p^1$

$$|\partial \mathcal{S}(\boldsymbol{v})|_p \le C |curl\, \boldsymbol{v}|_p.$$

In general, there is a constant C *so that for all* $\boldsymbol{v} \in \mathbb{H}_p^{s+1}$,

$$\begin{aligned} |\partial \mathcal{S}(\boldsymbol{v})|_{s,p} = |\mathcal{S}(\partial \boldsymbol{v})|_{s,p} &\le C |curl\, \boldsymbol{v}|_{s,p}, \\ |\mathcal{S}(\boldsymbol{v})|_{s+1,p} &\le C(|curl\, \boldsymbol{v}|_{s,p} + |\boldsymbol{v}|_{s,p}), \end{aligned}$$

or for all $\boldsymbol{v} \in \mathbb{H}^s_p$

$$|\mathcal{S}(\boldsymbol{v})|_{s,p} \le C(|curl\,\boldsymbol{v}|_{s-1,p} + |\boldsymbol{v}|_{s-1,p}).$$

Proof. Indeed, considering Fourier transforms, we easily find that for all $\boldsymbol{v} \in \mathbb{C}^\infty_0$,

$$\mathcal{S}(\boldsymbol{v}) = \boldsymbol{v} - \mathcal{G}(\boldsymbol{v}). = -\sum_j (R_j R_j \boldsymbol{v} - R_j R v^j) = -R(R \wedge \boldsymbol{v}),$$

$$\partial_m \mathcal{S}(\boldsymbol{v}) = -\sum_j R_m R_j (\partial_j \boldsymbol{v} - \boldsymbol{\nabla} v^j).$$

Also,

$$\begin{aligned}\partial_m J_s \mathcal{S}(\boldsymbol{v}) &= -\sum_j R_m R_j (\partial_j J_s \boldsymbol{v} - \boldsymbol{\nabla} J_s v^j)\\ &= -\sum_j R_m R_j J_s (\partial_j \boldsymbol{v} - \boldsymbol{\nabla} v^j),\end{aligned}$$

where $J_s = (1-\Delta)^{s/2}$. Since the Riesz transform is bounded in L_p and

$$|\boldsymbol{v}|_{s,p} + |\partial \boldsymbol{v}|_{s,p} \sim |\boldsymbol{v}|_{s+1,p},$$

the statement follows. □

REFERENCES

[1] V. Arnold, *Sur la geometrie differentielle des groupes de Lie de dimension infinie et ses applications a l'hydrodynamic des fluides parfaits*, Ann. Inst. Grenoble, 16 (1966), pp. 319–361.

[2] P. Baxendale and T. E. Harris, *Isotropic stochastic flows*, Ann. Probab., 14 (1986), pp. 1155–1179.

[3] A. Bensoussan and R. Temam, *Equatios stochastique du type Navier-Stokes*, J. Funct. Anal., 13 (1973), pp. 195–222.

[4] Z. Brzeźniak, M. Capiński, and F. Flandoli, *Stochastic partial differential equations and turbulence*, Math. Models Methods Appl. Sci., 1 (1991), pp. 41–59.

[5] Z. Brzezniak and S. Peszat, *Strong local and global solutions to stochastic Navier-Stokes equations*, in Infinite Dimensional Stochastic Analysis, Proceedings of the Coloquium of the Royal Netherlands Academy of Sciences, Amsterdam, 1999, North–Holland, Amsterdam, 2000, pp. 85–98.

[6] R. H. Cameron and W. T. Martin, *The orthogonal development of non-linear functionals in a series of Fourier-Hermite functionals*, Ann. of Math. (2), 48 (1947), pp. 385–392.

[7] M. Capiński and N. J. Cutland, *Stochastic Navier-Stokes equations*, Acta Appl. Math., 25 (1991), pp. 59–85.

[8] A. J. Chorin and J. Marsden, *A mathematical introduction to fluid mechanics*, Springer-Verlag, New York, 1990.

[9] A. J. Chorin, *Lectures on Turbulence Theory*, Berkeley Math. Lect. Notes 2, Berkeley, CA, 1993.

[10] S. C. Crow and G. H. Canavan, *Relationship between a Wiener-Hermite expansion and an energy cascade*, J. Fluid Mech., 41 (1970), pp. 387–403.

[11] G. DaPrato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, UK, 1992.

[12] H. B. da Veiga, *Existence and asymptotic behavior for the strong solutions of the Navier-Stokes equations in the whole space*, Indiana Univ. Math. J., 36 (1987), pp. 149–166.

[13] W. E. and Y. Sinai, *Recent results on mathematical & statistical hydrodynamics*, Russian Math. Surveys, 55 (2000), pp. 25–57.

[14] D. G. Ebin and J. Marsden, *Groups of diffeomorphisms and the notion of an incompressible fluid*, Ann. of Math. (2), 92 (1970), pp. 102–163.

[15] F. Flandoli and D. Gatarek, *Martingale and stationary solutions for stochastic Navier-Stokes equations*, Probab. Theory Related Fields, 102 (1995), pp. 367–391.
[16] K. Gawedzki and A. Kupiainen, *Universality in Turbulence: An Exactly Solvable Model*, in Low-Dimensional Models in Statistical Physics and Quantum Field Theory, H. Grosse and L. Pittner, eds., Springer-Verlag, Berlin, 1996, pp. 71–105.
[17] K. Gawedzki and M. Vergassola, *Phase transition in the passive scalar advection*, Phys. D, 138 (2000), pp. 63–90.
[18] Y. Giga and T. Miyakawa, *Solutions in L_r of the Navier-Stokes initial value problem*, Arch. Ration. Mech. Anal., 89 (1985), pp. 251–265.
[19] T. Hida, H. H. Kuo, J. Potthoff, and L. Streit, *White Noise*, Kluwer Academic, Dordrecht, The Netherlands, 1993.
[20] H. Holden, B. Oksendal, J. Uboe, and T. Zhang, *Stochastic Partial Differential Equations. A Modeling, White Noise Functional Approach*, Birkhäuser Boston, Boston, 1996.
[21] A. Inoue and T. Funaki, *On a new derivation of the Navier-Stokes equation*, Comm. Math. Phys., 6 (1979), pp. 83–90.
[22] J. Jacod, *Calcul stochastic et problemes de martingales*, Lecture Notes in Math. 714, Springer-Verlag, New York, 1979.
[23] M. Jardak, C.-H. Su, and G. E. Karniadakis, *Spectral polynomial solutions of the stochastic advection equation*, J. Sci. Comput., 17 (2002), pp. 319–338.
[24] T. Kato and G. Ponce, *Well-posedness of the Euler and Navier-Stokes equations in the Lebesgue spaces $L^p_s(R^2)$*, Rev. Mat. Iberoamericana, 2 (1986), pp. 73–88.
[25] T. Kato and G. Ponce, *On nonstationary flows of viscous and ideal fluids in $L^p_s(R^2)$*, Duke Math. J., 55 (1987), pp. 487–489.
[26] R. H. Kraichnan, *Small-scale structure of a scalar field convected by turbulence*, Phys. Fluids, 11 (1968), pp. 945–963.
[27] N. V. Krylov, *On L_p-theory of stochastic partial differential equations in the whole space*, SIAM J. Math. Anal., 27 (1996), pp. 313–340.
[28] N. V. Krylov, *An analytic approach to SPDEs*, in Stochastic Partial Differential Equations: Six Perspectives, Math. Surveys Monogr. 64, R. Carmona and B. Rozovskii, eds., AMS, Providence, RI, 1999, pp. 185–242.
[29] H. Kunita, *Stochastic Flows and Stochastic Differential Equations*, Cambridge University Press, New York, 1990.
[30] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics*, Addison-Wesley, Reading, MA, 1959.
[31] Y. Le Jan and O. Raimond, *Integration of Brownian Vector Fields*, Ann. Probab., 30 (2002) pp. 826–873.
[32] O. P. Le Maître, O. M. Kino, H. N. Najim, and R. G. Ghanem, *A stochastic projection method for fluid flow*, J. Comput. Phys., 173 (2001), pp. 481–511.
[33] S. Lototsky, R. Mikulevicius, and B. L. Rozovskii, *Nonlinear filtering revisited: A spectral approach,* SIAM J. Control Optim., 35 (1997), pp. 435–461.
[34] P. Major, *Multiple Wiener-Ito Integrals*, 2nd ed., Lecture Notes in Math. 849, Springer-Verlag, Berlin, 1981.
[35] J. Mattingly, *The Stochastic Navier-Stokes Equation. Energy Estimates and Phase Space Construction*, Thesis, Princeton University, Princeton, NJ, 1998.
[36] P.-A. Meyer, *Quantum Probability for Probabilists*, 2nd ed., Lecture Notes in Math. 1538, Springer-Verlag, Berlin, 1995.
[37] R. Mikulevicius, *On the Cauchy problem for stochastic Stokes equations*, SIAM J. Math. Anal., 34 (2002), pp. 121–141.
[38] R. Mikulevicius and B. L. Rozovskii, *A note on Krylov's L_p-theory for systems of SPDEs*, Electron. J. Probab., 6 (2001), pp. 1–35.
[39] R. Mikulevicius and B. L. Rozovskii, *Stochastic Navier-Stokes equations. Propagation of chaos and statistical moments*, in Optimal Control and Partial Differential Equations, J. L. Menaldi, E. Rofmann, and A. Sulem, eds., IOS Press, Amsterdam, 2001, pp. 258–267.
[40] R. Mikulevicius and B. L. Rozovskii, *On equations of stochastic fluid mechanics*, in Stochastics in Finite and Infinite Dimensions, T. Hida, R. Karandikar, H. Kunita, B. Rajput, S. Watanabe, and J. Xiong, eds., Birkhäuser Boston, Boston, MA, 2001, pp. 285–302.
[41] R. Mikulevicius and B. L. Rozovskii, *Martingale problems for stochastic PDE's*, in Stochastic Partial Differential Equations: Six Perspectives, Math. Surveys Monogr. 64, R. Carmona and B. L. Rozovskii, eds., AMS, Providence, RI, 1998, pp. 243–325.
[42] R. Mikulevicius and B. L. Rozovskii, *Linear parabolic stochastic PDE's and Wiener chaos*, SIAM J. Math. Anal., 29 (1998), pp. 452–480.
[43] R. Mikulevicius and B. L. Rozovskii, *Global L_2-solutions of stochastic Navier-Stokes equations*, Ann. Probab., to appear.

[44] R. Mikulevicius and G. Valiukevicius, *On stochastic Euler equations in R^d*, Electron. J. Probab., 5 (2000), pp. 1–20.
[45] D. Nualart and B. L. Rozovskii, *Weighted Wiener chaos and linear SPDE's driven by a space-time white noise*, J. Funct. Anal., 149 (1997), pp. 200–225.
[46] S. A. Orszag and L. R. Bissonnette, *Dynamical properties of truncated Wiener-Hermite expansions*, Phys. Fluids, 10 (1967), pp. 2603–2613.
[47] B. L. Rozovskii, *Stochastic Evolution Systems*, Kluwer Academic, Dordrecht, Boston, 1990.
[48] A. Siegel, T. Imamura, and W. C. Meecham, *Wiener-Hermite expansion in model turbulence in the late decay stage*, J. Math. Phys., 6 (1965), pp. 707–721.
[49] E. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, 1970.
[50] H. Triebel, *Theory of Function Spaces*, Birkhäuser Verlag, Basel, 1983.
[51] M. Viot, *Solutions faibles d'equation aux derives partielles stochastiques non lineaires*, Thèse de doctorat, University of Paris VI, Paris, France, 1976.
[52] M. I. Vishik and A. V. Fursikov, *Mathematical Problems of Statistical Hydromechanics*, Kluwer Academic Press, Dordrecht, Boston, London, 1979.

SIAM J. MATH. ANAL.
Vol. 35, No. 5, pp. 1311–1346

FAMILIES OF PERIODIC ORBITS FOR THE SPATIAL ISOSCELES 3-BODY PROBLEM*

MONTSERRAT CORBERA† AND JAUME LLIBRE‡

Abstract. We study the families of periodic orbits of the spatial isosceles 3-body problem (for small enough values of the mass lying on the symmetry axis) coming via the analytic continuation method from periodic orbits of the circular Sitnikov problem. Using the first integral of the angular momentum, we reduce the dimension of the phase space of the problem by two units. Since periodic orbits of the reduced isosceles problem generate invariant two-dimensional tori of the nonreduced problem, the analytic continuation of periodic orbits of the (reduced) circular Sitnikov problem at this level becomes the continuation of invariant two-dimensional tori from the circular Sitnikov problem to the nonreduced isosceles problem, each one filled with periodic or quasi-periodic orbits. These tori are not KAM tori but just isotropic, since we are dealing with a three-degrees-of-freedom system. The continuation of periodic orbits is done in two different ways, the first going directly from the reduced circular Sitnikov problem to the reduced isosceles problem, and the second one using two steps: first we continue the periodic orbits from the reduced circular Sitnikov problem to the reduced elliptic Sitnikov problem, and then we continue those periodic orbits of the reduced elliptic Sitnikov problem to the reduced isosceles problem. The continuation in one or two steps produces different results. This work is merely analytic and uses the variational equations in order to apply Poincaré's continuation method.

Key words. periodic orbits, quasi-periodic orbits, 3-body problem, analytic continuation method

AMS subject classifications. 70F15, 37N05

DOI. 10.1137/S0036141002407880

1. Introduction.

We consider a special case of the spatial 3-body problem, the *spatial isosceles 3-body problem*, or simply the *isosceles problem*. This problem consists of describing the motion of two equally massive bodies, $m_1 = m_2 = 1/2$, having initial conditions and velocities symmetric with respect to a straight line which passes through their center of mass, and a third body, with mass $m_3 = \mu$, having initial position and velocity on this straight line. This problem is called the isosceles problem because the three bodies form an isosceles triangle at any time, eventually degenerated to a segment.

The most interesting application of the spatial isosceles 3-body problem was given by Xia in [25]. He used two spatial isosceles 3-body problems to prove that five bodies can escape to infinity in a finite time without collision. Other works on the spatial isosceles 3-body problem are [16] and the references therein. If in the spatial isosceles 3-body problem the initial positions and velocities of the three bodies are contained in a plane, then the motion remains always in this plane, and we have the so-called *planar isosceles 3-body problem*. There are several papers about the planar isosceles 3-body problem, for instance, [9], [17], etc.

When the third body of the isosceles 3-body problem has infinitesimal mass (i.e., $\mu = 0$) then we obtain the *restricted isosceles problems*. Depending on the motion

*Received by the editors May 17, 2002; accepted for publication (in revised form) May 30, 2003; published electronically January 30, 2004. This work was partially supported by MCYT grants BFM 2002–04236–C02–02 and by CIRIT grant SGR 2001 00173.

http://www.siam.org/journals/sima/35-5/40788.html

†Departament d'Informàtica i Matemàtica, Universitat de Vic, Laura 13, 08500 Vic, Barcelona, Spain (montserrat.corbera@uvic.es).

‡Departament de Matemàtiques, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain (jllibre@mat.uab.es).

of the primaries m_1 and m_2 we have seven different cases for the restricted isosceles problems. Here we consider, due to their richness in periodic orbits, only the cases in which the primaries move in circular or elliptic orbits of the 2-body problem, the *circular and elliptic* restricted isosceles problems, also called the *circular and elliptic Sitnikov problems*.

The isosceles problem and the restricted isosceles problems possess the first integral of the angular momentum. In section 3 we will prove that the phase portrait of any of these problems on each level of the angular momentum c with $c \neq 0$ is the same. Notice that the angular momentum $c = 0$ contains the triple and the double collision orbits, but collision orbits are not treated in this work. With a fixed value of the angular momentum $c \neq 0$, we reduce by two dimensions (an angle and its derivative) the phase space of the isosceles problem, obtaining the *reduced isosceles problem*. In particular, we see that each periodic orbit of the reduced isosceles problem gives an invariant two-dimensional torus of the isosceles problem, filled with either periodic or quasi-periodic orbits, which is not a KAM tori. We note that the circular and elliptic Sitnikov problems that appear in the literature are essentially our reduced circular and elliptic Sitnikov problems.

The main objective of this work is to prove that the invariant two-dimensional tori of the restricted isosceles problem that come from the known periodic orbits of the reduced circular Sitnikov problem persist when we pass from the restricted isosceles problem to the isosceles problem for $\mu > 0$ sufficiently small. Consequently these tori persist inside the general spatial 3-body problem. The main tool for proving this result will be the classical Poincaré analytic continuation method of periodic orbits. In particular, we continue the known periodic orbits of the reduced circular Sitnikov problem to periodic orbits of the reduced isosceles problem for $\mu > 0$ sufficiently small. In order to do that, we will use the symmetries of the problem. The isosceles problem is invariant under the time reversibility (t-symmetry), and it is also invariant under a symmetry with respect to the plane defined by the motion of m_1 and m_2 (r-symmetry). These symmetries will allow us to find r- and t-symmetric periodic orbits for the reduced isosceles problem. We still distinguish another type of symmetric periodic orbits, the *doubly symmetric periodic orbits*, which are simultaneously r- and t-symmetric periodic orbits.

Using the analytical continuation method of Poincaré, we will continue the known periodic orbits of the reduced circular Sitnikov problem (where $\mu = 0$), which are doubly symmetric periodic orbits, to symmetric periodic orbits of the reduced isosceles problem for $\mu > 0$ sufficiently small. Those periodic orbits are continued in two different ways. The first goes directly from the reduced circular Sitnikov problem to the reduced isosceles problem. The second uses two steps: first we continue the periodic orbits from the reduced circular Sitnikov problem to symmetric periodic orbits of the reduced elliptic Sitnikov problem (where $\mu = 0$) for small values of the eccentricity e, and then we continue those symmetric periodic orbits of the reduced elliptic Sitnikov problem to the reduced isosceles problem for small values of $\mu > 0$.

A key point in this work is the knowledge of an analytical expression for the solution of the variational equations of the reduced circular (elliptic) Sitnikov problem along the periodic solution that we want to continue. We must remark that all results presented in this paper are analytical results.

The main results about continuation of periodic orbits from the reduced circular Sitnikov problem to the reduced isosceles problem are summarized in the following result.

THEOREM A. *Let γ be a periodic orbit of the reduced circular Sitnikov problem with period $T > \pi/\sqrt{2}$, and let $f(e) = (1-e^2)^{3/2}$. Then γ can be continued to the following families of periodic orbits of the reduced isosceles problem with angular momentum $c = 1/4$ and $\mu > 0$ sufficiently small:*

1. *Case $T = 2\pi\omega$ with $\omega > 1/(2\sqrt{2})$ an irrational number.*
   (a) *γ can be continued directly to one 2-parameter family (on μ and τ) of doubly symmetric periodic orbits with period τ sufficiently close to T.*
2. *Case $T = 2\pi p/q$ for some $p, q \in \mathbb{N}$ coprime with $p > q/(2\sqrt{2})$.*
   (a) *p odd:*
      i. *γ can be continued directly to one 2-parameter family (on μ and τ) of doubly symmetric periodic orbits with period τ sufficiently close to T.*
      ii. *γ can be continued by two steps to two 2-parameter families (on μ and e) of r-symmetric periodic orbits with period $qT\,f(e)$ where $e > 0$ is sufficiently small.*
      iii. *γ can be continued by two steps to two 2-parameter families (on μ and e) of t-symmetric periodic orbits with period $qT\,f(e)$ where $e > 0$ is sufficiently small.*
   (b) *p even and $q \neq 1$:*
      i. *γ can be continued directly to one 2-parameter family (on μ and τ) of doubly symmetric periodic orbits with period τ sufficiently close to T.*
      ii. *γ can be continued by two steps to two 2-parameter families (on μ and e) of doubly symmetric periodic orbits of period $qT\,f(e)$ where $e > 0$ is sufficiently small.*
   (c) *p even and $q = 1$:*
      i. *γ can be continued by two steps to two 2-parameter families (on μ and e) of doubly symmetric periodic orbits of period $qT\,f(e)$ where $e > 0$ is sufficiently small.*

Using direct continuation we can continue all periodic orbits of the reduced circular Sitnikov problem except the ones that have period multiple of 4π. In particular, we can continue the periodic orbits, with period $2\pi\omega$ and ω irrational. These periodic orbits become quasi-periodic orbits in the restricted isosceles problem. So, in fact we have continued quasi-periodic orbits of the restricted isosceles problem to either periodic or quasi-periodic orbits of the isosceles problem for $\mu > 0$ sufficiently small.

The continuation in two steps allows us to continue only periodic orbits of the reduced circular Sitnikov problem with period $T = 2\pi p/q$ for all $p, q \in \mathbb{N}$ coprime and $p > q/(2\sqrt{2})$. These periodic orbits become periodic orbits of the restricted isosceles problem. We note that the periodic orbits of the reduced circular Sitnikov problem that cannot be continued directly can be continued in two steps. Moreover the rest of the periodic orbits with period $T = 2\pi p/q$ can be continued in both ways, obtaining different periodic orbits for the reduced isosceles problem.

Since each periodic orbit of the reduced isosceles problem gives an invariant two-dimensional torus of the isosceles problem, in particular we have continued the invariant two-dimensional tori of the circular restricted isosceles problem ($\mu = 0$) to invariant two-dimensional tori of the isosceles problem for $\mu > 0$ sufficiently small. In section 13 we state Theorem A translated to the language of tori for the isosceles problem.

This paper is organized as follows. In section 2 we give the equations of motion

of the isosceles problem in appropriate cylindrical coordinates; these coordinates will allow us to define the reduced isosceles problem in section 3. In section 4 we give the relationships between the orbits of the reduced isosceles problem and the isosceles problem. In particular, we see that if $\overline{\varphi}$ is an orbit for the reduced isosceles problem, then $\overline{\varphi} \times \mathbb{S}^1$ is an invariant manifold for the isosceles problem (for more details see Theorem 4.1). In section 5 we analyze the symmetries of the reduced isosceles problem. In section 6 we define the restricted isosceles problems and the reduced restricted isosceles problems. In this work, we will consider only the circular and elliptic restricted isosceles problems, which are treated in sections 7 and 8, respectively. In particular, we are interested in the invariant two-dimensional tori of these problems that come from periodic orbits of the corresponding reduced problems. In section 7.1, we summarize the basic properties given in [8] of the periodic solutions of the circular Sitnikov problem. In section 8.1 we summarize the basic properties of the periodic solutions of the elliptic Sitnikov problem and give the basic results on continuation of periodic solutions from the circular Sitnikov problem ($e = 0$) to the elliptic Sitnikov problem for $e > 0$ sufficiently small. These results have also been extracted from [8]. In section 9 we analyze the variational equations of the reduced circular and elliptic Sitnikov problem and explicitly give the solution of the variational equations of the Kepler problem along a circular or elliptic periodic solution and the solution of the variational equations of the circular Sitnikov problem. In section 10 we analyze the direct continuation of periodic solutions from the reduced circular Sitnikov problem to the isosceles problem for $\mu > 0$ sufficiently small; in particular, we prove statements 1(a), 2(a)i, and 2(b)i of Theorem A (see Theorem 10.1). In section 11 we analyze the continuation of the symmetric periodic solutions of the reduced elliptic Sitnikov problem that we give in section 8 to the isosceles problem for $\mu > 0$ sufficiently small. The continuation by two steps from the reduced circular Sitnikov problem to the reduced isosceles problem is analyzed in section 12; in particular, we prove the remaining statements of Theorem A (see Theorem 12.8). In section 13 we summarize the basic results on continuation of invariant two-dimensional tori from the circular restricted isosceles problem to the isosceles problem for $\mu > 0$ small.

2. Coordinates and equations of motion of the isosceles problem. Let P_1 and P_2 be two particles, with equal masses $m_1 = m_2$, having initial positions and velocities symmetric with respect to a straight line that passes through their center of mass. Let P_3 be a third particle, with mass m_3, having initial position and velocity on this straight line. The *spatial isosceles* 3*-body problem*, or simply the *isosceles problem* in this work, consists of describing the motion of these three particles under their mutual Newtonian gravitational attraction. We note that the solutions of the isosceles problem are in fact solutions of the general spatial 3-body problem.

We choose an inertial coordinate system (X, Y, Z) in such a way that the Z-axis is the straight line that contains the particle P_3. The initial positions of the particles P_1, P_2, and P_3 in this coordinate system are (X, Y, Z_2), $(-X, -Y, Z_2)$, $(0, 0, Z_1)$, respectively, and their respective velocities are $(\dot{X}, \dot{Y}, \dot{Z}_2)$, $(-\dot{X}, -\dot{Y}, \dot{Z}_2)$, and $(0, 0, \dot{Z}_1)$ (see Figure 2.1). Of course, the dot denotes the derivative with respect to the time t.

In order to develop our analysis we will use the cylindrical coordinates $(r, z, \theta) \in \mathbb{R}^+ \times \mathbb{R} \times \mathbb{S}^1$ introduced as follows. Here $\mathbb{R}^+$ denotes the open interval $(0, \infty)$. First we put the origin $\mathbf{0}$ of the coordinate system at the center of mass of m_1, m_2, and m_3, which implies taking $Z_2 = -m_3 Z_1$. Then we define a new variable $z = Z_1 - Z_2 \in \mathbb{R}$ which denotes the distance between the third particle P_3 and the orthogonal plane to the Z-axis that contains the particles P_1 and P_2 with the convenient sign (positive

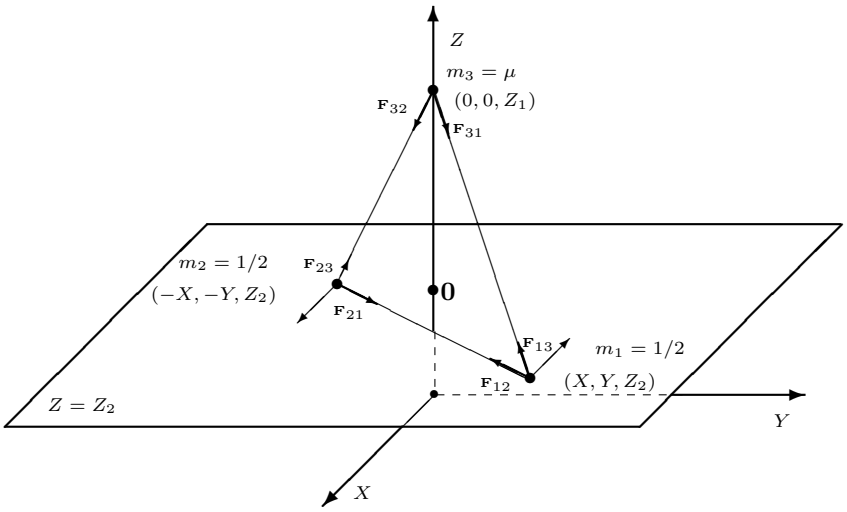


FIG. 2.1. *The isosceles problem.*

if $Z_1 > Z_2$ and negative if $Z_1 < Z_2$). Finally we consider polar coordinates, $(r, \theta) \in \mathbb{R}^+ \times \mathbb{S}^1$, in the above orthogonal plane by taking $X = r\cos\theta$ and $Y = r\sin\theta$.

We choose the unit of mass in such a way that $m_1 = m_2 = 1/2$ and $m_3 = \mu$, and the unit of length is chosen so that the gravitational constant is one. Then the *kinetic energy* and the *potential energy* in the coordinate system $(r, \dot{r}, z, \dot{z}, \theta, \dot{\theta})$ are given, respectively, by

$$T = \frac{1}{2}\left(\dot{r}^2 + r^2\dot{\theta}^2 + \frac{\mu}{1+\mu}\,\dot{z}^2\right) \qquad \text{and} \qquad U = -\frac{1}{8\,r} - \frac{\mu}{(z^2+r^2)^{1/2}}.$$

Therefore the *Lagrangian equations* of motion for the isosceles problem are

$$\begin{aligned} \frac{d}{dt}(\dot{r}) &= r\dot{\theta}^2 - \frac{1}{8\,r^2} - \frac{\mu r}{(z^2+r^2)^{3/2}}\ , \\ \frac{d}{dt}\left(\frac{\mu}{1+\mu}\,\dot{z}\right) &= -\frac{\mu z}{(z^2+r^2)^{3/2}}\ , \\ \frac{d}{dt}(r^2\dot{\theta}) &= 0. \end{aligned} \tag{2.1}$$

We note that the third equation of system (2.1) can be integrated directly, obtaining the first integral of the *angular momentum*

$$C = r^2\dot{\theta}. \tag{2.2}$$

Of course, system (2.1) also has the first integral given by the *energy* $H = T + U$.

3. The reduced isosceles problem. To avoid singular situations, throughout this work we consider only solutions of system (2.1) having nonzero angular momentum (i.e., in particular, we do not consider solutions with collision between the masses, either triple or double). We note that under this assumption it is sufficient to consider solutions of (2.1) having a fixed value of the angular momentum $C = c$ for some $c \neq 0$,

because the phase portrait of the isosceles problem on each angular momentum level $c \neq 0$ is the same as that shown in the following proposition.

PROPOSITION 3.1. *Let $(r(t), \dot{r}(t), z(t), \dot{z}(t), \theta(t), \dot{\theta}(t))$ be a solution of the isosceles problem (2.1) with angular momentum $C = c$ for some $c \neq 0$. If we take $\alpha^{1/2} = \overline{c}/c \neq 0$, then*

$$\varphi(t) = \left(\alpha r(\alpha^{3/2}t), \frac{\dot{r}(\alpha^{3/2}t)}{\alpha^{1/2}}, \alpha z(\alpha^{3/2}t), \frac{\dot{z}(\alpha^{3/2}t)}{\alpha^{1/2}}, \theta(\alpha^{3/2}t), \frac{\dot{\theta}(\alpha^{3/2}t)}{\alpha^{3/2}} \right)$$

is a solution of (2.1) with angular momentum $\overline{c}$.

Proof. It is easy to see that system (2.1) is invariant under the transformation

$$(t, r, \dot{r}, z, \dot{z}, \theta, \dot{\theta}) \longmapsto (\alpha^{3/2}t, \alpha r, \alpha^{-1/2}\dot{r}, \alpha z, \alpha^{-1/2}\dot{z}, \theta, \alpha^{-3/2}\dot{\theta}).$$

Thus $\varphi(t)$ is a solution of (2.1). Moreover the angular momentum of $\varphi(t)$ is given by

$$\alpha^2 r^2(\alpha^{3/2}t)\alpha^{-3/2}\dot{\theta}(\alpha^{3/2}t) = \alpha^{1/2}c = \overline{c}.$$

Then $\varphi(t)$ is a solution of (2.1) with angular momentum $\overline{c}$. □

Assuming that the value of the angular momentum is fixed at $C = c$ for some $c \neq 0$, we can reduce by two units the dimension of the phase space. Indeed, the variable θ does not appear explicitly in system (2.1); moreover from (2.2), $\dot{\theta} = c/r^2$, and thus we need to consider only the first two equations of (2.1) with $\dot{\theta}$ replaced by c/r^2. That is, we need to consider only the *reduced isosceles problem*

$$\ddot{r} = \frac{c^2}{r^3} - \frac{1}{8\,r^2} - \frac{\mu r}{(z^2 + r^2)^{3/2}}\,, \qquad \ddot{z} = -\frac{(1+\mu)z}{(z^2 + r^2)^{3/2}}. \tag{3.1}$$

4. Relationships between the reduced isosceles problem and the isosceles problem. Let $\varphi(t) = (r(t), \dot{r}(t), z(t), \dot{z}(t))$ be a solution of the reduced isosceles problem (3.1) for a fixed $c \neq 0$ with initial conditions $r(0) = r_0$, $\dot{r}(0) = \dot{r}_0$, $z(0) = z_0$, $\dot{z}(0) = \dot{z}_0$. For each $\theta_0 \in \mathbb{S}^1$, the solution $\varphi(t)$ gives rise to a solution $\gamma_{\varphi,\theta_0,c}(t) = (r(t), \dot{r}(t), z(t), \dot{z}(t), \theta(t), \dot{\theta}(t))$ of the isosceles problem (2.1) with angular momentum c, having initial conditions $r(0) = r_0$, $\dot{r}(0) = \dot{r}_0$, $z(0) = z_0$, $\dot{z}(0) = \dot{z}_0$, $\theta(0) = \theta_0 \pmod{2\pi}$, $\dot{\theta}(0) = c/r_0^2$, where $\dot{\theta}(t)$ and $\theta(t)$ are given by

$$\dot{\theta}(t) = \frac{c}{r^2(t)} \qquad \text{and} \qquad \theta(t) = \int_0^t \frac{c}{r^2(\tau)}\, d\tau + \theta_0 = F(t) + \theta_0. \tag{4.1}$$

It is well known that all solutions of the 3-body problem, except those that end in collision, are defined for all $t \in \mathbb{R}$ (see [18] or [21]). Since our isosceles problem is a particular case of the general 3-body problem, all its solutions with angular momentum $c \neq 0$ are defined for all $t \in \mathbb{R}$.

Fixing a value of $c \neq 0$, the union of the orbits $\gamma_{\varphi,\theta_0,c} = \{\gamma_{\varphi,\theta_0,c}(t) : t \in \mathbb{R}\}$, varying $\theta_0 \in \mathbb{S}^1$, is an invariant submanifold $\mathcal{E}_{\varphi,c}$ of the phase space of the isosceles problem $\mathcal{E} = \{(r, \dot{r}, z, \dot{z}, \theta, \dot{\theta}) \in \mathbb{R}^+ \times \mathbb{R}^3 \times \mathbb{S}^1 \times \mathbb{R}\}$. In particular, $\mathcal{E}_{\varphi,c}$ is an invariant submanifold of $\mathcal{E}_c = \{(r, \dot{r}, z, \dot{z}, \theta, \dot{\theta}) \in \mathcal{E} : r^2\dot{\theta} = c\}$. Note that $\mathcal{E}_c$, called the *angular momentum level* $C = c$, is a submanifold of dimension 5 of $\mathcal{E}$ because $c \neq 0$. The invariant submanifold $\mathcal{E}_{\varphi,c}$ is called the *relative set* associated to the orbit $\overline{\varphi} = \{\varphi(t) : t \in \mathbb{R}\}$, and it is diffeomorphic to $\overline{\varphi} \times \mathbb{S}^1$.

By the qualitative theory of differential equations we know that the orbits of the reduced isosceles problem (3.1) can be either equilibrium points, periodic orbits, or

orbits diffeomorphic to $\mathbb{R}$. Thus if $\overline{\varphi}$ is an equilibrium point, then the corresponding relative set is diffeomorphic to a circle $\mathbb{S}^1$ (a *relative periodic orbit*). If $\overline{\varphi}$ is a periodic orbit (i.e., a closed curve diffeomorphic to $\mathbb{S}^1$), then the corresponding relative set is diffeomorphic to a two-dimensional torus $\mathbb{S}^1 \times \mathbb{S}^1$ (a *relative torus*). This relative torus can be filled with either periodic or quasi-periodic orbits (in this last case the orbits are dense on the torus). We note that these kinds of tori are not KAM tori (see, for instance, [1]), because they are two-dimensional tori of a problem with three degrees of freedom, and the KAM tori of such a system have dimension 3. Finally if $\overline{\varphi}$ is neither an equilibrium point nor a periodic orbit, then the corresponding relative set is diffeomorphic to a cylinder $\mathbb{R} \times \mathbb{S}^1$. In particular, we have the following result.

THEOREM 4.1. *Let* $\overline{\varphi} = \{\varphi(t) = (r(t), \dot{r}(t), z(t), \dot{z}(t)) \ : \ t \in \mathbb{R}\}$ *be an orbit of the reduced isosceles problem* (3.1) *for a fixed value of* $c \neq 0$*; and let* $\gamma_{\varphi,\theta_0,c} = \{\gamma_{\varphi,\theta_0,c}(t) = (r(t), \dot{r}(t), z(t), \dot{z}(t), \theta(t), \dot{\theta}(t)) \ : \ t \in \mathbb{R}\}$ *be the orbit of the isosceles problem* (2.1) *with* $\theta(t) = F(t) + \theta_0$ *(see* (4.1)*) for a fixed* $\theta_0 \in \mathbb{S}^1$*. Then* $\mathcal{E}_{\varphi,c}$ *is diffeomorphic to one of the following manifolds:*

1. *A circle* $\mathbb{S}^1 \subset \mathcal{E}_c$ *formed by a periodic orbit of* (2.1) *with period* $128\pi c^3/(1 + 8\mu)^2$ *if* $\overline{\varphi} = (8c^2/(1 + 8\mu), 0, 0, 0)$ *is the equilibrium point of* (3.1)*. This periodic orbit is known as the collinear solution of Euler for the* 3*-body problem (for more details see* [21]*).*
2. *A two-dimensional torus* $\mathbb{S}^1 \times \mathbb{S}^1 \subset \mathcal{E}_c$ *if* $\overline{\varphi}$ *is a* T*-periodic orbit. Moreover this torus is formed by the union of*
   (a) *periodic orbits of period* mT *if* $F(T) = 2\pi l/m$ *with* $l \in \mathbb{Z}$*,* $m \in \mathbb{N}$ *and* l*,* m *coprime;*
   (b) *quasi-periodic orbits if* $F(T) = \omega 2\pi$ *with* ω *an irrational number.*
3. *A cylinder* $\mathbb{S}^1 \times \mathbb{R} \subset \mathcal{E}_c$ *if* $\overline{\varphi}$ *is neither the equilibrium point nor a periodic orbit.*

5. Symmetries. It is easy to check that the equations of motion of the reduced isosceles problem (3.1) are invariant under the symmetry

$$(t, r, \dot{r}, z, \dot{z}) \longmapsto (-t, r, -\dot{r}, -z, \dot{z}). \tag{5.1}$$

This means that if $\varphi(t) = (r(t), \dot{r}(t), z(t), \dot{z}(t))$ is a solution of system (3.1), then also $\psi(t) = (r(-t), -\dot{r}(-t), -z(-t), \dot{z}(-t))$ is a solution. We note that in the configuration space $\{(r, z) \in \mathbb{R}^+ \times \mathbb{R}\}$ this symmetry corresponds to a symmetry with respect to the r-axis, so in what follows it will be denoted by the *r-symmetry*. On the other hand, in the configuration space $\{(r, z, \theta) \in \mathbb{R}^+ \times \mathbb{R} \times \mathbb{S}^1\}$ the r-symmetry would correspond to a symmetry with respect to the plane defined by the motion of the particles P_1 and P_2.

This symmetry can be used, in a standard way, to find periodic solutions as follows. Suppose that $\varphi(t)$ crosses orthogonally the r-axis at a time $t = 0$; that is, $z(0) = 0$ and $\dot{r}(0) = 0$. Using symmetry (5.1) we have that the two solutions $\varphi(t)$ and $\psi(t)$ coincide at $t = 0$; then by the theorem of uniqueness of solutions of an ordinary differential equation they must be the same. If there is another time such that the solution $\varphi(t)$ crosses the r-axis orthogonally, then by symmetry (5.1) the orbit of $\varphi(t)$ must be closed, and $\varphi(t)$ is called an *r-symmetric periodic solution*.

Since system (3.1) is autonomous, the origin of time can be chosen arbitrarily. Thus, if $\gamma(t)$ is a solution of (3.1) that crosses the r-axis in a point $\mathbf{p}$ at $t = t_0$, then $\varphi(t) = \gamma(t + t_0)$ is a solution of (3.1) that crosses the r-axis in the point $\mathbf{p}$ at $t = 0$. Therefore we have proved the following well-known result.

PROPOSITION 5.1. *Let* $\varphi(t) = (r(t), \dot{r}(t), z(t), \dot{z}(t))$ *be a solution of the reduced*

isosceles problem (3.1). *If $\dot r(t)$ and $z(t)$ are zero at $t = t_0$ and at $t = t_0 + T/2$ but are not simultaneously zero at any value of $t \in (t_0, t_0 + T/2)$, then $\varphi(t)$ is an r-symmetric periodic solution of period T.*

Equations (3.1) are also invariant under the symmetry

$$(t, r, \dot r, z, \dot z) \longmapsto (-t, r, -\dot r, z, -\dot z), \tag{5.2}$$

i.e., the time reversibility symmetry, which will be denoted in what follows by the *t-symmetry*. As in the r-symmetry we can introduce the notion of *t-symmetric periodic solutions*, which are characterized as follows.

PROPOSITION 5.2. *Let $\varphi(t) = (r(t), \dot r(t), z(t), \dot z(t))$ be a solution of the reduced isosceles problem* (3.1). *If $\dot r(t)$ and $\dot z(t)$ are zero at $t = t_0$ and at $t = t_0 + T/2$ but are not simultaneously zero at any value of $t \in (t_0, t_0 + T/2)$, then $\varphi(t)$ is a t-symmetric periodic solution of period T.*

We note that there could be periodic solutions of (3.1) that are simultaneously r- and t-symmetric. These periodic solutions will be called *doubly symmetric periodic solutions* (see, for instance, [12] for more information about doubly symmetric periodic orbits) and are characterized by the following result.

PROPOSITION 5.3. *Let $\varphi(t) = (r(t), \dot r(t), z(t), \dot z(t))$ be a solution of the reduced isosceles problem* (3.1).

1. *If $\dot r(t)$ and $z(t)$ are zero at $t = t_0$ and $\dot r(t)$ and $\dot z(t)$ are zero at $t = t_0 + T/4$ but are not simultaneously zero at any value of $t \in (t_0, t_0 + T/4)$, then $\varphi(t)$ is a doubly symmetric periodic solution of period T.*
2. *If $\dot r(t)$ and $\dot z(t)$ are zero at $t = t_0$ and $\dot r(t)$ and $z(t)$ are zero at $t = t_0 + T/4$ but are not simultaneously zero at any value of $t \in (t_0, t_0 + T/4)$, then $\varphi(t)$ is a doubly-symmetric periodic solution of period T.*

6. Restricted isosceles problems. To obtain the restricted isosceles problems we assume that the value of the mass m_3 is infinitesimally small (i.e., $\mu = 0$). Then the equations of motion of the restricted isosceles problem become

$$\ddot r = r\dot\theta^2 - \frac{1}{8\,r^2}, \qquad \ddot z = -\frac{z}{(z^2 + r^2)^{3/2}}, \qquad \frac{d}{dt}(r^2\dot\theta) = 0. \tag{6.1}$$

Notice that the first and the third equations of (6.1) do not depend on z; moreover they are the equations of motion of a 2-body problem in polar coordinates. This means that the particles P_1 and P_2 (the *primaries*) move on the plane $z = 0$ describing a solution of this 2-body problem. Moreover the particle P_3 that lies on the straight line orthogonal to the plane containing P_1 and P_2 that passes through their center of mass moves under the gravitational attraction of the previous two but does not influence their motion. Thus, for every solution $(r(t), \theta(t))$ of that 2-body problem, system (6.1) defines a different restricted isosceles problem; it can be a circular, elliptic, parabolic, hyperbolic, elliptic collision, parabolic collision, or hyperbolic collision restricted isosceles problem depending on the nature of the solution $(r(t), \theta(t))$.

As in the isosceles problem (2.1) if we assume that the value of the angular momentum is fixed at $C = c$ for some $c \neq 0$, then we can reduce the dimension of the phase space by two, obtaining the *reduced restricted isosceles problem*

$$\ddot r = \frac{c^2}{r^3} - \frac{1}{8\,r^2}\,, \qquad \ddot z = -\frac{z}{(z^2 + r^2)^{3/2}}. \tag{6.2}$$

In this work we are interested only in the periodic solutions of system (6.2) for $c \neq 0$. So, we will consider only the reduced circular and elliptic restricted isosceles

problems, which we will call *reduced circular Sitnikov problem* and *reduced elliptic Sitnikov problem*, respectively.

7. On the circular restricted isosceles problem. Without loss of generality we can assume that the primaries describe a circular orbit of radius 1/2 (or, equivalently, a circular orbit of period 2π). This corresponds to fixing the value of the angular momentum to $c = 1/4$. Then the equation of motion for the infinitesimal mass becomes

$$\ddot{z} = -\frac{z}{(z^2 + 1/4)^{3/2}}, \tag{7.1}$$

which is the equation of the known *circular Sitnikov problem*.

Assume that $(z(t), \dot{z}(t))$ is a solution of (7.1) with arbitrary initial conditions $z(0) = z_0$ and $\dot{z}(0) = \dot{z}_0$. Then it is clear that $\varphi(t) = (r(t) = 1/2, \dot{r}(0) = 0, z(t), \dot{z}(t))$ is a solution of the reduced circular Sitnikov problem

$$\ddot{r} = \frac{1}{16\, r^3} - \frac{1}{8\, r^2}, \qquad \ddot{z} = -\frac{z}{(z^2 + r^2)^{3/2}}, \tag{7.2}$$

with initial conditions $r(0) = 1/2$, $\dot{r}(0) = 0$, $z(0) = z_0$, $\dot{z}(0) = \dot{z}_0$. Next we analyze the periodic solutions of this problem.

Since we have taken $r(t) = 1/2$, $\dot{r}(t) = 0$, it's clear that $\varphi(t) = (r(t), \dot{r}(t), z(t), \dot{z}(t))$ is a periodic solution of the reduced circular Sitnikov problem with period T if and only if $(z(t), \dot{z}(t))$ is a periodic solution of the circular Sitnikov problem (7.1) with period T. So, we start summarizing the basic results about periodic solutions of the circular Sitnikov problem (7.1) that are needed for the development of this work. Then we will analyze the periodic solutions of the reduced circular Sitnikov problem and their relationship with the corresponding solutions of the circular restricted isosceles problem.

7.1. Periodic solutions of the circular Sitnikov problem. Equation (7.1) defines an integrable Hamiltonian system of one degree of freedom with Hamiltonian $H = v^2/2 - \left(z^2 + 1/4\right)^{-1/2}$, where $v = \dot{z}$. The orbits for the circular Sitnikov problem in the energy level h are described by the curve $H = h$, where h varies in $[-2, \infty)$.

The circular Sitnikov problem has been studied by several authors. In 1907 Pavanini [19] expressed its solutions by means of Weierstrassian elliptic functions. Four years later MacMillan [14] expressed the solutions in terms of Jacobian elliptic functions (a detailed description of this work can be found in Stumpff [22]). Some other analytical expressions for the solutions of this problem can be found, for instance, in [23], [2], and [24]. In particular, in this paper we will use the analytical expressions of the solutions of the circular Sitnikov problem for $h > -2$ that appear in [2], which are given in terms of Jacobian elliptic functions. A detailed description of all Jacobian elliptic functions to be used in this paper can be found in [4] and [8].

We remark that the knowledge of an analytic expression for the solutions of the circular Sitnikov problem plays a key role in our analysis, because it allows us to prove our results analytically.

In what follows we use the following notation for the Jacobian elliptic functions: $\operatorname{sn}\nu = \operatorname{sn}(\nu, k)$, $\operatorname{cn}\nu = \operatorname{cn}(\nu, k)$, $\operatorname{dn}\nu = \operatorname{dn}(\nu, k)$ are the *sine*, *cosine*, and *delta amplitude* Jacobian elliptic functions, respectively; $F(\nu) = F(\operatorname{am}(\nu), k)$, $E(\nu) = E(\operatorname{am}(\nu), k)$, $\Pi(\nu, 2k^2) = \Pi(\operatorname{am}(\nu), 2k^2, k)$ are the *normal elliptic integral of the first*, *second*, and *third kind*, respectively; $\operatorname{am}(\nu)$ is the *amplitude* Jacobian elliptic function;

and finally $K = K(k) = F(\pi/2, k)$, $E = E(k) = E(\pi/2, k)$, $\Pi(\alpha^2, k) = \Pi(\pi/2, \alpha^2, k)$ are the *complete elliptic integrals of the first*, *second*, and *third kind*, respectively (see [4] or [8] for the precise definitions).

Using the analytic expression for the solutions of the circular Sitnikov problem given in Theorem A of [2], we see that the periodic solutions of that problem can be written as follows (see [8] for more details).

LEMMA 7.1. *The periodic solutions of the circular Sitnikov problem have energy* $-2 < h < 0$ *and can be written as*

$$(z(t), \dot{z}(t)) = \left(\frac{k \operatorname{sn}\nu \operatorname{dn}\nu}{1 - 2k^2 \operatorname{sn}^2\nu} \ , \ 2\sqrt{2}k \operatorname{cn}\nu \right), \tag{7.3}$$

where $k = \sqrt{2+h}/2$ *and* ν *is the function of* t *defined implicitly by*

$$\begin{aligned} t &= \frac{\sqrt{2}}{8(1-2k^2)} \left[2E(\nu) - \nu + \Pi(\nu, 2k^2) - 4k^2 \frac{\operatorname{sn}\nu \operatorname{cn}\nu \operatorname{dn}\nu}{1 - 2k^2 \operatorname{sn}^2\nu} \right] + C \\ &= \tau(\nu, k) + C. \end{aligned}$$

Here C *is an integration constant whose value depends on the initial conditions of the periodic solution* $(z(t), \dot{z}(t))$.

Since $\operatorname{sn}\nu$ and $\operatorname{cn}\nu$ are periodic functions of period $4K$ and $\operatorname{dn}\nu$ is a periodic function of period $2K$ (see formulas 122 in [4]), from (7.3) we see that the period in the new time ν is $4K$, where $K = K(k)$ is the complete elliptic integral of the first kind and $k = \sqrt{2+h}/2$. Moreover the period in the real time t is given by

$$T = \frac{\sqrt{2}}{2(1-2k^2)} [2E(k) - K(k) + \Pi(2k^2, k)]; \tag{7.4}$$

for more details see Theorem 2.3 in [8].

We note that (7.1) is autonomous, so the origin of time can be chosen arbitrarily. In particular, in this paper we are interested only in periodic solutions $(z(t), \dot{z}(t))$ having initial conditions either $z(0) = 0$ or $\dot{z}(0) = 0$. The following lemma, taken from [8], gives the values of the integration constant C for those initial conditions.

LEMMA 7.2. *Let* T *be the period of the periodic solution* $(z(t), \dot{z}(t))$ *given in* (7.4).

1. *If* $(z(t), \dot{z}(t))$ *has initial conditions* $z(0) = 0$ *and* $\dot{z}(0) = \sqrt{2h+4}$, *then taking* $\nu(0) = 0$, *we have* $t = \tau(\nu, k)$.
2. *If* $(z(t), \dot{z}(t))$ *has initial conditions* $z(0) = 0$ *and* $\dot{z}(0) = -\sqrt{2h+4}$, *then taking* $\nu(0) = 2K$, *we have* $t = \tau(\nu, k) - T/2$.
3. *If* $(z(t), \dot{z}(t))$ *has initial conditions* $z(0) = \sqrt{\frac{1}{h^2} - \frac{1}{4}}$ *and* $\dot{z}(0) = 0$, *then taking* $\nu(0) = K$, *we have* $t = \tau(\nu, k) - T/4$.
4. *If* $(z(t), \dot{z}(t))$ *has initial conditions* $z(0) = -\sqrt{\frac{1}{h^2} - \frac{1}{4}}$ *and* $\dot{z}(0) = 0$, *then taking* $\nu(0) = 3K$, *we have* $t = \tau(\nu, k) - 3T/4$.

In order to simplify computations we will usually work with the new time ν instead of the real time t, but always keeping in mind that ν is a function of t via Lemma 7.2. The two following lemmas taken also from [8] give some relationships between the real time t and the new time ν that will be useful later on.

LEMMA 7.3. *Let* T *be the period of the periodic solution* $(z(t), \dot{z}(t))$.

1. $\nu(t + qT) = \nu(t) + q4K$ *for all* $t \in \mathbb{R}$ *and for all* $q \in \mathbb{N}$.
2. $\nu(t + qT/2) = \nu(t) + q2K$ *for all* $t \in \mathbb{R}$ *and for all* $q \in \mathbb{N}$.

LEMMA 7.4. *Let T be the period of the solution $(z(t), \dot{z}(t))$. If $\nu(0) = lK$ for $l = 0, 1, 2, 3$, then $\nu(qT/4) = (l + q)K$ for all $q \in \mathbb{N}$.*

The following result gives the properties of the function $T = T(h)$.

THEOREM 7.5. *The period T satisfies*

1. $\lim_{h\to -2} T(h) = \pi/\sqrt{2}$;
2. $\lim_{h\to 0} T(h) = \infty$;
3. $dT/dh > 0$ *for all* $h \in (-2, 0)$;
4. $\lim_{h\to -2} dT/dh = \pi(1 + 4\sqrt{2})/16$;
5. $\lim_{h\to 0} dT/dh = \infty$.

Proof. See the proof of Theorem C in [2]. □

Theorem 7.5 assures the existence of periodic orbits of the circular Sitnikov problem with period $T = T(h)$ for all $T > \pi/\sqrt{2}$. In fact, since $T = T(h)$ is an injective function there is a one-to-one correspondence between $h \in (-2, 0)$ and $T \in (\pi/\sqrt{2}, \infty)$, so we can characterize the periodic orbits either by the period or by the energy.

7.2. Periodic solutions of the reduced circular Sitnikov problem. Notice that equations (7.2) are invariant under symmetries (5.1) and (5.2). These symmetries can be used to obtain symmetric periodic solutions for the reduced circular Sitnikov problem. It is not difficult to prove the next result.

PROPOSITION 7.6. *All periodic orbits of the reduced circular Sitnikov problem are doubly symmetric periodic orbits.*

We note that the periodic solutions of the reduced circular Sitnikov problem are periodic solutions for the infinitesimal mass, but in general they are not periodic solutions involving the three masses; that is, they are not periodic solutions of the circular restricted isosceles problem. Since the primaries describe a circular solution of a 2-body problem with period 2π, the only periodic orbits of the circular Sitnikov problem that give periodic orbits involving the three masses are the ones that have a period commensurable with 2π; that is, $T = T(h) = 2\pi p/q$ for some $p, q \in \mathbb{N}$ coprime. In this case the period of the corresponding orbit involving the three masses is $\tau = 2\pi p = qT(h)$. That is, during a period τ, the primaries have completed p revolutions and the infinitesimal mass has completed q revolutions.

7.3. Invariant tori of the circular restricted isosceles problem. From section 7.2, we have the following result, which can be obtained easily from Theorem 4.1.

PROPOSITION 7.7. *Let $\{(z_h(t), \dot{z}_h(t)) : t \in \mathbb{R}\}$ be a periodic orbit of the circular Sitnikov problem with energy h for some $h \in (-2, 0)$; and let $\overline{\varphi}_h = \{\varphi_h(t) = (r(t) = 1/2, \dot{r}(t) = 0, z_h(t), \dot{z}_h(t)) : t \in \mathbb{R}\}$ be its corresponding orbit of the reduced circular Sitnikov problem. Then the relative set of the circular restricted isosceles problem associated to the orbit $\overline{\varphi}_h$ is diffeomorphic to a two-dimensional torus $\mathbb{S}^1 \times \mathbb{S}^1$. Moreover, this relative torus is formed by the union of*

1. *periodic orbits of period qT if $T = T(h) = 2\pi p/q$ for some $p, q \in \mathbb{N}$ coprime and $p > q/(2\sqrt{2})$;*
2. *quasi-periodic orbits if $T = T(h) = 2\pi\omega$ for some irrational $\omega > 1/(2\sqrt{2})$.*

8. On the elliptic restricted isosceles problem. We assume that the primaries are describing an elliptic orbit of the 2-body problem with period 2π and eccentricity e. This corresponds to fixing the value of the angular momentum to $c = c_e = \sqrt{1 - e^2}/4$. Then, choosing conveniently the origin of time, a solution of the reduced elliptic Sitnikov problem is a solution $\varphi(t) = (r(t), \dot{r}(t), z(t), \dot{z}(t))$ of

$$\ddot{r} = \frac{1 - e^2}{16\, r^3} - \frac{1}{8\, r^2}\,, \qquad \ddot{z} = -\frac{z}{(z^2 + r^2)^{3/2}}, \tag{8.1}$$

with initial conditions $r(0) = (1 \pm e)/2$, $\dot{r}(0) = 0$, $z(0) = z_0$, $\dot{z}(0) = \dot{z}_0$ for some $z_0, \dot{z}_0 \in \mathbb{R}$.

Since $r(t)$ is a 2π-periodic function, the periodic solutions of the reduced elliptic Sitnikov problem must have period that is a multiple of 2π. Moreover $\varphi(t) = (r(t), \dot{r}(t), z(t), \dot{z}(t))$ is a periodic solution of the reduced elliptic Sitnikov problem with period $T = 2k\pi$ for some $k \in \mathbb{N}$ if and only if $(z(t), \dot{z}(t))$ is a periodic solution with period $T = 2k\pi$ of the elliptic Sitnikov problem

$$\ddot{z} = -\frac{z}{(z^2 + r(t)^2)^{3/2}}.$$

It is clear that equations (8.1) are invariant under symmetries (5.1) and (5.2). These symmetries can be used to obtain symmetric periodic solutions for the reduced elliptic Sitnikov problem. We remark that symmetries (5.1) and (5.2) for the reduced elliptic Sitnikov problem correspond to the r- and the t-symmetry of the elliptic Sitnikov problem defined in [7] and [8].

8.1. Periodic solutions of the reduced elliptic Sitnikov problem. In section 7.2 we have seen that all periodic orbits of the reduced circular Sitnikov problem are doubly symmetric periodic orbits. This fact does not occur when we consider the reduced elliptic Sitnikov problem, as follows from the next result.

PROPOSITION 8.1. *For the reduced elliptic Sitnikov problem there exist four different types of periodic orbits: nonsymmetric periodic orbits, doubly symmetric periodic orbits, and r- and t-symmetric periodic orbits that are not doubly symmetric.*

Proof. See the proof of Propositions 12, 15, and 23 in [7]. □

On the other hand, [8] gives initial conditions for some symmetric periodic solutions of the elliptic Sitnikov problem (or, equivalently, the reduced elliptic Sitnikov problem) with sufficiently small values of the eccentricity $e > 0$. These initial conditions are obtained from the analytic continuation of the known periodic solutions of the reduced circular Sitnikov problem to symmetric periodic solutions of the reduced elliptic Sitnikov problem for sufficiently small values of the eccentricity e. Later on, in section 11, the symmetric periodic solutions of the reduced elliptic Sitnikov problem given in [8] will be continued to the reduced isosceles problem for sufficiently small values of $\mu > 0$. Here we summarize the main results of [8] about symmetric periodic orbits of the reduced elliptic Sitnikov problem.

In what follows $\varphi_c(t; \mathbf{x}_0, \mu) = (r(t; \mathbf{x}_0, \mu), \dot{r}(t; \mathbf{x}_0, \mu), z(t; \mathbf{x}_0, \mu), \dot{z}(t; \mathbf{x}_0, \mu))$, with $\mathbf{x}_0 = (r_0, \dot{r}_0, z_0, \dot{z}_0)$, denotes the solution of the reduced isosceles problem (3.1) with angular momentum $C = c \neq 0$, satisfying the initial conditions $r(0; r_0, \dot{r}_0, z_0, \dot{z}_0, \mu) = r_0$, $\dot{r}(0; r_0, \dot{r}_0, z_0, \dot{z}_0, \mu) = \dot{r}_0$, $z(0; r_0, \dot{r}_0, z_0, \dot{z}_0, \mu) = z_0$, $\dot{z}(0; r_0, \dot{r}_0, z_0, \dot{z}_0, \mu) = \dot{z}_0$.

THEOREM 8.2. *Given $p, q \in \mathbb{N}$ coprime with $p > q/(2\sqrt{2})$, let $\varphi_{1/4}(t; r_0 = 1/2, \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{z}_0^* = \pm\sqrt{2h+4}, \mu = 0)$ be a periodic solution of the reduced circular Sitnikov problem with period $T = 2\pi p/q$.*

1. *This solution can be continued to two families $\varphi_{c_e}(t; r_0 = (1-e)/2, \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{z}_0^P = \dot{z}_0^* + O(e), \mu = 0)$ and $\varphi_{c_e}(t; r_0 = (1+e)/2, \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{z}_0^A = \dot{z}_0^* + O(e), \mu = 0)$ of r-symmetric periodic solutions of the reduced elliptic Sitnikov problem having period $\tau = 2\pi p = qT$ for $e > 0$ sufficiently small.*
2. *If p is odd, then those r-symmetric periodic solutions are not doubly symmetric, whereas if p is even, then they are doubly symmetric.*

Proof. See the proof of Theorem 4.4 in [8]. □

THEOREM 8.3. *Given $p, q \in \mathbb{N}$ coprime with $p > q/(2\sqrt{2})$, let $\varphi_{1/4}(t; r_0 = 1/2,$*

$\dot r_0 = 0, z_0 = z_0^* = \pm\sqrt{\frac{1}{h^2} - \frac{1}{4}}, \dot z_0 = 0, \mu = 0)$ *be a periodic solution of the reduced circular Sitnikov problem with period* $T = 2\pi p/q$.

1. *This solution can be continued to two families* $\varphi_{c_e}(t; r_0 = (1-e)/2, \dot r_0 = 0, z_0 = z_0^P = z_0^* + O(e), \dot z_0 = 0, \mu = 0)$ *and* $\varphi_{c_e}(t; r_0 = (1+e)/2, \dot r_0 = 0, z_0 = z_0^A = z_0^* + O(e), \dot z_0 = 0, \mu = 0)$ *of* t*-symmetric periodic solutions of the reduced elliptic Sitnikov problem having period* $\tau = 2\pi p = qT$ *for* $e > 0$ *sufficiently small.*
2. *If* p *is odd, then those* t*-symmetric periodic solutions are not doubly symmetric, whereas if* p *is even, then they are doubly symmetric.*

Proof. See the proof of Theorem 4.6 in [8]. □

We note that in Theorems 8.2 and 8.3 we continue four different initial conditions of the periodic orbit of the reduced circular Sitnikov problem with period $T = 2\pi p/q$ for given $p, q \in \mathbb{N}$ coprime, $p > q/(2\sqrt{2})$; they are $\varphi_{1/4}\left(t; 1/2, 0, 0, \sqrt{2h+4}, 0\right)$ and $\varphi_{1/4}\left(t; 1/2, 0, 0, -\sqrt{2h+4}, 0\right)$ in Theorem 8.2, and $\varphi_{1/4}(t; 1/2, 0, \sqrt{\frac{1}{h^2} - \frac{1}{4}}, 0, 0)$ and $\varphi_{1/4}(t; 1/2, 0, -\sqrt{\frac{1}{h^2} - \frac{1}{4}}, 0, 0)$ in Theorem 8.3. These four initial conditions are continued to eight families of periodic orbits of the reduced elliptic Sitnikov problem for $e > 0$ sufficiently small. The following theorem says how many of these eight families of periodic orbits are really different (see [8] for more details).

THEOREM 8.4. *The periodic solutions of the reduced circular Sitnikov problem with period* $T = 2\pi p/q$, *for given* $p, q \in \mathbb{N}$ *coprime* $p > q/(2\sqrt{2})$, *can be continued to*

1. *two families of* r*-symmetric periodic orbits and two families of* t*-symmetric periodic orbits (that are not doubly symmetric) of the reduced elliptic Sitnikov problem with period* $\tau = 2\pi p = qT$, *for* $e > 0$ *sufficiently small, when* p *is odd;*
2. *two families of doubly symmetric periodic orbits of the reduced elliptic Sitnikov problem with period* $\tau = 2\pi p = qT$, *for* $e > 0$ *sufficiently small, when* p *is even.*

Proof. See the proof of Theorem 4.15 in [8]. □

8.2. Invariant tori of the elliptic restricted isosceles problem. From Theorem 4.1(2)(a), the next result follows.

PROPOSITION 8.5. *Let* $\overline{\varphi} = \{\varphi(t) = (r(t), \dot r(t), z(t), \dot z(t)) : t \in \mathbb{R}\}$ *be a periodic orbit of the reduced elliptic Sitnikov problem with period* $\tau = 2\pi n$ *for some* $n \in \mathbb{N}$. *Then the relative set of the restricted isosceles problem associated to the orbit* $\overline{\varphi}$ *is diffeomorphic to a two-dimensional torus* $\mathbb{S}^1 \times \mathbb{S}^1 \subset \mathcal{E}_{c_e}$, *which is formed by periodic orbits of period* τ.

We remark that the orbits of the circular restricted isosceles problem coming from periodic orbits of the reduced circular Sitnikov problem are not in general periodic orbits (see Proposition 7.7).

By means of Propositions 7.7 and 8.5, Theorem 8.4 can be extended to the restricted isosceles problem, obtaining the following result.

THEOREM 8.6. *Let* Γ_{pq} *be the periodic two-dimensional tori of the circular restricted isosceles problem that comes from the periodic orbit of the reduced circular Sitnikov problem with period* $T = p2\pi/q$, $p, q \in \mathbb{N}$ *coprime and* $p > q/2\sqrt{2}$. *Then* Γ_{pq} *can be continued to two or four families (two for even* p *and four for odd* p*) of periodic two-dimensional tori of the elliptic restricted isosceles problem.*

9. Variational equations. The main objective of this work is to continue the known symmetric periodic orbits of the reduced circular and elliptic Sitnikov problems to symmetric periodic orbits of the reduced isosceles problem for $\mu > 0$ sufficiently

small. Those periodic orbits will be continued by using the classical analytic continuation method of Poincaré (for details see [21] or [15]). In order to apply this method to our problem we must know the solution of the variational equations of the reduced circular and elliptic Sitnikov problems along the periodic solutions that we want to continue. In this section we will analyze those variational equations.

Let $(r(t), R(t), z(t), Z(t))$ be a solution of the reduced circular ($e = 0$) or elliptic ($0 < e < 1$) Sitnikov problem

$$\dot r = R, \qquad \dot R = \frac{1-e^2}{16\, r^3} - \frac{1}{8\, r^2}, \qquad \dot z = Z, \qquad \dot Z = -\frac{z}{(z^2+r^2)^{3/2}}, \tag{9.1}$$

with initial conditions $r(0) = r_0 = (1 \pm e)/2$, $R(0) = R_0 = 0$, $z(0) = z_0$, $Z(0) = Z_0$. In particular, $(r(t), R(t))$ is a circular or elliptic solution of the Kepler problem

$$\dot r = R\ , \qquad \dot R = \frac{1-e^2}{16\, r^3} - \frac{1}{8\, r^2}; \tag{9.2}$$

and $(z(t), Z(t))$ is a solution of the circular or elliptic Sitnikov problem

$$\dot z = Z\ , \qquad \dot Z = -\frac{z}{(z^2+r^2(t))^{3/2}} \tag{9.3}$$

(see sections 7 and 8).

The variational equations of system (9.1) along the solution curve $(r(t), R(t), z(t), Z(t))$ are given by the matrix differential equation

$$\frac{d}{dt} A = B(t)\, A, \tag{9.4}$$

with initial condition $A(0) = I$ (the 4×4 identity matrix), where

$$A = \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix}\ , \qquad B(t) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ b_1(t) & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ b_2(t) & 0 & b_3(t) & 0 \end{pmatrix},$$

with A_1, A_2, A_3, and A_4 given by

$$\begin{pmatrix} \dfrac{\partial r}{\partial r_0} & \dfrac{\partial r}{\partial R_0} \\ \dfrac{\partial R}{\partial r_0} & \dfrac{\partial R}{\partial R_0} \end{pmatrix}, \quad \begin{pmatrix} \dfrac{\partial r}{\partial z_0} & \dfrac{\partial r}{\partial Z_0} \\ \dfrac{\partial R}{\partial z_0} & \dfrac{\partial R}{\partial Z_0} \end{pmatrix}, \quad \begin{pmatrix} \dfrac{\partial z}{\partial r_0} & \dfrac{\partial z}{\partial R_0} \\ \dfrac{\partial Z}{\partial r_0} & \dfrac{\partial Z}{\partial R_0} \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} \dfrac{\partial z}{\partial z_0} & \dfrac{\partial z}{\partial Z_0} \\ \dfrac{\partial Z}{\partial z_0} & \dfrac{\partial Z}{\partial Z_0} \end{pmatrix},$$

respectively, and

$$b_1(t) = -\frac{3(1-e^2)}{16\, r^4(t)} + \frac{1}{4\, r^3(t)},\ b_2(t) = \frac{3\, r(t)\, z(t)}{(z^2(t)+r^2(t))^{5/2}},\ b_3(t) = \frac{2\, z^2(t) - r^2(t)}{(z^2(t)+r^2(t))^{5/2}}.$$

If we denote $q_1 = r_0$, $q_2 = R_0$, $q_3 = z_0$, and $q_4 = Z_0$ system (9.4) can be written like the linear system of differential equations,

$$\begin{cases} \dfrac{d}{dt}\left(\dfrac{\partial r}{\partial q_i}\right) = \dfrac{\partial R}{\partial q_i}\ , \\ \dfrac{d}{dt}\left(\dfrac{\partial R}{\partial q_i}\right) = \left(-\dfrac{3(1-e^2)}{16\, r^4(t)} + \dfrac{1}{4\, r^3(t)}\right) \dfrac{\partial r}{\partial q_i}\ , \end{cases} \tag{9.5}$$

$$\text{(9.6)}\qquad \begin{cases} \dfrac{d}{dt}\left(\dfrac{\partial z}{\partial q_i}\right) = \dfrac{\partial Z}{\partial q_i}\ , \\ \dfrac{d}{dt}\left(\dfrac{\partial Z}{\partial q_i}\right) = \dfrac{3\,r(t)z(t)}{(z^2(t)+r^2(t))^{5/2}}\dfrac{\partial r}{\partial q_i} + \dfrac{2\,z^2(t)-r^2(t)}{(z^2(t)+r^2(t))^{5/2}}\dfrac{\partial z}{\partial q_i}\ , \end{cases}$$

with initial conditions

$$\frac{\partial r}{\partial q_i}(0) = \delta_{1,i}\ , \qquad \frac{\partial R}{\partial q_i}(0) = \delta_{2,i}\ , \qquad \frac{\partial z}{\partial q_i}(0) = \delta_{3,i}\ , \qquad \frac{\partial Z}{\partial q_i}(0) = \delta_{4,i}\ ,$$

where $i = 1, \ldots, 4$, $\delta_{i,j} = 1$ if $i = j$ and $\delta_{i,j} = 0$ if $i \neq j$.

Since equations (9.5) do not depend on $\partial z/\partial q_i$ and $\partial Z/\partial q_i$, they can be solved separately. Thus, the derivatives $\partial r/\partial r_0$, $\partial r/\partial R_0$, $\partial R/\partial r_0$, and $\partial R/\partial R_0$ are given by the solution of the matrix differential equation

$$\text{(9.7)}\qquad \frac{d}{dt}\ A_1 = \begin{pmatrix} 0 & 1 \\ -\dfrac{3(1-e^2)}{16\,r^4(t)} + \dfrac{1}{4\,r^3(t)} & 0 \end{pmatrix} A_1\ ,$$

with initial condition $A_1(0) = I$ (the 2×2 identity matrix); that is, they are given by the solution of the variational equations of the Kepler problem (9.2) along the solution curve $(r(t), R(t))$.

On the other hand,

$$\text{(9.8)}\qquad \frac{\partial r}{\partial z_0}(t) = 0\ , \qquad \frac{\partial R}{\partial z_0}(t) = 0\ , \qquad \frac{\partial r}{\partial Z_0}(t) = 0\ , \qquad \frac{\partial R}{\partial Z_0}(t) = 0,$$

because the first two equations of (9.1) do not depend on z and Z; consequently changes on the initial conditions z_0 and Z_0 do not affect the solution $(r(t), R(t))$.

By (9.6) and (9.8), the derivatives $\partial z/\partial z_0$, $\partial z/\partial Z_0$, $\partial Z/\partial z_0$, and $\partial Z/\partial Z_0$ are given by the solution of the matrix differential equation

$$\text{(9.9)}\qquad \frac{d}{dt}\ A_4 = \begin{pmatrix} 0 & 1 \\ -\dfrac{2\,z^2(t)-r^2(t)}{(z^2(t)+r^2(t))^{5/2}} & 0 \end{pmatrix} A_4,$$

with initial condition $A_4(0) = I$ (the 2×2 identity matrix); that is, they are given by the solution of the variational equations of the circular or elliptic Sitnikov problem (9.3) along the solution curve $(z(t), Z(t))$.

We note that we do not know an exact expression for the symmetric periodic solutions of the nonautonomous elliptic Sitnikov problem, and thus their variational equations cannot be solved explicitly. However, since the eccentricity e is sufficiently small, the solution of these variational equations may be expressed as a power series of the eccentricity e. We have computed analytically the terms of zero order in e. They are given by the variational equations of the circular Sitnikov problem.

Finally the derivatives $\partial z/\partial r_0$, $\partial z/\partial R_0$, $\partial Z/\partial r_0$, and $\partial Z/\partial R_0$ are obtained by solving the nonhomogeneous linear system of differential equations that comes from replacing in (9.6) $\partial r/\partial q_i$ and $\partial R/\partial q_i$, $i = 1, 2$, by the solutions $(\partial r/\partial q_i)(t)$ and $(\partial R/\partial q_i)(t)$ of the variational equations of the Kepler problem (9.2) along the solution curve $(r(t), R(t))$. If we know a fundamental matrix $\Phi(t)$ of the variational equations of the circular or elliptic Sitnikov problem (9.3) along the solution curve $(z(t), Z(t))$ (i.e., a fundamental matrix of the homogeneous system), then we can solve

the nonhomogeneous one using the method of variation of constants (see, for instance, [11, p. 81]). Thus, for $i = 1, 2$, we have that

$$\begin{pmatrix} \dfrac{\partial z}{\partial q_i}(t) \\ \dfrac{\partial Z}{\partial q_i}(t) \end{pmatrix} = \Phi(t) \int_0^t \Phi^{-1}(s) \begin{pmatrix} 0 \\ \dfrac{3\, r(s) z(s)}{(z^2(s) + r^2(s))^{5/2}} \dfrac{\partial r}{\partial q_i}(s) \end{pmatrix} ds.$$

In order to compute the solution of the variational equations of the Kepler problem (9.2), for $0 \leqslant e < 1$, and of the circular Sitnikov problem (9.3) with $r(t) = 1/2$, we could use a theorem of Diliberto [10] on the integration of the homogeneous variational equations of a plane autonomous differential system in terms of geometric quantities along a given solution curve of the system (see also the paper of Chicone [5], where, in addition to using the Diliberto theorem to address his problem, he corrects a flaw in the theorem). But we compute here the solution of those variational equations directly using a result that appears in [8].

We note that the Kepler problem (9.2) and the circular Sitnikov problem (9.3) with $r(t) = 1/2$ can be written like a second order differential equation of the form

$$\ddot{x} = f(x). \tag{9.10}$$

The solution of the variational equations of (9.10) along a given nonconstant solution curve $x(t)$ are given by the following result.

PROPOSITION 9.1. *The linear variational equations of* (9.10) *along a nonconstant solution curve* $x(t)$ *have a fundamental matrix* $\Phi(t)$*, satisfying that* $\det(\Phi(0)) = 1$*, which is given by*

$$\Phi(t) = \begin{pmatrix} \dot{x}(t) & g(t) \\ f(x(t)) & \dot{g}(t) \end{pmatrix},$$

where $g(t) = \dot{x}(t) \int \frac{dt}{\dot{x}^2(t)}$ *without the constant due to integration.*

Moreover, the solution of these variational equations is given by

$$\begin{pmatrix} \dfrac{\partial x}{\partial x_0}(t) & \dfrac{\partial x}{\partial y_0}(t) \\ \dfrac{\partial y}{\partial x_0}(t) & \dfrac{\partial y}{\partial y_0}(t) \end{pmatrix} = \Phi(t)\Phi^{-1}(0),$$

where $y = \dot{x}$*,* $x_0 = x(0)$*, and* $y_0 = \dot{x}(0)$*.*

Proof. See the proof of Proposition B.1 in [8]. □

9.1. Variational equations of the Kepler problem. We start computing a fundamental matrix of the variational equations (9.7) of the Kepler problem (9.2) for $0 \leqslant e < 1$ along an arbitrary elliptic solution (a circular solution if $e = 0$)

$$r(t) = \tfrac{1}{2}(1 - e\cos u). \tag{9.11}$$

As usual u is the eccentric anomaly which is a function of t via the Kepler's equation

$$u - e\sin u = t - \tau = M, \tag{9.12}$$

where M is the mean anomaly and τ is the time of pericenter passage. Later on we will give the solution of those variational equations when $(r(t), R(t))$ is the solution with initial conditions $r(0) = (1 \pm e)/2$ and $R = \dot{r}(0) = 0$.

We note that when $e = 0$ we cannot apply Proposition 9.1 to solve the variational equations (9.7) of the Kepler problem (9.2) along the solution curve (9.11), because $r(t) = 1/2$ is constant.

PROPOSITION 9.2. *When $e = 0$, the solution of the variational equations* (9.7) *of the Kepler problem* (9.2) *along the solution curve* (9.11) *is given by*

$$A_1(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}.$$

Proof. The proof follows easily, noting that the solution of the variational equations (9.7) when $e = 0$ is a matrix whose columns are the solutions of the differential equation

$$\frac{d}{dt}\begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix},$$

with initial conditions $\omega_1(0) = 1,\ \omega_2(0) = 0$ and $\omega_1(0) = 0,\ \omega_2(0) = 1$, respectively. □

When $0 < e < 1$, to solve the variational equations (9.7) of the Kepler problem (9.2) along the solution curve (9.11), we apply Proposition 9.1. Thus a fundamental matrix of those variational equations is given by

$$\Phi(t) = \begin{pmatrix} \Phi_{11}(t) & \Phi_{12}(t) \\ \Phi_{21}(t) & \Phi_{22}(t) \end{pmatrix} = \begin{pmatrix} \dot r(t) & g(t) \\ \dfrac{1-e^2}{16\,r^3(t)} - \dfrac{1}{8\,r^2(t)} & \dot g(t) \end{pmatrix},$$

where $g(t) = \dot r(t) \int \frac{dt}{\dot r^2(t)}$.

In order to simplify our computations we will work with the eccentric anomaly, u, instead of the real time, t, but keeping in mind that u is a function of t via (9.12) when it is necessary.

Replacing $r(t)$ by (9.11) in $\Phi_{21}(t)$ and simplifying we get that

$$\Phi_{21}(t) = -\frac{e^2 - e\cos u}{2(1-e\cos u)^3}\ . \tag{9.13}$$

Differentiating (9.11) with respect to t we obtain

$$\Phi_{11}(t) = \dot r(t) = \frac{dr}{du}\frac{du}{dt} = \frac{e\sin u}{2(1-e\cos u)}. \tag{9.14}$$

Substituting $\dot r(t)$ into $g(t)$ and working with the variable u instead of the variable t, we have that

$$\begin{aligned} \Phi_{12}(t) = g(t) &= \frac{2\sin u}{e(1-e\cos u)}\int \frac{(1-e\cos u)^3}{\sin^2 u}\,du \\ &= \frac{2}{e(1-e\cos u)}[-(1+3e^2)\cos u - 3e^2 u\sin u + e^3\sin^2 u + 3e + e^3]. \end{aligned} \tag{9.15}$$

Finally, differentiating $g(t)$ we obtain

$$\begin{aligned} \Phi_{22}(t) = \dot g(t) = \frac{dg}{du}\frac{du}{dt} &= \frac{2}{e(1-e\cos u)^3}[\sin u(1-3e^2-3e^4) - 3e^2 u\cos u \\ &\quad + 5e^3\sin u\cos u + 3e^3 u + e^4\sin^3 u]. \end{aligned} \tag{9.16}$$

In short, we have proved the following result.

PROPOSITION 9.3. *A fundamental matrix of the variational equations* (9.7) *of the Kepler problem* (9.2) *along the solution curve* (9.11)*, when* $0 < e < 1$*, is* $\Phi(t) = (\Phi_{ij}(t))$*, where* $\Phi_{ij}(t)$*, with* $i, j = 1, 2$*, are given by* (9.13), (9.14), (9.15), *and* (9.16)*, and* u *is the eccentric anomaly as a function of* t *via the Kepler equation* (9.12)*. Moreover the solution of these variational equations is*

$$A_1(t) = \Phi(t)\Phi^{-1}(0). \tag{9.17}$$

Now we compute the solution of the variational equations (9.7) of the Kepler problem (9.2) along the elliptic solution $(r(t), R(t))$ with initial conditions $r(0) = (1 \pm e)/2$ and $R(0) = 0$.

Case $r(0) = (1 - e)/2$. Without loss of generality, we can assume that $u(0) = 0$. Then the Kepler equation (9.12) becomes

$$u - e \sin u = t. \tag{9.18}$$

Therefore, by Proposition 9.3, the fundamental matrix $\Phi(t)$ evaluated at $t = 0$ (or, equivalently, at $u = 0$) is given by $\Phi_{11}(0) = \Phi_{22}(0) = 0$, $\Phi_{12}(0) = -2(1 - e)^2/e$, $\Phi_{21}(0) = e/(2(1-e)^2)$. Therefore, from (9.17) after doing some computations, we get

$$\begin{aligned}
\frac{\partial r}{\partial r_0}(t) &= \frac{(1 + 3e^2)\cos u + 3e^2 u \sin u - e^3 \sin^2 u - 3e - e^3}{(1 - e)^2(1 - e\cos u)},\\
\frac{\partial r}{\partial R_0}(t) &= (1 - e)^2 \frac{\sin u}{(1 - e\cos u)},\\
\frac{\partial R}{\partial r_0}(t) &= -\frac{1}{(1 - e)^2(1 - e\cos u)^3}[(1 - 3e^2 - 3e^4)\sin u\\
&\quad - 3e^2 u \cos u + 5e^3 \sin u \cos u + 3e^3 u + e^4 \sin^3 u],\\
\frac{\partial R}{\partial R_0}(t) &= (1 - e)^2 \frac{(\cos u - e)}{(1 - e\cos u)^3},
\end{aligned} \tag{9.19}$$

and u is the eccentric anomaly as a function of time via (9.18).

Case $r(0) = (1 + e)/2$. Without loss of generality, we can assume that $u(0) = \pi$, and the Kepler equation (9.12) becomes

$$u - e \sin u = t + \pi. \tag{9.20}$$

By Proposition 9.3, the fundamental matrix $\Phi(t)$ evaluated at $t = 0$, or, equivalently, at $u = \pi$, is given by $\Phi_{11}(0) = 0$, $\Phi_{12}(0) = 2(1+e)^2/e$, $\Phi_{21}(0) = -e/(2(1+e)^2)$, $\Phi_{22}(0) = 6e\pi/(1 + e)^2$. Thus, by (9.17) we have

$$\begin{aligned}
\frac{\partial r}{\partial r_0}(t) &= -\frac{(1 + 3e^2)\cos u + 3e^2 u \sin u - e^3 \sin^2 u - 3e - e^3 - 3e^2\pi \sin u}{(1 + e)^2(1 - e\cos u)},\\
\frac{\partial r}{\partial R_0}(t) &= -(1 + e)^2 \frac{\sin u}{1 - e\cos u},\\
\frac{\partial R}{\partial r_0}(t) &= \frac{1}{(1 + e)^2(1 - e\cos u)^3}[(1 - 3e^2 - 3e^4)\sin u - 3e^2 u \cos u\\
&\quad + 5e^3 \sin u \cos u + 3e^3 u + e^4 \sin^3 u + 3e^2\pi \cos u - 3e^3\pi],\\
\frac{\partial R}{\partial R_0}(t) &= -(1 + e)^2 \frac{(\cos u - e)}{(1 - e\cos u)^3},
\end{aligned} \tag{9.21}$$

and u is the eccentric anomaly as a function of time via (9.20).

9.2. Variational equations of the circular Sitnikov problem. The variational equations (9.9) of the circular Sitnikov problem (9.3) with $r = 1/2$ along a given periodic solution curve $(z(t), Z(t))$ were solved in [8]; therefore we will refer to the corresponding results in this paper when it is necessary.

9.3. Variational equations of the elliptic Sitnikov problem for small values of the eccentricity. We consider the elliptic Sitnikov problem (9.3) where $r(t)(1 - e\cos u)/2$ is the elliptic solution of the Kepler problem (9.2), $0 < e < 1$, and u is the eccentric anomaly, which is a function of t via equation (9.12).

If the eccentricity e is small, then $r(t)$ may be expanded in terms of the mean anomaly M and of the eccentricity e, and $r(t) = (1 - e\cos M)/2 + O(e^2)$ (see, for instance, [3]). Thus, system (9.3) may be written as

$$\dot z = Z, \qquad \dot Z = -\frac{z}{(z^2+1/4)^{3/2}} - e\left[\frac{3}{4}\frac{z}{(z^2+1/4)^{5/2}}\cos M\right] + O(e^2). \tag{9.22}$$

Let $(z(t), Z(t))$ be a periodic solution of system (9.22). If the eccentricity e is sufficiently small, then by the Poincaré expansion theorem (see, for instance, [20] or [13]) $(z(t), Z(t))$ may be expanded in power series of e and

$$(z(t), Z(t)) = (z_{(0)}(t) + z_{(1)}(t)e + O(e^2), Z_{(0)}(t) + Z_{(1)}(t)e + O(e^2)),$$

where $(z_{(0)}(t), Z_{(0)}(t))$ is a given solution of the circular Sitnikov problem (or, equivalently, a solution of (9.22) for $e = 0$).

We analyze here the solution of the variational equations of the elliptic Sitnikov problem (9.3) along the solution curve $(z(t), Z(t))$ for $e > 0$ sufficiently small. These variational equations are given by the matrix differential equation

$$\frac{d}{dt}\,A_4 = \begin{pmatrix} 0 & 1 \\ \overline{b}(t) & 0 \end{pmatrix} A_4,$$

with initial condition $A_4(0) = I$ (the 2×2 identity matrix), where

$$\begin{aligned}\overline{b}(t) &= \frac{2z^2(t) - 1/4}{(z^2(t)+1/4)^{5/2}} + e\left[\frac{3}{4}\frac{4z^2(t)-1/4}{(z^2(t)+1/4)^{7/2}}\cos M\right] + O(e^2)\\ &= \frac{2{z_{(0)}}^2(t) - 1/4}{({z_{(0)}}^2(t)+1/4)^{5/2}} + eF(t) + O(e^2),\end{aligned}$$

and

$$F(t) = \frac{-6{z_{(0)}}^3(t)z_{(1)}(t) + 9z_{(0)}(t)z_{(1)}(t)/4 + 3\cos M(4{z_{(0)}}^2(t) - 1/4)/4}{({z_{(0)}}^2(t)+1/4)^{7/2}}.$$

Thus the derivatives $(\partial z/\partial z_0, \partial Z/\partial z_0)$ and $(\partial z/\partial Z_0, \partial Z/\partial Z_0)$ are given by the solution of system

$$\frac{dx}{dt} = y, \qquad \frac{dy}{dt} = \left(\frac{2{z_{(0)}}^2(t) - 1/4}{({z_{(0)}}^2(t)+1/4)^{5/2}} + F(t)e + O(e^2)\right)x,$$

with initial conditions $x(0) = 1$, $y(0) = 0$ and $x(0) = 0$, $y(0) = 1$, respectively. By the Poincaré expansion theorem this solution may be expanded in power series of e and

$$\begin{pmatrix} \dfrac{\partial z}{\partial z_0}(t) & \dfrac{\partial z}{\partial Z_0}(t) \\ \dfrac{\partial Z}{\partial z_0}(t) & \dfrac{\partial Z}{\partial Z_0}(t) \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{n=0}^{\infty} x_{1(n)}(t)e^n & \displaystyle\sum_{n=0}^{\infty} x_{2(n)}(t)e^n \\ \displaystyle\sum_{n=0}^{\infty} y_{1(n)}(t)e^n & \displaystyle\sum_{n=0}^{\infty} y_{2(n)}(t)e^n \end{pmatrix}, \tag{9.23}$$

where $\begin{pmatrix} x_{1(0)}(t) & x_{2(0)}(t) \\ y_{1(0)}(t) & y_{2(0)}(t) \end{pmatrix}$ is the solution of the variational equations (9.9) of the circular Sitnikov problem along the solution curve $(z_{(0)}(t), Z_{(0)}(t))$ and

$$\begin{pmatrix} x_{1(n)}(t) & x_{2(n)}(t) \\ y_{1(n)}(t) & y_{2(n)}(t) \end{pmatrix} = \begin{pmatrix} \dfrac{\partial^n}{\partial e^n}\left(\dfrac{\partial z}{\partial z_0}\right)(t)\Big|_{e=0} & \dfrac{\partial^n}{\partial e^n}\left(\dfrac{\partial z}{\partial Z_0}\right)(t)\Big|_{e=0} \\ \dfrac{\partial^n}{\partial e^n}\left(\dfrac{\partial Z}{\partial z_0}\right)(t)\Big|_{e=0} & \dfrac{\partial^n}{\partial e^n}\left(\dfrac{\partial Z}{\partial Z_0}\right)(t)\Big|_{e=0} \end{pmatrix}.$$

We remark that the solution of the variational equations (9.9) of the circular Sitnikov problem along the solution curve $(z_{(0)}(t), Z_{(0)}(t))$ is unbounded when t goes to infinity. Therefore, with a fixed value of e, (9.23) is valid only for t less than a constant which depends on the value of e.

10. Continuation of periodic orbits from the reduced circular Sitnikov problem to the reduced isosceles problem. In this section we will use the analytic continuation method of Poincaré to continue the periodic orbits of the reduced circular Sitnikov problem to symmetric periodic orbits of the reduced isosceles problem for $\mu > 0$ sufficiently small.

Choosing conveniently the origin of time, the periodic orbit of the reduced circular Sitnikov problem with period $T > \pi/\sqrt{2}$ is the orbit associated to the periodic solution with initial conditions $\varphi_{1/4}(t; r_0 = 1/2, \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{z}_0^* = \sqrt{2h+4}, \mu = 0)$. Here $h \in (-2, 0)$ is the energy of the periodic orbit of period $T \in (\pi/\sqrt{2}, \infty)$ (see Theorem 7.5 for details). We remark that the notation used here is the one defined in section 8.

Since the reduced isosceles problem is autonomous, if we continue using different initial conditions defining the same periodic orbit, then we will obtain the same continued periodic orbits. So, it will be sufficient to continue periodic solutions with initial conditions $\varphi_{1/4}(t; 1/2, 0, 0, \dot{z}_0^*, 0)$ for $-2 < h < 0$. We note that these periodic solutions are doubly symmetric, so we can investigate their continuation to periodic solutions of the reduced isosceles problem for $\mu > 0$ small that are either doubly symmetric, r-symmetric, or t-symmetric. Here we analyze only the continuation to doubly symmetric periodic solutions. We have also analyzed the continuation to r- and to t-symmetric periodic solutions, but these two types of continuation provide again the same families of doubly symmetric periodic orbits of the reduced isosceles problem for $\mu > 0$ small (for details see [6]).

By Proposition 5.3(1), if we can find initial conditions r_0 and $\dot{z}_0$ such that the solution $\varphi_{1/4}(t; r_0, 0, 0, \dot{z}_0, \mu) = (r(t; r_0, \dot{z}_0, \mu), \dot{r}(t; r_0, \dot{z}_0, \mu), z(t; r_0, \dot{z}_0, \mu), \dot{z}(t; r_0, \dot{z}_0, \mu))$ of the reduced isosceles problem (3.1) with $c = 1/4$ satisfies

$$\dot{r}(\tau/4; r_0, \dot{z}_0, \mu) = 0, \qquad \dot{z}(\tau/4; r_0, \dot{z}_0, \mu) = 0, \tag{10.1}$$

and $\dot{r}$,$\dot{z}$ are not simultaneously zero for $t \in (0, \tau/4)$, then $\varphi_{1/4}(t; r_0, 0, 0, \dot{z}_0, \mu)$ is a doubly symmetric periodic solution with period τ.

Observe that $\tau = T = T(h)$, $r_0 = 1/2$, $\dot{z}_0 = \dot{z}_0^* = \sqrt{2h+4}$, and $\mu = 0$ is a solution of (10.1) for each $-2 < h < 0$. It corresponds to the doubly symmetric periodic solution $\varphi_{1/4}(t; 1/2, 0, 0, \dot{z}_0^*, 0)$ of the reduced circular Sitnikov problem. Our aim is to find solutions of (10.1) near the known solution $\tau = T$, $r_0 = 1/2$, $\dot{z}_0 = \dot{z}_0^*$, and $\mu = 0$. For this goal, we will apply the implicit function theorem to (10.1) in a neighborhood of that point, choosing $(r_0, \dot{z}_0)$ as the dependent variables and (μ, τ) as the independent ones.

We note that there are five other choices for the dependent (independent) variables. Since we want to continue periodic solutions from $\mu = 0$ to $\mu > 0$ small, we are interested in solutions of (10.1) depending on μ. So, the other possible choices for the independent variables are (μ, r_0) and $(\mu, \dot{z}_0)$. Since the reduced isosceles problem possesses the first integral of the energy, we also could be interested in expressing the solutions of (10.1) as a function of μ and of the energy $\widetilde{h}$. We have analyzed these other possible choices for the independent variables, then saw that the implicit function theorem using either (μ, r_0) or $(\mu, \widetilde{h})$ as the independent variables cannot be applied to this problem because the corresponding determinant vanishes. Moreover, if we apply the implicit function theorem using $(\mu, \dot{z}_0)$ as the independent variables, we obtain the same solutions of (10.1) as we do using (μ, τ). The difference is that these solutions are parameterized by $(\mu, \dot{z}_0)$ instead of (μ, τ).

We apply the implicit function theorem to system (10.1) in a neighborhood of the point $\tau = T$, $r_0 = 1/2$, $\dot{z}_0 = \dot{z}_0^*$, and $\mu = 0$, choosing μ and τ as the independent variables. If

$$\det \begin{pmatrix} \dfrac{\partial \dot{r}(\tau/4; r_0, \dot{z}_0, \mu)}{\partial r_0} & \dfrac{\partial \dot{r}(\tau/4; r_0, \dot{z}_0, \mu)}{\partial \dot{z}_0} \\ \dfrac{\partial \dot{z}(\tau/4; r_0, \dot{z}_0, \mu)}{\partial r_0} & \dfrac{\partial \dot{z}(\tau/4; r_0, \dot{z}_0, \mu)}{\partial \dot{z}_0} \end{pmatrix}_{\big|(\mu=0, \tau=T, r_0=1/2, \dot{z}_0=\dot{z}_0^*)} \neq 0, \tag{10.2}$$

then for each (μ, τ) in a sufficiently small neighborhood of $(0, T)$, there exist two unique functions $r_0 = r_0(\mu, \tau)$ and $\dot{z}_0 = \dot{z}_0(\mu, \tau)$ such that $r_0(0, T) = 1/2$, $\dot{z}_0(0, T) = \dot{z}_0^*$, and r_0, $\dot{z}_0$ satisfy system (10.1). We note that the negative values of μ do not have physical meaning. Therefore, if condition (10.2) is satisfied, then for each (μ, τ) in a sufficiently small neighborhood of $(0, T)$ with $\mu \geqslant 0$, $\varphi_{1/4}(t; r_0(\mu, \tau), 0, 0, \dot{z}_0(\mu, \tau), \mu)$ is a doubly symmetric periodic solution of the reduced isosceles problem (3.1) for $c = 1/4$ with period τ. Since the functions that appear in system (10.1) are analytic, the functions $r_0(\mu, \tau)$ and $\dot{z}_0(\mu, \tau)$ are also analytic and may be expanded in power series of μ and $\overline{\tau} = \tau - T$ in U, a sufficiently small neighborhood of $(0, 0)$; that is, $r_0 = 1/2 + O(\mu, \overline{\tau})$ and $\dot{z}_0 = \dot{z}_0^* + O(\mu, \overline{\tau})$.

Now we compute the value of the determinant (10.2). The derivatives that appear in this determinant are obtained by evaluating at time $t = T/4$ the corresponding solution of the variational equations of the reduced restricted isosceles problem (6.2) for $c = 1/4$ along the solution curve $\varphi_{1/4}(t; 1/2, 0, 0, \dot{z}_0^*, 0)$. These variational equations were solved in section 9. Then, from (9.8),

$$\left.\frac{\partial \dot{r}(\tau/4; r_0, \dot{z}_0, \mu)}{\partial \dot{z}_0}\right|_{(\mu=0, \tau=T, r_0=1/2, \dot{z}_0=\dot{z}_0^*)} = 0.$$

The value of the derivative $\partial \dot{r}(\tau/4; r_0, \dot{z}_0, \mu)/\partial r_0$ evaluated at $\mu = 0$, $\tau = T$, $r_0 = 1/2$, and $\dot{z}_0 = \dot{z}_0^*$ can be obtained by evaluating at $t = T/4$ the corresponding solution of the variational equations of the Kepler problem (9.2), with $e = 0$, along the solution curve $(r(t) = 1/2, \dot{r}(t) = 0)$. Thus, by Proposition 9.2, we get

$$\left.\frac{\partial \dot{r}(\tau/4; r_0, \dot{z}_0, \mu)}{\partial r_0}\right|_{(\mu=0 \tau=T, r_0=1/2, \dot{z}_0=\dot{z}_0^*)} = -\sin(T/4),$$

which is different from zero if and only if the period T is a nonmultiple of 4π.

It remains only to find the value of $\partial \dot{z}(\tau/4; r_0, \dot{z}_0, \mu)/\partial \dot{z}_0$ at $\mu = 0$, $\tau = T$, $r_0 = 1/2$, and $\dot{z}_0 = \dot{z}_0^*$. This value can be obtained by evaluating at $t = T/4$ the

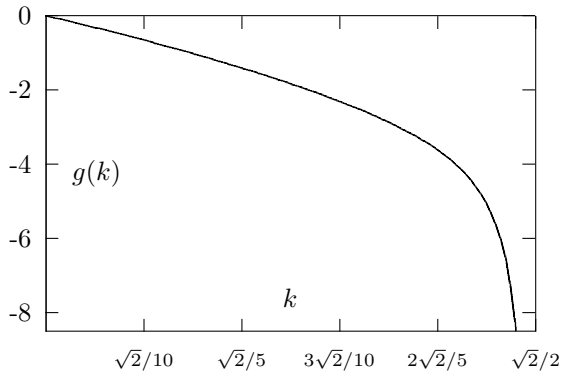


FIG. 10.1. *The graphic of* $g(k)$.

corresponding solution of the variational equations of the circular Sitnikov problem along the solution curve $(z(t;1/2,0,0,\dot{z}_0^*,0),\dot{z}(t;1/2,0,0,\dot{z}_0^*,0))$. The solution of those variational equations is given by formula (B.12) of [8]. In particular, the derivative $\partial\dot{z}(t;r_0,\dot{z}_0,\mu)/\partial\dot{z}_0$ evaluated at $\mu=0$, $r_0=1/2$, and $\dot{z}_0=\dot{z}_0^*$ is

$$
\begin{aligned}
&\frac{(1-2k^2\,\mathrm{sn}^2\nu)^2}{(2k^2-1)^2k'^2}\Bigg[-k^2\,\mathrm{sn}^2\nu\,\mathrm{cn}\,\nu+\mathrm{dn}^2\nu\,\mathrm{cn}\,\nu-\mathrm{sn}\,\nu\,\mathrm{dn}\,\nu\Bigg(k'^2(k^2+1)\nu\\
&-(2k^2k'^2+1)E(\nu)-3k^2k'^2\Pi(\nu,2k^2)+4k^4k'^2\frac{\mathrm{sn}\,\nu\,\mathrm{cn}\,\nu\,\mathrm{dn}\,\nu}{1-2k^2\,\mathrm{sn}^2\nu}\Bigg)\\
&+\mathrm{cn}\,\nu\Bigg(k'^2(k^2+1)-(2k^2k'^2+1)\,\mathrm{dn}^2\nu-\frac{3k^2k'^2}{1-2k^2\,\mathrm{sn}^2\nu}\\
&+4k^4k'^2\frac{(\,\mathrm{cn}^2\nu\,\mathrm{dn}^2\nu-\mathrm{sn}^2\nu\,\mathrm{dn}^2\nu-k^2\,\mathrm{sn}^2\nu\,\mathrm{cn}^2\nu)}{1-2k^2\,\mathrm{sn}^2\nu}\\
&+16k^6k'^2\frac{\mathrm{sn}^2\nu\,\mathrm{cn}^2\nu\,\mathrm{dn}^2\nu}{(1-2k^2\,\mathrm{sn}^2\nu)^2}\Bigg)\Bigg],
\end{aligned}
\tag{10.3}
$$

where ν is a function of t via Lemma 7.2(1), $k=\sqrt{2+h}/2$, and $k'=\sqrt{1-k^2}$.

By Lemmas 7.2(1) and 7.4, we have that $\nu(0)=0$ and $\nu(T/4)=K$, respectively. Then, by formula 122.02 of [4] we have that $\mathrm{sn}\,\nu(T/4)=1$, $\mathrm{cn}\,\nu(T/4)=0$, $\mathrm{dn}\,\nu(T/4)=k'$, and by formula (A.5) of [8] we have that $E(\nu(T/4))=E$ and $\Pi(\nu(T/4),2k^2)=\Pi(2k^2,k)$. Therefore,

$$
\left.\frac{\partial\dot{z}(\tau/4;r_0,\dot{z}_0,\mu)}{\partial\dot{z}_0}\right|_{(\mu=0,\tau=T,r_0=1/2,\dot{z}_0=\dot{z}_0^*)}=-\frac{1}{k'}g(k),
$$

where

$$
g(k)=k'^2(k^2+1)K-(2k^2k'^2+1)E-3k^2k'^2\Pi(2k^2,k).
\tag{10.4}
$$

Since $-2<h<0$, we have that $k\in(0,\sqrt{2}/2)$. We plot the function $g(k)$ in the range $0<k<\sqrt{2}/2$, obtaining Figure 10.1. Therefore $g(k)$ is always different from zero except when $k=0$, but this case is not considered here because it corresponds to the equilibrium point of the reduced circular Sitnikov problem.

In short, if the period $T=T(h)$ is a nonmultiple of 4π, then determinant (10.2) is different from zero. This proves the following theorem.

THEOREM 10.1. *For any $T > \pi/\sqrt{2}$, with $T \neq 4\pi n$ for all $n \in \mathbb{N}$, the periodic orbit of the reduced circular Sitnikov problem with period T can be continued to a 2-parameter family (on μ and τ) of doubly symmetric periodic orbits of the reduced isosceles problem* (3.1) *with angular momentum $c = 1/4$, which have period τ for (μ, τ) in a sufficiently small neighborhood of $(0, T)$ with $\mu \geqslant 0$.*

10.1. Remarks. We note that Theorem 10.1 also gives periodic orbits of the reduced isosceles problem for $\mu = 0$. One might think that this theorem could be used to find new symmetric periodic orbits of the reduced elliptic Sitnikov problem. But this is not the case because the symmetric periodic orbits for $\mu = 0$ that we obtain in this way are periodic orbits of the reduced circular Sitnikov problem, which are already known. This follows from the fact that the functions $r_0(\mu, \tau)$ and $\dot{z}_0(\mu, \tau)$ are unique and that $\varphi_{1/4}(t; r_0 = 1/2, 0, 0, \dot{z}_0 = \sqrt{2h(\tau) + 4}, 0)$ is a periodic solution of the reduced circular Sitnikov problem.

On the other hand, Theorem 10.1 does not allow us to continue the periodic orbits of the reduced circular Sitnikov problem that have period T that is a multiple of 4π. Later on, in section 12, we will see that these periodic solutions can be continued in two steps to two different families of doubly symmetric periodic solutions of the reduced isosceles problem (3.1) with angular momentum $c = 1/4$ and $\mu > 0$ sufficiently small, having period τ near T (see Theorem 12.8). The fact that the continuation is to two families explains why we have not been able here to continue these periodic orbits using only the implicit function theorem.

Often when we analyze a problem of continuation of periodic solutions we are interested in families of periodic solutions with the same period or with the same energy (these last families are called *isoenergetic families*). We could also consider families of periodic solutions with a fixed initial condition. In order to obtain these kinds of families in our problem we would fix one of the variables (it could be T, $\widetilde{h}$, r_0, or $\dot{z}_0$) in system (10.1), and then we would continue, in function of μ, the known periodic solutions of the reduced circular Sitnikov problem. We have done that and seen that the periodic solutions of the reduced circular Sitnikov problem with period T, nonmultiple of 4π, can be continued to a 1-parameter family (on μ) of doubly symmetric periodic solutions of the reduced isosceles problem having fixed period T, and another 1-parameter family having fixed initial condition $\dot{z}_0 = \dot{z}_0^*$. Clearly these two families are contained in the 2-parameter family of doubly symmetric periodic orbits of the reduced isosceles problem obtained in Theorem 10.1. Finally, the continuation fixing either the initial condition r_0 or the energy $\widetilde{h}$ is not possible because the corresponding determinants vanish.

Theorem 10.1 is improved by the following result.

THEOREM 10.2. *For any interval $[T_1, T_2]$ with $T_1 > \pi/\sqrt{2}$ and such that $4\pi n \notin [T_1, T_2]$ for all $n \in \mathbb{N}$, there exist $\mu_0 > 0$ and two unique analytic functions $r_0(\mu, \tau)$ and $\dot{z}_0(\mu, \tau)$, defined for all $\mu \in [0, \mu_0)$ and $\tau \in [T_1, T_2]$, such that $\varphi_{1/4}(t; r_0(\mu, \tau), 0, 0, \dot{z}_0(\mu, \tau), \mu)$ is a double symmetric periodic solution, with period τ, of the reduced isosceles problem* (3.1) *with angular momentum $c = 1/4$. Moreover $r_0(0, \tau) = 1/2$ and $\dot{z}_0(0, \tau) = \sqrt{2h(\tau) + 4}$, where $h(\tau)$ is the value of the energy of the periodic orbit of the circular Sitnikov problem having period τ.*

Proof. Fixed $\tau^* \in [T_1, T_2]$, the implicit function theorem assures the existence of two unique analytic functions $r_0(\mu, \tau)$ and $\dot{z}_0(\mu, \tau)$ for (μ, τ) in a sufficiently small neighborhood of $(0, \tau^*)$. Due to the compactness of $[T_1, T_2]$ and the uniqueness of $r_0(\mu, \tau)$ and $\dot{z}_0(\mu, \tau)$, we can find $\mu_0 > 0$ such that, for $0 \leqslant \mu < \mu_0$, these functions are defined for all $\tau^* \in [T_1, T_2]$, which proves the result. □

11. Continuation of symmetric periodic orbits from the reduced elliptic Sitnikov problem to the reduced isosceles problem. In this section we will continue the known symmetric periodic solutions of the reduced elliptic Sitnikov problem with eccentricity e (meaning the symmetric periodic solutions given in section 8) to symmetric periodic solutions of the reduced isosceles problem with $c = c_e$ and $\mu > 0$ sufficiently small. In particular we will prove the following result.

THEOREM 11.1. *Let γ_{e_0} be a symmetric periodic orbit of the reduced elliptic Sitnikov problem with eccentricity e_0 given by Theorems* 8.2 *or* 8.3 *that has period $\tau^\diamond = 2\pi p = qT$ for fixed values of $p, q \in \mathbb{N}$ coprime with $p > q/(2\sqrt{2})$. If the eccentricity e_0 is sufficiently small, then γ_{e_0} can be continued to a 2-parameter family (on μ and τ) of symmetric periodic orbits of the reduced isosceles problem* (3.1) *with angular momentum $c = \sqrt{1-e_0^2}/4$ and $\mu \geqslant 0$ that have period τ for (μ, τ) in a sufficiently small neighborhood of $(0, \tau^\diamond)$. Moreover the continued periodic orbits satisfy the same symmetry as the initial orbit γ_{e_0}.*

Apart from the symmetric periodic orbits of the reduced elliptic Sitnikov problem given by Theorems 8.2 and 8.3, we know the existence of infinitely many symmetric periodic orbits of the reduced elliptic Sitnikov problem for all $0 < e < 1$ (see Propositions 12 and 15 in [7]); unfortunately we do not know analytical expressions for their initial conditions. Nevertheless we will give sufficient conditions in order to continue an arbitrary symmetric periodic orbit of the reduced elliptic Sitnikov problem to symmetric periodic orbits of the reduced isosceles problem for $\mu > 0$ sufficiently small.

We start analyzing the continuation of doubly symmetric periodic orbits of the reduced elliptic Sitnikov problem, after which we will analyze the continuation of r- and t-symmetric periodic orbits.

Choosing conveniently the origin of time, the symmetric periodic orbits of the reduced elliptic Sitnikov problem can be seen as the orbits associated to symmetric periodic solutions with initial conditions either $\varphi_{c_e}(t; r_0 = r_0^\diamond = (1 \pm e)/2, \dot r_0 = 0, z_0 = 0, \dot z_0 = \dot z_0^\diamond, \mu = 0)$ or $\varphi_{c_e}(t; r_0 = r_0^\diamond = (1 \pm e)/2, \dot r_0 = 0, z_0 = z_0^\diamond, \dot z_0 = 0, \mu = 0)$. So, we will study only the continuation of symmetric periodic solutions of these types. Of course, if we continue different initial conditions defining the same periodic orbit, then we will obtain the same periodic orbit of the reduced isosceles problem.

11.1. Continuation of doubly symmetric periodic solutions. As in section 10, by Proposition 5.3(1), the solution $\varphi_{c_e}(t; r_0, 0, 0, \dot z_0, \mu) = (r(t; r_0, \dot z_0, \mu), \dot r(t; r_0, \dot z_0, \mu), z(t; r_0, \dot z_0, \mu), \dot z(t; r_0, \dot z_0, \mu))$ is a doubly symmetric periodic solution of the reduced isosceles problem (3.1) with $c = c_e$ having period τ if it satisfies

$$\dot r(\tau/4; r_0, \dot z_0, \mu) = 0, \qquad \dot z(\tau/4; r_0, \dot z_0, \mu) = 0, \tag{11.1}$$

and $\dot r$, $\dot z$ are not simultaneously zero for $t \in (0, \tau/4)$.

Let $\varphi_{c_e}(t; r_0 = r_0^\diamond = (1 \pm e)/2, \dot r_0 = 0, z_0 = 0, \dot z_0 = \dot z_0^\diamond, \mu = 0)$ be a doubly symmetric periodic solution of the reduced elliptic Sitnikov problem for a fixed $0 < e < 1$ and let $\tau^\diamond = 2\pi p$, with $p \in \mathbb{N}$ even, be its period. This is equivalent to saying that $\tau = \tau^\diamond$, $r_0 = r_0^\diamond$, $\dot z_0 = \dot z_0^\diamond$, and $\mu = 0$ is a solution of system (11.1).

Applying the implicit function theorem to (11.1) in a neighborhood of the point $\tau = \tau^\diamond$, $r_0 = r_0^\diamond$, $\dot z_0 = \dot z_0^\diamond$, and $\mu = 0$, and choosing μ and τ as the independent variables, if

$$\det \left(\begin{array}{cc} \dfrac{\partial \dot r(\tau/4; r_0, \dot z_0, \mu)}{\partial r_0} & \dfrac{\partial \dot r(\tau/4; r_0, \dot z_0, \mu)}{\partial \dot z_0} \\ \dfrac{\partial \dot z(\tau/4; r_0, \dot z_0, \mu)}{\partial r_0} & \dfrac{\partial \dot z(\tau/4; r_0, \dot z_0, \mu)}{\partial \dot z_0} \end{array} \right)_{\big|(\mu=0, \tau=\tau^\diamond, r_0=r_0^\diamond, \dot z_0=\dot z_0^\diamond)} \neq 0,$$

then for each $(\mu, \overline{\tau} = \tau - \tau^\diamond)$ in a sufficiently small neighborhood W_d of $(0,0)$ with $\mu \geqslant 0$, we can find two unique analytic functions $r_0(\mu,\tau) = r_0^\diamond + O(\mu,\overline{\tau})$ and $\dot{z}_0(\mu,\tau) = \dot{z}_0^\diamond + O(\mu,\overline{\tau})$ such that $\varphi_{c_e}(t; r_0^\diamond + O(\mu,\overline{\tau}), 0, 0, \dot{z}_0^\diamond + O(\mu,\overline{\tau}), \mu)$ is a doubly symmetric periodic solution of period τ for the reduced isosceles problem (3.1) for $c = c_e$ and $\mu \geqslant 0$ small enough.

The derivatives that appear in this determinant are obtained by evaluating at time $t = \tau^\diamond/4$ the corresponding solution of the variational equations of the reduced restricted isosceles problem (6.2) for $c = c_e$ along the solution curve $\varphi_{c_e}(t; r_0^\diamond, 0, 0, \dot{z}_0^\diamond, 0)$ with $r_0^\diamond = (1 \pm e)/2$. The solution of these variational equations has been studied in section 9.1.

By (9.8), the derivative $\partial \dot{r}(\tau/4; r_0, \dot{z}_0, \mu)/\partial \dot{z}_0$ evaluated at $\mu = 0$, $\tau = \tau^\diamond$, $r_0 = r_0^\diamond$, and $\dot{z}_0 = \dot{z}_0^\diamond$ equals zero. The value of the derivative $\partial \dot{r}(\tau/4; r_0, \dot{z}_0, \mu)/\partial r_0$ evaluated at $\mu = 0$, $\tau = \tau^\diamond$, $r_0 = r_0^\diamond$, and $\dot{z}_0 = \dot{z}_0^\diamond$ can be obtained by evaluating at $t = T/4$ the corresponding solution of the variational equations of the Kepler problem (9.2) along the solution curve $(r(t; r_0^\diamond, \dot{z}_0^\diamond, 0), \dot{r}(t; r_0^\diamond, \dot{z}_0^\diamond, 0))$.

If $r_0^\diamond = (1-e)/2$—that is, $t = 0$ corresponds to the minimum value of $r(t; r_0^\diamond, \dot{z}_0^\diamond, 0)$—then from Kepler's equation (9.18), $u(\tau^\diamond/4) = u(m\pi) = m\pi$. Moreover, since p is even, $p = 2m$ for some $m \in \mathbb{N}$. Therefore, from (9.19),

$$\left.\frac{\partial \dot{r}(\tau/4; r_0, \dot{z}_0, \mu)}{\partial r_0}\right|_{(\mu=0, \tau=\tau^\diamond, r_0=\frac{1-e}{2}, \dot{z}_0=\dot{z}_0^\diamond)} = \frac{3e^2 m\pi((-1)^m - e)}{(1-(-1)^m e)^3(1-e)^2}, \tag{11.2}$$

which is different from zero because $e \neq 0$ and $e \neq 1$.

If $r_0^\diamond = (1+e)/2$—that is, $t = 0$ corresponds to the maximum value of $r(t; r_0^\diamond, \dot{z}_0^\diamond, 0)$—then from Kepler's equation (9.20), $u(\tau^\diamond/4) = u(m\pi) = (m+1)\pi$. Thus by (9.21),

$$\left.\frac{\partial \dot{r}(\tau/4; r_0, \dot{z}_0, \mu)}{\partial r_0}\right|_{(\mu=0, \tau=\tau^\diamond, r_0=\frac{1+e}{2}, \dot{z}_0=\dot{z}_0^\diamond)} = \frac{3e^2 m\pi(e - (-1)^{m+1})}{(1-(-1)^{m+1} e)^3(1+e)^2}, \tag{11.3}$$

which is also different from zero. In short, we have proved the following result.

THEOREM 11.2. *Let $\varphi_{c_e}(t; r_0 = r_0^\diamond = (1 \pm e)/2, \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{z}_0^\diamond, \mu = 0)$ be a doubly symmetric periodic solution of the reduced elliptic Sitnikov problem for a fixed $0 < e < 1$ and let $\tau^\diamond = 2\pi p$ with $p \in \mathbb{N}$ even be its period. If*

$$\left.\frac{\partial \dot{z}(\tau/4; r_0, \dot{z}_0, \mu)}{\partial \dot{z}_0}\right|_{(\mu=0, \tau=\tau^\diamond, r_0=r_0^\diamond, \dot{z}_0=\dot{z}_0)} \neq 0, \tag{11.4}$$

then this solution can be analytically continued to a 2-parameter family (on μ and τ) $\varphi_{c_e}(t; r_0 = r_0^\diamond + O(\mu,\overline{\tau}), \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{z}_0^\diamond + O(\mu,\overline{\tau}), \mu)$ of doubly symmetric periodic solutions of the reduced isosceles problem, with angular momentum $c = c_e$ and $\mu \geqslant 0$, that have period τ for $(\mu, \overline{\tau} = \tau - \tau^\diamond) \in W_d$, with W_d a sufficiently small neighborhood of $(0,0)$.

11.1.1. Application of Theorem 11.2. Now we apply Theorem 11.2 to continue the doubly symmetric periodic solutions of the reduced elliptic Sitnikov problem given by Theorem 8.2. Let $\varphi_{c_e}(t; r_0 = r_0^\diamond = (1 \pm e)/2, \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{z}_0^\diamond = \dot{z}_0^* + O(e), \mu = 0)$, with $\dot{z}_0^* = \pm\sqrt{2h+4}$, be one of these periodic solutions for a fixed $e > 0$ sufficiently small and fixed $p, q \in \mathbb{N}$ coprime with p even and $p > q/(2\sqrt{2})$. By Theorem 11.2, this doubly symmetric periodic solution can be continued to doubly symmetric periodic solutions of the reduced isosceles problem for $\mu > 0$ if (11.4) holds. The value of the derivative (11.4) is obtained from the solution, evaluated at $t = \tau^\diamond/4$,

of the variational equations of the elliptic Sitnikov problem (9.3) along the solution curve $(z(t),\dot z(t)) = (z(t;r_0^\diamond,\dot z_0=\dot z_0^\diamond=\dot z_0^*+O(e),0),\dot z(t;r_0^\diamond,\dot z_0=\dot z_0^\diamond=\dot z_0^*+O(e),0))$. We note that if the eccentricity e is sufficiently small, then by the Poincaré expansion theorem, the solution $(z(t),\dot z(t))$ may be expanded in power series of $\dot z_0^\diamond-\dot z_0^*$ and e. Since $\dot z_0^\diamond-\dot z_0^*=O(e)$, we have that $(z(t),\dot z(t)) = (z_{(0)}(t)+O(e),\dot z_{(0)}(t)+O(e))$, where $(z_{(0)}(t),\dot z_{(0)}(t))$ is the solution of the circular Sitnikov problem with initial conditions $z_{(0)}(0)=0$ and $\dot z_{(0)}(0)=\dot z_0^*$. So, the solution of the variational equations of the elliptic Sitnikov problem along that solution curve $(z(t),\dot z(t))$ is given by the solution of the variational equations of the reduced circular Sitnikov problem along the solution curve $(z_{(0)}(t),\dot z_{(0)}(t))$ plus terms of at least order one in e (see section 9.3). Since $\dot z_0^* = \pm\sqrt{2h+4}$, the solution of these last variational equations is given by formula (B.12) of [8].

We assume that e is small enough so that (9.23) is valid at least for $0\leqslant t\leqslant \tau^\diamond/4$. From formula (B.12) of [8] and (9.23), the derivative $\partial\dot z(t;r_0,\dot z_0,\mu)/\partial\dot z_0$ evaluated at $\mu=0$, $r_0=r_0^\diamond$, and $\dot z_0=\dot z_0^\diamond$ is given by (10.3) plus terms of at least order one in e. On the other hand, from Lemmas 7.2(1) and 7.4, we have that $\nu(0)=0$ and $\nu(\tau/4)=qK$, respectively. We consider that $q=2l+1$ for some $l\in\mathbb{N}$ (we note that q is odd because p is even and p and q are coprime). By formulas 122.02 and 122.04 of [4] we have that $\operatorname{sn}\nu(\tau/4)=(-1)^l$, $\operatorname{cn}\nu(\tau/4)=0$, $\operatorname{dn}\nu(\tau/4)=k'$; moreover by formula (A.5) of [8] we have that $E(\nu(\tau/4))=qE$ and $\Pi(\nu(\tau/4),2k^2)=q\Pi(2k^2,k)$. Hence

$$\left.\frac{\partial\dot z(\tau/4;r_0,\dot z_0,\mu)}{\partial\dot z_0}\right|_{(\mu=0,\tau=\tau^\diamond,r_0=r_0^\diamond,\dot z_0=\dot z_0^\diamond)} = -\frac{(-1)^l q}{k'}g(k)+O(e),$$

where $l\in\mathbb{N}$ is such that $q=2l+1$, and $g(k)$ is given by (10.4). Since $g(k)$ is always different from zero, if the eccentricity e is small enough, then the derivative $\partial\dot z(\tau/4;r_0,\dot z_0,\mu)/\partial\dot z_0$ evaluated at $\mu=0$, $\tau=\tau^\diamond$, $r_0=r_0^\diamond$, and $\dot z_0=\dot z_0^\diamond$ is different from zero. Thus we have the following result.

COROLLARY 11.3. *For fixed $e>0$ sufficiently small, let $\varphi_{c_e}(t;r_0=r_0^\diamond=(1\pm e)/2,\dot r_0=0,z_0=0,\dot z_0=\dot z_0^\diamond=\pm\sqrt{2h+4}+O(e),\mu=0)$ be one of the doubly symmetric periodic solutions of the reduced elliptic Sitnikov problem given by Theorem 8.2 that has period $\tau^\diamond=2\pi p=qT$ for given values of $p,q\in\mathbb{N}$ coprime with p even and $p>q/(2\sqrt{2})$. Then this solution can be analytically continued to a 2-parameter family (on μ and τ) $\varphi_{c_e}(t;r_0=r_0^\diamond+O(\mu,\overline\tau),\dot r_0=0,z_0=0,\dot z_0=\dot z_0^\diamond+O(\mu,\overline\tau),\mu)$ of doubly symmetric periodic solutions of the reduced isosceles problem, with angular momentum $c=c_e$ and $\mu\geqslant 0$, that have period τ for $(\mu,\overline\tau=\tau-\tau^\diamond)\in W_d$, with W_d a sufficiently small neighborhood of $(0,0)$.*

11.2. Continuation of r-symmetric periodic solutions. By Proposition 5.1, $\varphi_{c_e}(t;r_0,0,0,\dot z_0,\mu)$ is an r-symmetric periodic solution of the reduced isosceles problem (3.1) with $c=c_e$ having period τ if it satisfies

$$\dot r(\tau/2;r_0,\dot z_0,\mu)=0\ ,\qquad z(\tau/2;r_0,\dot z_0,\mu)=0, \tag{11.5}$$

and $\dot r$, z are not simultaneously zero for $t\in(0,\tau/2)$.

Let $\varphi_{c_e}(t;r_0=r_0^\diamond=(1\pm e)/2,\dot r_0=0,z_0=0,\dot z_0=\dot z_0^\diamond,\mu=0)$ be an r-symmetric periodic solution of the reduced elliptic Sitnikov problem for a fixed $0<e<1$ and let $\tau^\diamond=2\pi p$ with $p\in\mathbb{N}$ be its period. Or, equivalently, let $\tau=\tau^\diamond$, $r_0=r_0^\diamond$, $\dot z_0=\dot z_0^\diamond$, and $\mu=0$ be a solution of system (11.5). Applying the implicit function theorem to system (11.5) in a neighborhood of that solution, choosing μ and τ as the independent

variables, if

$$\det\left(\begin{array}{cc} \dfrac{\partial \dot r(\tau/2;r_0,\dot z_0,\mu)}{\partial r_0} & \dfrac{\partial \dot r(\tau/2;r_0,\dot z_0,\mu)}{\partial \dot z_0} \\ \dfrac{\partial z(\tau/2;r_0,\dot z_0,\mu)}{\partial r_0} & \dfrac{\partial z(\tau/2;r_0,\dot z_0,\mu)}{\partial \dot z_0} \end{array}\right)\Bigg|_{(\mu=0,\tau=\tau^\diamond,r_0=r_0^\diamond,\dot z_0=\dot z_0)} \neq 0,$$

then for each $(\mu,\overline{\tau}=\tau-\tau^\diamond)$ in a sufficiently small neighborhood W_r of $(0,0)$ with $\mu\geqslant 0$, we can find two unique analytic functions $r_0(\mu,\tau)=r_0^\diamond+O(\mu,\overline{\tau})$ and $\dot z_0(\mu,\tau)=\dot z_0^\diamond+O(\mu,\overline{\tau})$ such that $\varphi_{c_e}(t;r_0^\diamond+O(\mu,\overline{\tau}),0,0,\dot z_0^\diamond+O(\mu,\overline{\tau}),\mu)$ is an r-symmetric periodic solution of period τ for the reduced isosceles problem (3.1) for $c=c_e$ and $\mu\geqslant 0$ small.

The derivatives that appear in this determinant are obtained by evaluating at time $t=\tau^\diamond/2$ the corresponding solution of the variational equations of the reduced restricted isosceles problem (6.2) for $c=c_e$ along the solution curve $\varphi_{c_e}(t;r_0^\diamond,0,0,\dot z_0^\diamond,0)$. Thus from (9.8),

$$\frac{\partial \dot r(\tau/2;r_0,\dot z_0,\mu)}{\partial \dot z_0}\bigg|_{(\mu=0,\tau=\tau^\diamond,r_0=r_0^\diamond,\dot z_0=\dot z_0^\diamond)}=0. \tag{11.6}$$

On the other hand, if $r_0^\diamond=(1-e)/2$, then from (9.18), $u(\tau^\diamond/2)=u(p\pi)=p\pi$; and if $r_0^\diamond=(1+e)/2$, then from (9.20), $u(\tau^\diamond/2)=u(p\pi)=(p+1)\pi$. Therefore, taking p instead of m in (11.2) and (11.3), we have that the derivative $\partial\dot r(\tau/2;r_0,\dot z_0,\mu)/\partial r_0$ evaluated at $\mu=0$, $\tau=\tau^\diamond$, $\dot z_0=\dot z_0^\diamond$, and $r_0=(1-e)/2$ (respectively, $r_0=(1+e)/2$) is given by

$$\frac{3e^2p\pi((-1)^p-e)}{(1-(-1)^pe)^3(1-e)^2}\qquad\left(\text{respectively, }\frac{3e^2p\pi(e-(-1)^{p+1})}{(1-(-1)^{p+1}e)^3(1+e)^2}\right), \tag{11.7}$$

which is different from zero. In short, we have proved the following result.

THEOREM 11.4. *Let $\varphi_{c_e}(t;r_0=r_0^\diamond=(1\pm e)/2,\dot r_0=0,z_0=0,\dot z_0=\dot z_0^\diamond,\mu=0)$ be an r-symmetric periodic solution of the reduced elliptic Sitnikov problem for a fixed $0<e<1$ and let $\tau^\diamond=2\pi p$ with $p\in\mathbb{N}$ be its period. If*

$$\frac{\partial z(\tau/2;r_0,\dot z_0,\mu)}{\partial \dot z_0}\bigg|_{(\mu=0,\tau=\tau^\diamond,r_0=r_0^\diamond,\dot z_0=\dot z_0^\diamond)}\neq 0, \tag{11.8}$$

then this solution can be analytically continued to a 2-parameter family (on μ and τ) $\varphi_{c_e}(t;r_0=r_0^\diamond+O(\mu,\overline{\tau}),\dot r_0=0,z_0=0,\dot z_0=\dot z_0^\diamond+O(\mu,\overline{\tau}),\mu)$ of r-symmetric periodic solutions of the reduced isosceles problem, with angular momentum $c=c_e$ and $\mu\geqslant 0$, that have period τ for $(\mu,\overline{\tau}=\tau-\tau^\diamond)\in W_r$, with W_r a sufficiently small neighborhood of $(0,0)$.

11.2.1. Application of Theorem 11.4. Now let $\varphi_{c_e}(t;r_0=r_0^\diamond=(1\pm e)/2,\dot r_0=0,z_0=0,\dot z_0=\dot z_0^\diamond=\pm\sqrt{2h+4}+O(e),\mu=0)$ be one of the r-symmetric periodic solutions of the reduced elliptic Sitnikov problem given by Theorem 8.2 for fixed $e>0$ small and $\tau^\diamond=2\pi p=qT$ with $p,q\in\mathbb{N}$ coprime and $p>q/(2\sqrt{2})$. By Theorem 11.4, the r-symmetric periodic solution $\varphi_{c_e}(t;r_0^\diamond,0,0,\dot z_0^\diamond,0)$ can be continued if (11.8) holds. The value of the derivative (11.8) is obtained from the solution, evaluated at $t=\tau^\diamond/2$, of the variational equations of the elliptic Sitnikov problem (9.3) along the solution curve $(z(t),\dot z(t))=(z(t;r_0^\diamond,\dot z_0=\dot z_0^\diamond=\dot z_0^*+O(e),0),\dot z(t;r_0^\diamond,\dot z_0=\dot z_0^\diamond=\dot z_0^*+O(e),0))$. We have seen that if e is sufficiently small, then the solution of those variational

equations is given by the solution of the variational equations of the reduced circular Sitnikov problem along the solution curve $(z_{(0)}(t), \dot{z}_{(0)}(t))$ plus terms of at least order one in e, where $(z_{(0)}(t), \dot{z}_{(0)}(t))$ is the solution of the circular Sitnikov problem with initial conditions $z_{(0)}(0) = 0$, $\dot{z}_{(0)}(0) = \dot{z}_0^*$. Proceeding as in the continuation of doubly symmetric periodic solutions (see section 11.1.1), we can see that

$$\left.\frac{\partial z(\tau/2; r_0, \dot{z}_0, \mu)}{\partial \dot{z}_0}\right|_{(\mu=0, \tau=\tau^\diamond, r_0=r_0^\diamond, \dot{z}_0=\dot{z}_0^\diamond)} = \frac{(-1)^q q}{\sqrt{2}(2k^2-1)^2 k'^2} g(k) + O(e),$$

which is different from zero if the eccentricity e is small enough.

In short, if the eccentricity e is sufficiently small, then $\varphi_{c_e}(t; r_0^\diamond, 0, 0, \dot{z}_0^\diamond, 0)$ can be continued to a family $\varphi_{c_e}(t; r_0 = r_0^\diamond + O(\mu, \overline{\tau}), \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{z}_0^\diamond + O(\mu, \overline{\tau}), \mu)$ of r-symmetric periodic solutions of the reduced isosceles problem, with angular momentum $c = c_e$ and $\mu \geqslant 0$, that have period τ for $(\mu, \overline{\tau} = \tau - \tau^\diamond) \in W_r$.

We note that if p is even, then $\varphi_{c_e}(t; r_0^\diamond, 0, 0, \dot{z}_0^\diamond, 0)$ is a doubly symmetric periodic solution. Thus, if the eccentricity e is sufficiently small, then it can also be continued to a family $\varphi_{c_e}(t; r_0 = r_0^\diamond + O(\mu, \overline{\tau}), \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{z}_0^\diamond + O(\mu, \overline{\tau}), \mu)$ of doubly symmetric periodic solutions of the reduced isosceles problem, with $c = c_e$ and $\mu \geqslant 0$, that have period τ for $(\mu, \overline{\tau}) \in W_d$ (see Corollary 11.3). Due to the uniqueness of the functions $r_0(\mu, \tau)$ and $\dot{z}_0(\mu, \tau)$ given by the implicit function theorem we have that if p is even and $(\mu, \overline{\tau}) \in W_d \cap W_r$, then the r-symmetric periodic solutions $\varphi_{c_e}(t; r_0^\diamond + O(\mu, \overline{\tau}), 0, 0, \dot{z}_0^\diamond + O(\mu, \overline{\tau}), \mu)$ are doubly symmetric periodic solutions.

If p is odd, then $\varphi_{c_e}(t; r_0^\diamond, 0, 0, \dot{z}_0^\diamond, 0)$ is an r-symmetric periodic solution that is not doubly symmetric because $\dot{r}(\tau^\diamond/4, r_0^\diamond, \dot{z}_0^\diamond, 0) \neq 0$. So, $\dot{r}(\tau/4, r_0^\diamond + O(\mu, \overline{\tau}), \dot{z}_0^\diamond + O(\mu, \overline{\tau}), \mu) \neq 0$ for $(\mu, \overline{\tau} = \tau - \tau^\diamond)$ in a sufficiently small neighborhood of $(0, 0)$. Consequently the r-symmetric periodic solutions $\varphi_{c_e}(t; r_0^\diamond + O(\mu, \overline{\tau}), 0, 0, \dot{z}_0^\diamond + O(\mu, \overline{\tau}), \mu)$ are not doubly symmetric periodic solutions. In short, we have proved the following result.

THEOREM 11.5. *For fixed $e > 0$ sufficiently small, let $\varphi_{c_e}(t; r_0 = r_0^\diamond = (1 \pm e)/2, \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{z}_0^\diamond = \pm\sqrt{2h+4} + O(e), \mu = 0)$ be one of the r-symmetric periodic solutions of the reduced elliptic Sitnikov problem given by Theorem 8.2 that has period $\tau^\diamond = 2\pi p = qT$ for given values of $p, q \in \mathbb{N}$ coprime and $p > q/(2\sqrt{2})$.*

1. *This solution can be continued to a 2-parameter family (on μ and τ) $\varphi_{c_e}(t; r_0 = r_0^\diamond + O(\mu, \overline{\tau}), \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{z}_0^\diamond + O(\mu, \overline{\tau}), \mu)$ of r-symmetric periodic solutions of the reduced isosceles problem, with angular momentum $c = c_e$ and $\mu \geqslant 0$, that have period τ for $(\mu, \overline{\tau} = \tau - \tau^\diamond)$ in a sufficiently small neighborhood of $(0, 0)$.*
2. *If p is odd, then the r-symmetric periodic solutions $\varphi_{c_e}(t; r_0^\diamond + O(\mu, \overline{\tau}), 0, 0, \dot{z}_0^\diamond + O(\mu, \overline{\tau}), \mu)$ are not doubly symmetric, whereas if p is even, they are doubly symmetric.*

11.3. Continuation of t-symmetric periodic solutions. By Proposition 5.2, $\varphi_{c_e}(t; r_0, 0, z_0, 0, \mu) = (r(t; r_0, z_0, \mu), \dot{r}(t; r_0, z_0, \mu), z(t; r_0, z_0, \mu), \dot{z}(t; r_0, z_0, \mu))$ is a t-symmetric periodic solution of the reduced isosceles problem (3.1), with $c = c_e$ having period τ, if it satisfies

$$\dot{r}(\tau/2; r_0, z_0, \mu) = 0, \qquad \dot{z}(\tau/2; r_0, z_0, \mu) = 0, \tag{11.9}$$

and $\dot{r}$, $\dot{z}$ are not simultaneously zero for $t \in (0, \tau/2)$.

Let $\varphi_{c_e}(t; r_0 = r_0^\diamond = (1 \pm e)/2, \dot{r}_0 = 0, z_0 = z_0^\diamond, \dot{z}_0 = 0, \mu = 0)$ be a t-symmetric periodic solution of the reduced elliptic Sitnikov problem for a fixed $0 < e < 1$ and let

$\tau^\diamond = 2\pi p$ with $p \in \mathbb{N}$ be its period. That is, let $\tau = \tau^\diamond$, $r_0 = r_0^\diamond$, $z_0 = z_0^\diamond$, and $\mu = 0$ be a solution of system (11.9). Applying the implicit function theorem to system (11.9) in a neighborhood of that solution, choosing μ and τ as the independent variables, if

$$\det \left(\begin{array}{cc} \dfrac{\partial \dot r(\tau/2; r_0, z_0, \mu)}{\partial r_0} & \dfrac{\partial \dot r(\tau/2; r_0, z_0, \mu)}{\partial z_0} \\ \dfrac{\partial \dot z(\tau/2; r_0, z_0, \mu)}{\partial r_0} & \dfrac{\partial \dot z(\tau/2; r_0, z_0, \mu)}{\partial z_0} \end{array} \right)_{\big|(\mu=0, \tau=\tau^\diamond, r_0=r_0^\diamond, z_0=z_0^\diamond)} \neq 0,$$

then for each $(\mu, \overline{\tau} = \tau - \tau^\diamond)$ in a sufficiently small neighborhood W_t of $(0,0)$ with $\mu \geqslant 0$, we can find two unique analytic functions $r_0(\mu, \tau) = r_0^\diamond + O(\mu, \overline{\tau})$ and $z_0(\mu, \tau) = z_0^\diamond + O(\mu, \overline{\tau})$ such that $\varphi_{c_e}(t; r_0^\diamond + O(\mu, \overline{\tau}), 0, z_0^\diamond + O(\mu, \overline{\tau}), 0, \mu)$ is a t-symmetric periodic solution of period τ for the reduced isosceles problem (3.1) for $c = c_e$ and $\mu \geqslant 0$ small enough. The derivatives that appear in this determinant are obtained by evaluating at time $t = \tau^\diamond/2$ the corresponding solutions of the variational equations of the reduced restricted isosceles problem (6.2) for $c = c_e$ along the solution curve $\varphi_{c_e}(t; r_0^\diamond, 0, z_0^\diamond, 0, 0)$. The solution of these variational equations was studied in section 9. Since the first equation of (6.2) does not depend on z and $\dot z$, $r(t; r_0, z_0, 0)$ and $\dot r(t; r_0, z_0, 0)$ do not depend on the initial conditions $z(0; r_0, z_0, 0)$ and $\dot z(0; r_0, z_0, 0)$. So, using the computations made in section 11.2 (see (11.6) and (11.7)) we can prove the following result.

THEOREM 11.6. *Let $\varphi_{c_e}(t; r_0 = r_0^\diamond = (1 \pm e)/2, \dot r_0 = 0, z_0 = z_0^\diamond, \dot z_0 = 0, \mu = 0)$ be a t-symmetric periodic solution of the reduced elliptic Sitnikov problem for a fixed $0 < e < 1$ and let $\tau^\diamond = 2\pi p$ with $p \in \mathbb{N}$ be its period. If*

$$\frac{\partial \dot z(\tau/2; r_0, z_0, \mu)}{\partial z_0}\bigg|_{(\mu=0, \tau=\tau^\diamond, r_0=r_0^\diamond, z_0=z_0^\diamond)} \neq 0, \tag{11.10}$$

then this solution can be analytically continued to a 2-parameter family (on μ and τ) $\varphi_{c_e}(t; r_0 = r_0^\diamond + O(\mu, \overline{\tau}), \dot r_0 = 0, z_0 = z_0^\diamond + O(\mu, \overline{\tau}), \dot z_0 = 0, \mu)$ of t-symmetric periodic solutions of the reduced isosceles problem, with angular momentum $c = c_e$ and $\mu \geqslant 0$, that have period τ for $(\mu, \overline{\tau} = \tau - \tau^\diamond) \in W_t$, with W_t a sufficiently small neighborhood of $(0,0)$.

11.3.1. Application of Theorem 11.6. Now we apply Theorem 11.6 to continue the t-symmetric periodic solutions of the reduced elliptic Sitnikov problem given by Theorem 8.3. Let $\varphi_{c_e}(t; r_0 = r_0^\diamond = (1 \pm e)/2, \dot r_0 = 0, z_0 = z_0^\diamond = z_0^* + O(e), \dot z_0 = 0, \mu = 0)$, with $z_0^* = \pm\sqrt{\frac{1}{h^2} - \frac{1}{4}}$, be one of these periodic solutions for a fixed $e > 0$ sufficiently small and fixed $p, q \in \mathbb{N}$ coprime with $p > q/(2\sqrt{2})$. The t-symmetric periodic solution $\varphi_{c_e}(t; r_0^\diamond, 0, z_0^\diamond, 0, 0)$ can be continued if (11.10) holds. Proceeding in a similar way to that of sections 11.1 and 11.2, we can see that if the eccentricity e is sufficiently small, then

$$\frac{\partial \dot z(\tau/2; r_0, \dot z_0, \mu)}{\partial z_0}\bigg|_{(\mu=0, \tau=\tau^\diamond, r_0=r_0^\diamond, z_0=z_0^\diamond)} = (-1)^{q+1} 4\sqrt{2}(1 - 2k^2)^2 q\, g(k) + O(e) \neq 0.$$

In short, if the eccentricity e is sufficiently small, then $\varphi_{c_e}(t; r_0^\diamond, 0, z_0^\diamond, 0, 0)$ can be continued to a family $\varphi_{c_e}(t; r_0 = r_0^\diamond + O(\mu, \overline{\tau}), \dot r_0 = 0, z_0 = z_0^\diamond + O(\mu, \overline{\tau}), \dot z_0 = 0, \mu)$ of t-symmetric periodic solutions of the reduced isosceles problem, with angular momentum $c = c_e$ and $\mu \geqslant 0$, that have period τ for $(\mu, \overline{\tau} = \tau - \tau^\diamond) \in W_t$. Moreover, due to the uniqueness of the functions $r_0(\mu, \tau)$ and $z_0(\mu, \tau)$ given by

the implicit function theorem we can see that the t-symmetric periodic solutions $\varphi_{c_e}(t; r_0^\diamond + O(\mu,\overline{\tau}), 0, z_0^\diamond + O(\mu,\overline{\tau}), 0, \mu)$ are doubly symmetric when p is even, and they are not doubly symmetric when p is odd (see the arguments of section 11.2.1). Therefore we have proved the following result.

THEOREM 11.7. *For fixed $e > 0$ sufficiently small, let $\varphi_{c_e}(t; r_0 = r_0^\diamond = (1 \pm e)/2, \dot{r}_0 = 0, z_0 = z_0^\diamond = z_0^* + O(e), \dot{z}_0 = 0, \mu = 0)$, with $z_0^* = \pm\sqrt{\frac{1}{h^2} - \frac{1}{4}}$, be one of the t-symmetric periodic solutions of the reduced elliptic Sitnikov problem given by Theorem 8.3 that has period $\tau^\diamond = 2\pi p = qT$ for given values of $p, q \in \mathbb{N}$ coprime and $p > q/(2\sqrt{2})$.*

1. *This solution can be continued to a 2-parameter family (on μ and τ) $\varphi_{c_e}(t; r_0 = r_0^\diamond + O(\mu,\overline{\tau}), \dot{r}_0 = 0, z_0 = z_0^\diamond + O(\mu,\overline{\tau}), \dot{z}_0 = 0, \mu)$ of t-symmetric periodic solutions of the reduced isosceles problem, with angular momentum $c = c_e$ and $\mu \geqslant 0$, that have period τ for $(\mu, \overline{\tau} = \tau - \tau^\diamond)$ in a sufficiently small neighborhood of $(0,0)$.*
2. *If p is odd, then the t-symmetric periodic solutions $\varphi_{c_e}(t; r_0^\diamond + O(\mu,\overline{\tau}), 0, z_0^\diamond + O(\mu,\overline{\tau}), 0, \mu)$ are not doubly symmetric, whereas if p is even, they are doubly symmetric.*

11.4. Remarks. In Theorems 11.5 and 11.7, we continued eight periodic solutions of the reduced elliptic Sitnikov problem: $\varphi_{c_e}(t; r_0 = r_0^\diamond = \frac{(1\pm e)}{2}, \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \pm\sqrt{2h+4} + O(e), \mu = 0)$ and $\varphi_{c_e}(t; r_0 = r_0^\diamond = \frac{(1\pm e)}{2}, \dot{r}_0 = 0, z_0 = \pm\sqrt{\frac{1}{h^2} - \frac{1}{4}} + O(e), \dot{z}_0 = 0, \mu = 0)$. But not all eight periodic solutions give different periodic orbits of the reduced elliptic Sitnikov problem (see Theorem 8.4). Since the reduced isosceles problem is an autonomous system, if we continue different periodic solutions that define the same periodic orbit, then we will obtain the same periodic orbit of the reduced isosceles problem. Therefore, Corollary 11.3 and Theorems 11.5 and 11.7 prove Theorem 11.1.

We note that in order to continue the symmetric periodic solutions of the reduced elliptic Sitnikov problem, we applied the implicit function theorem, choosing μ and τ as the independent variables. As happened in the continuation of periodic solutions from the reduced circular Sitnikov problem (see section 10), there are other possible choices for the independent variables. These other possible choices are (μ, r_0), $(\mu, \dot{z}_0)$, and $(\mu, \widetilde{h})$ (respectively, (μ, r_0), (μ, z_0), $(\mu, \widetilde{h})$) when the starting initial condition that we continue is r-symmetric (respectively, t-symmetric). Here $\widetilde{h}$ is the energy of the periodic solution. We have analyzed these choices for the independent variables, but we have not obtained new periodic orbits. In particular, we have seen that the determinant that we must evaluate when we use $(\mu, \dot{z}_0)$ (respectively, (μ, z_0)) as the independent variables is more complicated than in the other cases because we do not know an explicit expression of some of the derivatives.

In particular, we also have analyzed the continuation of the symmetric periodic solutions of the reduced elliptic Sitnikov problem given by Theorems 8.2 and 8.3 to symmetric periodic solutions of the reduced isosceles problem by fixing either the period, one of the initial conditions, or the energy. We have seen that if the eccentricity e is sufficiently small, then these symmetric periodic solutions can be continued to families of symmetric periodic solutions of the reduced isosceles problem for $\mu > 0$ sufficiently small that have either the same period, the same initial condition r_0, or the same energy $\widetilde{h}$ as the initial orbit. We have also evaluated numerically for some periodic orbits the correspondent determinant when we continue by fixing the initial condition $\dot{z}_0$ (respectively, z_0), and we have seen that it is different from zero.

We note that in order to apply successfully the implicit function theorem it is very important to choose a good set of independent variables.

12. From reduced circular Sitnikov problem to reduced isosceles problem in two steps. In section 10 we have continued directly the periodic orbits of the reduced circular Sitnikov problem with period $T \neq 4\pi n$ for all $n \in \mathbb{N}$ to doubly symmetric periodic orbits of the reduced isosceles problem for $\mu > 0$ sufficiently small having period near T and fixed angular momentum $c = 1/4$. Now we continue, by using two steps, the periodic orbits of the reduced circular Sitnikov problem with rational period $T = 2\pi p/q$ for all $p, q \in \mathbb{N}$ coprime and $p > q/(2\sqrt{2})$ to symmetric periodic orbits of the reduced isosceles problem for $\mu > 0$ sufficiently small having period near $2\pi p$ and fixed angular momentum $c = 1/4$. First, we continue them to periodic orbits of the reduced elliptic Sitnikov problem for sufficiently small values of e, and then we continue the periodic orbits of the reduced elliptic Sitnikov problem to the reduced isosceles problem for $\mu > 0$ sufficiently small, always having fixed angular momentum $c = 1/4$. The main differences between direct continuation and continuation in two steps are analyzed at the end of this section.

LEMMA 12.1. *Let $\varphi(t) = (r(t), \dot{r}(t), z(t), \dot{z}(t))$ be a periodic solution of the reduced isosceles problem* (3.1) *with $c = c_e$ having initial conditions $r(0) = r_0$, $\dot{r}(0) = 0$, $z(0) = z_0$, $\dot{z}(0) = \dot{z}_0$ and period τ. If we set $\alpha = 1/(1-e^2)$, $\widetilde{r}(t) = \alpha r(\alpha^{3/2}t)$, $\dot{\widetilde{r}}(t) = \alpha^{-1/2}\dot{r}(\alpha^{3/2}t)$, $\widetilde{z}(t) = \alpha z(\alpha^{3/2}t)$, and $\dot{\widetilde{z}}(t) = \alpha^{-1/2}\dot{z}(\alpha^{3/2}t)$, then $\gamma(t) = (\widetilde{r}(t), \dot{\widetilde{r}}(t), \widetilde{z}(t), \dot{\widetilde{z}}(t))$ is a periodic solution of the reduced isosceles problem* (3.1) *with $c = 1/4$ having initial conditions $\widetilde{r}(0) = \alpha r_0$, $\dot{\widetilde{r}}(0) = 0$, $\widetilde{z}(0) = \alpha z_0$, $\dot{\widetilde{z}}(0) = \alpha^{-1/2}\dot{z}_0$ and period $\widetilde{\tau} = \alpha^{-3/2}\tau$.*

Proof. The proof is an immediate consequence of Proposition 3.1. □

Remark 12.2. We note that the period $\widetilde{\tau} = \widetilde{\tau}(e) = \tau(1-e^2)^{3/2}$ is a decreasing function in $(0,1)$, so in this interval the function $\widetilde{\tau}(e)$ has the inverse

$$e(\widetilde{\tau}) = \sqrt{1 - \left(\frac{\widetilde{\tau}}{\tau}\right)^{2/3}}.$$

Therefore, the solution $\gamma(t) = (\widetilde{r}(t), \dot{\widetilde{r}}(t), \widetilde{z}(t), \dot{\widetilde{z}}(t))$ can be parameterized by the period $\widetilde{\tau}$ instead of the eccentricity e.

Let γ_{pq} be the periodic orbit of the reduced circular Sitnikov problem with period $T = 2\pi p/q$ for given $p, q \in \mathbb{N}$ coprime with $p > q/(2\sqrt{2})$. Choosing conveniently the origin of time, γ_{pq} can be thought of as the orbit associated to either the solutions $\varphi_{1/4}\left(t; r_0 = 1/2, \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{z}_0^* = \pm\sqrt{2h+4}, \mu = 0\right)$ or the solutions $\varphi_{1/4}(t; r_0 = 1/2, \dot{r}_0 = 0, z_0 = z_0^* = \pm\sqrt{\frac{1}{h^2} - \frac{1}{4}}, \dot{z}_0 = 0, \mu = 0)$, where h is such that $T = T(h) = 2\pi p/q$.

We start analyzing the continuation in two steps of the periodic solutions $\varphi_{1/4}(t; 1/2, 0, 0, \dot{z}_0^* = \pm\sqrt{2h+4}, 0\big)$ to r-symmetric periodic solutions of the reduced isosceles problem with $c = 1/4$ and $\mu > 0$ sufficiently small. Afterward we will analyze the continuation in two steps to t-symmetric periodic solutions of the periodic solutions $\varphi_{1/4}(t; 1/2, 0, z_0^* = \pm\sqrt{\frac{1}{h^2} - \frac{1}{4}}, 0, 0)$. We note that it is not necessary to consider the continuation in two steps of the above periodic solutions to doubly symmetric periodic solutions, because it can be obtained from the continuation of either r- or t-symmetric periodic solutions having period $T = 2\pi p/q$ with p even.

By Theorem 8.2, each periodic solution $\varphi_{1/4}(t; 1/2, 0, 0, \dot{z}_0^* = \pm\sqrt{2h+4}, 0)$ can be continued to two families $\varphi_{c_e}(t; r_0 = r_0^P = (1-e)/2, \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{z}_0^P = \dot{z}_0^* +$

$O(e), \mu = 0)$ and $\varphi_{c_e}(t; r_0 = r_0^A = (1+e)/2, \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{z}_0^A = \dot{z}_0^* + O(e), \mu = 0)$ of r-symmetric periodic solutions of the reduced elliptic Sitnikov problem having period $\tau^\diamond = 2\pi p = qT$ for $e > 0$ sufficiently small. Moreover these two families are formed by doubly symmetric periodic solutions if p is even, and they are formed by r-symmetric periodic solutions that are not doubly symmetric if p is odd. Then, using Lemma 12.1, Theorem 8.2 can be stated as follows.

THEOREM 12.3 (reformulation of Theorem 8.2). *Let $\varphi_{1/4}\left(t; r_0 = 1/2, \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{z}_0^* = \pm\sqrt{2h+4}, \mu = 0\right)$ be a periodic solution of the reduced circular Sitnikov problem with period $T = 2\pi p/q$ for given $p, q \in \mathbb{N}$ coprime and $p > q/(2\sqrt{2})$. We denote*

$$\widetilde{r}_0^P(e) = \frac{r_0^P}{1-e^2} = \frac{1}{2(1+e)}\,, \qquad \widetilde{r}_0^A(e) = \frac{r_0^A}{1-e^2} = \frac{1}{2(1-e)}\,,$$
$$\dot{\widetilde{z}}_0^P(e) = \sqrt{1-e^2}\,\dot{z}_0^P\,, \qquad \dot{\widetilde{z}}_0^A(e) = \sqrt{1-e^2}\,\dot{z}_0^A\,.$$

1. *The solution $\varphi_{1/4}\left(t; 1/2, 0, 0, \dot{z}_0^*, 0\right)$ can be continued to two families $\varphi_{1/4}(t; r_0 = \widetilde{r}_0^P(e), \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{\widetilde{z}}_0^P(e), \mu = 0)$ and $\varphi_{1/4}(t; r_0 = \widetilde{r}_0^A(e), \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{\widetilde{z}}_0^A(e), \mu = 0)$ of r-symmetric periodic solutions of the reduced elliptic restricted isosceles problem with angular momentum $c = 1/4$ having period $\widetilde{\tau} = 2\pi p(1-e^2)^{3/2}$ for $e \in (0, \overline{e})$ with $\overline{e}$ sufficiently small.*
2. *If p is odd, the r-symmetric periodic solutions $\varphi_{1/4}(t; \widetilde{r}_0^{P,\,A}(e), 0, 0, \dot{\widetilde{z}}_0^{P,\,A}(e), 0)$ are not doubly symmetric, whereas if p is even, then they are doubly symmetric.*

Let $\varphi_{c_e}(t; r_0 = r_0^\diamond, \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{z}_0^\diamond, \mu = 0)$ be one of the r-symmetric periodic solutions of the reduced elliptic Sitnikov problem given by Theorem 8.2 for fixed values of p, q and $e > 0$ small. If e is sufficiently small, then from Theorem 11.5, this r-symmetric periodic solution can be continued to a 2-parameter family (on μ and τ) $\varphi_{c_e}(t; r_0 = r_0(\mu, \tau), \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{z}_0(\mu, \tau), \mu)$ of r-symmetric periodic solutions of the reduced isosceles problem, with angular momentum $c = c_e$ and $\mu \geqslant 0$, that have period τ for (μ, τ) in a sufficiently small neighborhood W of $(0, \tau^\diamond)$. Moreover $r_0(\mu, \tau)$ and $\dot{z}_0(\mu, \tau)$ are the two unique analytic functions defined in W such that $r_0(0, \tau^\diamond) = r_0^\diamond$ and $\dot{z}_0(0, \tau^\diamond) = \dot{z}_0^\diamond$. We note that, by Lemma 12.1,

$$\varphi_{1/4}\left(t; r_0 = \overline{r}_0 = \frac{r_0(\mu, \tau)}{1-e^2}, \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{\overline{z}}_0 = \sqrt{1-e^2}\,\dot{z}_0(\mu, \tau), \mu\right)$$

is an r-symmetric periodic solution of the reduced isosceles problem, with angular momentum $c = 1/4$ and $\mu \geqslant 0$, that has period $\overline{\tau} = \tau(1-e^2)^{3/2}$. In short, Theorem 11.5 can be stated as follows.

THEOREM 12.4 (reformulation of Theorem 11.5). *Let $\varphi_{1/4}(t; r_0 = \widetilde{r}_0^{P,\,A}, \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{\widetilde{z}}_0^{P,\,A}, \mu = 0)$ be one of the r-symmetric periodic solutions of the reduced elliptic restricted isosceles problem given by Theorem 12.3 for fixed $e > 0$ sufficiently small and $p, q \in \mathbb{N}$ coprime with $p > q/(2\sqrt{2})$.*

1. *This solution can be continued to a 2-parameter family (on μ and $\overline{\tau}$) $\varphi_{1/4}(t;\ r_0 = \overline{r}_0(\mu, \overline{\tau}), \dot{r}_0 = 0, z_0 = 0, \dot{z}_0 = \dot{\overline{z}}_0(\mu, \overline{\tau}), \mu)$ of r-symmetric periodic solutions of the reduced isosceles problem, with angular momentum $c = 1/4$ and $\mu \geqslant 0$, that have period $\overline{\tau}$ for $(\mu, \overline{\tau})$ in a sufficiently small neighborhood $\overline{W}$ of $\left(0, 2\pi p(1-e^2)^{3/2}\right)$. Moreover $\overline{r}_0(\mu, \overline{\tau})$ and $\dot{\overline{z}}_0(\mu, \overline{\tau})$ are the two unique analytic functions defined in $\overline{W}$ such that $\overline{r}_0\left(0, 2\pi p(1-e^2)^{3/2}\right) = \widetilde{r}_0^{P,\,A}$ and $\dot{\overline{z}}_0\left(0, 2\pi p(1-e^2)^{3/2}\right) = \dot{\widetilde{z}}_0^{P,\,A}$.*

2. *If p is odd, the r-symmetric periodic solutions $\varphi_{c_e}(t;\overline{r}_0(\mu,\overline{\tau}),0,0,\dot{\overline{z}}_0(\mu,\overline{\tau}),\mu)$ are not doubly symmetric, whereas if p is even, they are doubly symmetric.*

We note that using Remark 12.2, the solutions obtained from Theorems 12.3 and 12.4 can be parameterized by means of the period $\widetilde{\tau}$ and $\overline{\tau}$, respectively, instead of the eccentricity.

Using the period instead of the eccentricity as a parameter, the r-symmetric periodic solutions of the reduced restricted isosceles problem $\varphi_{1/4}(t;\widetilde{r}_0^{\,P,\,A}(e),0,0,\dot{\widetilde{z}}_0^{\,P,\,A}(e),0)$ given by Theorem 12.3 become $\varphi_{1/4}(t;\widehat{r}_0^{\,P,\,A}(\widetilde{\tau}),0,0,\dot{\widehat{z}}_0^{\,P,\,A}(\widetilde{\tau}),0)$, where $\widehat{r}_0^{\,P,\,A}(\widetilde{\tau}) = \widetilde{r}_0^{\,P,\,A}(e(\widetilde{\tau}))$ and $\dot{\widehat{z}}_0^{\,P,\,A}(\widetilde{\tau}) = \dot{\widetilde{z}}_0^{\,P,\,A}(e(\widetilde{\tau}))$, with

$$e(\widetilde{\tau}) = \sqrt{1-\left(\frac{\widetilde{\tau}}{2\pi p}\right)^{2/3}}$$

and $\widetilde{\tau} \in (\widetilde{\tau}_1,\widetilde{\tau}_2) = \left(\tau^\diamond(1-\overline{e}^2)^{3/2},\tau^\diamond\right)$ for $\overline{e}$ sufficiently small. On the other hand, from Theorem 12.4, we have that, for a fixed value of $\widetilde{\tau}^* \in (\widetilde{\tau}_1,\widetilde{\tau}_2)$, we can find two unique analytic functions $\overline{r}_0^{\,P,\,A}(\mu,\overline{\tau})$ and $\dot{\overline{z}}_0^{\,P,\,A}(\mu,\overline{\tau})$ in such a way that $\varphi_{1/4}(t;r_0 = \overline{r}_0^{\,P,\,A}(\mu,\overline{\tau}),\dot{r}_0 = 0,z_0 = 0,\dot{z}_0 = \dot{\overline{z}}_0^{\,P,\,A}(\mu,\overline{\tau}),\mu)$ is an r-symmetric periodic solution of the reduced isosceles problem, with angular momentum $c = 1/4$ and $\mu \geqslant 0$, that has period $\overline{\tau}$ for $(\mu,\overline{\tau})$ in a sufficiently small neighborhood $\overline{W}$ of $(0,\widetilde{\tau}^*)$. Moreover $\overline{r}_0^{\,P,\,A}(\mu,\overline{\tau})$ and $\dot{\overline{z}}_0^{\,P,\,A}(\mu,\overline{\tau})$ are the two unique analytic functions defined in $\overline{W}$ such that $\overline{r}_0^{\,P,\,A}(0,\widetilde{\tau}^*) = \widehat{r}_0^{\,P,\,A}(\widetilde{\tau}^*)$ and $\dot{\overline{z}}_0^{\,P,\,A}(0,\widetilde{\tau}^*) = \dot{\widehat{z}}_0^{\,P,\,A}(\widetilde{\tau}^*)$. In particular, $\overline{r}_0^{\,P,\,A}(0,\widetilde{\tau}) = \widehat{r}_0^{\,P,\,A}(\widetilde{\tau})$ and $\dot{\overline{z}}_0^{\,P,\,A}(0,\widetilde{\tau}) = \dot{\widehat{z}}_0^{\,P,\,A}(\widetilde{\tau})$ for all $(0,\widetilde{\tau}) \in \overline{W}$. Then using the compactness argument of Theorem 10.2 and working again with the parameter e instead of $\widetilde{\tau}$, Theorem 12.4 can be improved as follows.

THEOREM 12.5. *For fixed $p,q \in \mathbb{N}$ coprime with $p > q/(2\sqrt{2})$, for any interval $[e_1,e_2]$ with $0 < e_1 < e_2 < \overline{e}$ and $\overline{e}$ sufficiently small, we can find $\mu_0 > 0$ and analytic functions $r_0^P(\mu,e)$, $\dot{z}_0^P(\mu,e)$, $r_0^A(\mu,e)$, $\dot{z}_0^A(\mu,e)$ defined for all $\mu \in [0,\mu_0)$ and $e \in [e_1,e_2]$ such that $\varphi_{1/4}(t;r_0^P(\mu,e),0,0,\dot{z}_0^P(\mu,e),\mu)$ and $\varphi_{1/4}(t;r_0^A(\mu,e),0,0,\dot{z}_0^A(\mu,e),\mu)$ are r-symmetric periodic solutions of the reduced isosceles problem* (3.1)*, with angular momentum $c = 1/4$, that have period $\overline{\tau} = 2\pi p(1-e^2)^{3/2}$. Moreover*

$$r_0^P(0,e) = \frac{1}{2(1+e)}\ ,\quad \dot{z}_0^P(0,e) = \dot{\widetilde{z}}_0^P(e)\ ,\quad r_0^A(0,e) = \frac{1}{2(1-e)}\ ,\quad \dot{z}_0^A(0,e) = \dot{\widetilde{z}}_0^A(e)\ ,$$

where the functions $\dot{\widetilde{z}}_0^P(e)$ and $\dot{\widetilde{z}}_0^A(e)$ are the ones given by Theorem 12.3*.*

Moreover if p is even, then the continued periodic solutions are doubly symmetric, whereas if p is odd, then they are r- but not doubly symmetric.

In short, from Theorems 12.3 and 12.5, we have the following result.

THEOREM 12.6. *The two periodic solutions of the reduced circular Sitnikov problem $\varphi_{1/4}\left(t;1/2,0,0,\pm\sqrt{2h+4},0\right)$ having period $T = 2\pi p/q$ for given $p,q \in \mathbb{N}$ coprime with $p > q/(2\sqrt{2})$ can be continued by two steps to two 2-parameter families (on μ and e) $\varphi_{1/4}(t;r_0^P(\mu,e),0,0,\dot{z}_0^P(\mu,e),\mu)$ and $\varphi_{1/4}(t;r_0^A(\mu,e),0,0,\dot{z}_0^A(\mu,e),\mu)$ of r-symmetric periodic solutions of the reduced isosceles problem* (3.1)*, with angular momentum $c = 1/4$ and $\mu \geqslant 0$ sufficiently small, that have period $\overline{\tau} = 2\pi p(1-e^2)^{3/2}$ for $e > 0$ sufficiently small. Furthermore if p is even, then the continued periodic solutions are doubly symmetric, whereas if p is odd, then they are r- but not doubly symmetric.*

Applying to the t-symmetric periodic solutions $\varphi_{1/4}(t;1/2,0,z_0^* = \pm\sqrt{\frac{1}{h^2}-\frac{1}{4}},0,0)$

the arguments that we have used to continue the r-symmetric periodic solutions in two steps, we obtain the following result.

THEOREM 12.7. *The two periodic solutions of the reduced circular Sitnikov problem* $\varphi_{1/4}(t;1/2,0,\pm\sqrt{\frac{1}{h^2}-\frac{1}{4}},0,0)$ *having period* $T=2\pi p/q$ *for given* $p,q\in\mathbb{N}$ *coprime with* $p>q/(2\sqrt{2})$ *can be continued by two steps to two 2-parameter families (on* μ *and* e*)* $\varphi_{1/4}\,(t;r_0^P(\mu,e),0,z_0^P(\mu,e),0,\mu)$ *and* $\varphi_{1/4}\,(t;r_0^A(\mu,e),0,\ z_0^A(\mu,e),0,\mu)$ *of* t*-symmetric periodic solutions of the reduced isosceles problem* (3.1)*, with angular momentum* $c=1/4$ *and* $\mu\geqslant 0$ *sufficiently small, that have period* $\overline{\tau}=2\pi p(1-e^2)^{3/2}$ *for* $e>0$ *sufficiently small. Furthermore if* p *is even, then the continued periodic solutions are doubly symmetric, whereas if* p *is odd, then they are* t*-symmetric but not doubly symmetric.*

By Theorems 12.6 and 12.7 the periodic orbit of the reduced circular Sitnikov problem with period $T=2\pi p/q$ for given $p,q\in\mathbb{N}$ coprime and $p>q/(2\sqrt{2})$ can be continued in two steps to eight 2-parameter families (on μ and e) of symmetric periodic orbits of the reduced isosceles problem (3.1) with angular momentum $c=1/4$ and $\mu\geqslant 0$ small. But not all eight families of symmetric periodic orbits are different.

THEOREM 12.8. *Let* γ_{pq} *be the periodic orbit of the reduced circular Sitnikov problem with period* $T=2\pi p/q$ *for given* $p,q\in\mathbb{N}$ *coprime with* $p>q/(2\sqrt{2})$*.*

1. *If* p *is odd, then* γ_{pq} *can be continued by two steps to four 2-parameter families (on* μ *and* e*) of symmetric periodic orbits of the reduced isosceles problem* (3.1)*, with angular momentum* $c=1/4$ *and* $\mu\geqslant 0$ *sufficiently small, that have period* $\tau=2\pi p(1-e^2)^{3/2}$ *with* $e>0$ *sufficiently small. Moreover, two of these families are formed by* r*-symmetric periodic orbits that are not doubly symmetric, and the other two are formed by* t*-symmetric periodic orbits that are not doubly symmetric.*
2. *If* p *is even, then* γ_{pq} *can be continued by two steps to two 2-parameter families (on* μ *and* e*) of doubly symmetric periodic orbits of the reduced isosceles problem* (3.1)*, with angular momentum* $c=1/4$ *and* $\mu\geqslant 0$ *sufficiently small, that have period* $\tau=2\pi p(1-e^2)^{3/2}$ *with* $e>0$ *sufficiently small.*

Proof. From Lemma 12.1, we can see easily that different periodic orbits of the reduced isosceles problem with $c=c_e$ correspond to different periodic orbits of the reduced isosceles problem with $c=1/4$. Thus the proof follows immediately from Theorems 8.4, 11.1, 12.6, and 12.7. □

We remark that the periodic orbits γ_{p1} of the reduced circular Sitnikov problem with period $T=2\pi p$ for some even $p\in\mathbb{N}$ cannot be continued by direct continuation. They can only be continued by using two steps. The periodic orbits γ_{pq} with $q\neq 1$ and the ones with p odd and $q=1$ can be continued in both ways, that is, using direct continuation and using continuation in two steps. We note that if we use direct continuation, then γ_{pq} can be continued to a family of doubly symmetric periodic orbits with period near $T=2\pi p/q$. On the other hand, using continuation in two steps, γ_{pq} can be continued to two or four families of symmetric periodic orbits with period near $\tau^\diamond=2\pi p=qT$ (two families of doubly symmetric periodic orbits when p is even, and two families of r-symmetric plus two families of t-symmetric periodic orbits that are not doubly symmetric when p is odd). Therefore if $q\neq 1$, then the periodic orbits obtained from direct continuation and those obtained from continuation in two steps are always different, because they have different periods. Moreover, when p is odd, the orbits obtained from direct continuation are doubly symmetric, whereas the ones obtained from continuation in two steps are r- and t-symmetric, but not doubly symmetric. Therefore when p is odd and $q=1$ the direct continuation and the

continuation in two steps also give different periodic orbits. Finally the periodic orbits of the reduced circular Sitnikov problem with period $T = 2\pi\omega$, where $\omega > 1/(2\sqrt{2})$ is an irrational number, can be continued by direct continuation, but they cannot be continued in two steps.

13. Summary. The main results about continuation of the periodic orbits of the reduced circular Sitnikov problem to symmetric periodic orbits of the reduced isosceles problem for $\mu > 0$ sufficiently small—that is, Theorem 10.1 and Theorem 12.8—are summarized in Theorem A of the introduction.

In Remark 12.2 we have seen that we can work with the parameter $\tau = 2\pi p\, f(e)$ (the period) instead of the eccentricity e. Thus the 2-parameter families of periodic orbits of the reduced isosceles problem obtained from continuation in two steps of periodic orbits of the reduced circular Sitnikov problem with period $T = 2\pi p/q$ for $p, q \in \mathbb{N}$ coprime with $p > q/(2\sqrt{2})$ can be parameterized by means of μ and τ instead of μ and e. This means that Theorem A of the introduction can be stated using μ and τ as parameters instead of μ and e.

Next we give the extension of Theorem A to the full isosceles problem (see section 4 for more details about the relationship between the periodic orbits of the reduced isosceles problem and the orbits of the full isosceles problem).

Let Π_T denote the two-dimensional invariant torus of the restricted isosceles problem that comes from a periodic orbit of the reduced circular Sitnikov problem with period T. Then we have the following result.

THEOREM 13.1. *The torus of the circular restricted isosceles problem Π_T with $T > \pi/\sqrt{2}$ can be continued to the following families of two-dimensional tori of the isosceles problem with $\mu > 0$ sufficiently small. These tori are filled with either periodic or quasi-periodic orbits:*

1. *Case $T = 2\pi\omega$ with $\omega > 1/(2\sqrt{2})$ an irrational number.*
   (a) *Π_T can be continued directly to one 2-parameter family (on μ and τ with τ sufficiently close to T) of two-dimensional tori.*
2. *Case $T = 2\pi p/q$ for some $p, q \in \mathbb{N}$ coprime with $p > q/(2\sqrt{2})$.*
   (a) *p odd:*
      i. *Π_T can be continued directly to one 2-parameter family (on μ and τ with τ sufficiently close to T) of two-dimensional tori.*
      ii. *Π_T can be continued by two steps to four 2-parameter families (on μ and τ with τ sufficiently close to Tq) of two-dimensional tori.*
   (b) *p even and $q \neq 1$:*
      i. *Π_T can be continued directly to one 2-parameter family (on μ and τ with τ sufficiently close to T) of two-dimensional tori.*
      ii. *Π_T can be continued by two steps to two 2-parameter families (on μ and τ with τ sufficiently close to Tq) of two-dimensional tori.*
   (c) *p even and $q = 1$:*
      i. *Π_T can be continued by two steps to two 2-parameter families (on μ and τ with τ sufficiently close to Tq) of two-dimensional tori.*

By Proposition 7.7, the tori Π_T are filled with periodic orbits when $T = p2\pi/q$ for some $p, q \in \mathbb{N}$ coprime with $p > q/(2\sqrt{2})$; and they are filled with quasi-periodic orbits when $T = 2\pi\omega$ with $\omega > 1/(2\sqrt{2})$ an irrational number. So, in particular, we have continued tori filled with quasi-periodic orbits. The tori of the isosceles problem for $\mu > 0$ that we have obtained are filled with either periodic or quasi-periodic orbits of the isosceles problem.

Remember that the phase portrait of the isosceles problem on each angular mo-

mentum level c with $c \neq 0$ is the same (see Proposition 3.1). Therefore we have obtained invariant periodic and quasi-periodic two-dimensional tori on each angular momentum level $c \neq 0$.

REFERENCES

[1] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1978.

[2] E. Belbruno, J. Llibre, and M. Ollé, *On the families of periodic orbits which bifurcate from the circular Sitnikov motions*, Celestial Mech. Dynam. Astronom., 60 (1994), pp. 99–129.

[3] D. Brouwer and G. M. Clemence, *Methods of Celestial Mechanics*, Academic Press, New York, 1961.

[4] P. F. Byrd and M. D. Friedman, *Handbook of Elliptic Integrals for Engineers and Physicists*, Springer-Verlag, Berlin, 1954.

[5] C. Chicone, *Bifurcations of nonlinear oscillations and frequency entrainment near resonance*, SIAM J. Math. Anal., 23 (1992), pp. 1577–1608.

[6] M. Corbera, *Periodic and Quasi-periodic Motions for the Spatial Isosceles 3-Body Problem*, Ph.D. thesis, Universitat Autònoma de Barcelona, 1999.

[7] M. Corbera and J.Llibre, *Periodic orbits of the Sitnikov problem via a Poincaré map*, Celestial Mech. Dynam. Astronom., 77 (2000), pp. 273–303.

[8] M. Corbera and J. Llibre, *On symmetric periodic orbits of the elliptic Sitnikov problem via the analytic continuation method*, in Celestial Mechanics (Evanston, IL, 1999), Contemp. Math. 292, AMS, Providence, RI, 2002, pp. 91–127

[9] R. L. Devaney, *Motion near total collapse in the planar isosceles three-body problem*, Celestial Mech., 28 (1982), pp. 25–36.

[10] S. I. Diliberto, *On systems of ordinary differential equations*, in Contributions to the Theory of Nonlinear Oscillations, Ann. Math. Stud. 20, Princeton University Press, Princeton, NJ, 1950, pp. 1–38.

[11] J. K. Hale, *Ordinary Differential Equations*, Robert E. Krieger, Huntington, NY, 1980.

[12] R. C. Howison and K. R. Meyer, *Doubly-symmetric periodic solutions of the spatial restricted three-body problem*, J. Differential Equations, 163 (2000), pp. 174–197.

[13] S. Lefschetz, *Differential Equations: Geometric Theory*, Dover, New York, 1977.

[14] W. D. MacMillan, *An integrable case in the restricted problem of three bodies*, Astron. J., 27 (1913), p. 11.

[15] K. R. Meyer, *Periodic Solutions of the N-Body Problem*, Lecture Notes in Math. 1719, Springer-Verlag, Berlin, 1999.

[16] K. R. Meyer and Q. Wang, *The global phase structure of the three-dimensional isosceles three-body problem with zero energy*, in Hamiltonian Dynamical Systems (Cincinnati, OH, 1992), IMA Vol. Math. Appl. 63, Springer-Verlag, New York, 1995, pp. 265–282

[17] R. Moeckel, *Heteroclinic phenomena in the isosceles three-body problem*, SIAM J. Math. Anal., 15 (1984), pp. 857–876.

[18] P. Painlevé, *Leçons sur le théory analytique des équations différentielles*, Hermann, Paris, 1897.

[19] G. Pavanini, *Sopra una nuova categoria di soluzioni periodiche nel problema di tre corpi*, Annali di Mathematica, Serie III, Tomo XIII, 1907.

[20] H. Poincaré, *Les Mèthodes Nouvelles de la Mécanique Céleste*, 3 Vols., Gauthier-Villars, Paris 1892–1899; reprinted by Dover, New York, 1957.

[21] C. L. Siegel and J. K. Moser, *Lectures on Celestial Mechanics*, Springer-Verlag, Berlin, 1971.

[22] K. Stumpff, *Himmelsmeckanik*, Band II, Veb, Berlin, 1965, pp. 73–79.

[23] V. Szebehely, *Theory of Orbits*, Academic Press, New York, 1967.

[24] K. Wodnar, *Analytical approximations for Sitnikov's problem*, in From Newton to Chaos, A. E. Roy and B. A. Steves, eds., Plenum Press, New York, 1995, pp. 513–523.

[25] Z. Xia, *The existence of non collision singularities in Newtonian systems*, Ann. of Math., 135 (1992), pp. 411–468.

SIAM J. MATH. ANAL.
Vol. 35, No. 5, pp. 1347–1370

A POSTERIORI ERROR ESTIMATE FOR FRONT-TRACKING: SYSTEMS OF CONSERVATION LAWS*

M. LAFOREST†

Abstract. We demonstrate an a posteriori error estimate in the L^1 norm for front-tracking approximate solutions to hyperbolic systems of nonlinear conservation laws. Starting with the L^1-stability result of Bressan, Liu, and Yang, we use their L^1-equivalent functional for pairs of front-tracking approximations and identify the leading order contribution to the numerical error. Our measure for the error explicitly identifies the local sources of errors within front-tracking approximations. We also show that these local error estimators are necessary and sufficient global error estimators. We apply the estimate to a new and wider class of front-tracking approximations, which includes the approximations of Risebro.

Key words. hyperbolic, nonlinear, conservation laws, error, a posteriori, front-tracking

AMS subject classifications. 65M15, 35L65

DOI. 10.1137/S0036141002416870

1. Introduction.

We present an a posteriori error estimate for front-tracking approximate solutions to the Cauchy problem for nonlinear systems of conservation laws in one space dimension,

$$u_t + f(u)_x = 0. \tag{1.1}$$

Our estimate of the difference between an exact and an approximate solution to (1.1) is called a posteriori because it depends only on the approximate solution, the initial data, and the flux f. This result extends the L^1-stability estimate of Bressan, Liu, and Yang [6] by explicitly identifying the leading order contribution to the error in front-tracking approximations. The global error is bounded by local error estimators which measure the entropy production along discontinuities in the piecewise constant front-tracking approximations [17]. The front-tracking approximations developed in this paper generalize a construction of Risebro [22] by allowing some additional flexibility in the choice of the strengths, speeds, and interactions of its discontinuities.

A posteriori error estimates serve as upper bounds of the error within approximations and as criteria for local mesh refinement strategies. They are well established for elliptic and parabolic equations [1], while for hyperbolic equations their study has been hindered by the lack of stability estimates for most numerical methods. Previous a posteriori estimates were either for scalar conservation laws, linear systems of conservation laws, or systems with some stabilizing mechanism. For nonlinear scalar conservation laws, we mention the estimates of Nessyahu and Tadmor [20], Cockburn and Gau [7], and Gosse and Makridakis [11]. As we showed in [17], our estimates restricted to scalar conservation laws coincide with those of Cockburn and Gau but with a proof fashioned upon Keyfitz's L^1-stability result [14]. One popular approach

*Received by the editors October 26, 2002; accepted for publication (in revised form) May 30, 2003; published electronically January 30, 2004. This work was supported by the Fonds pour la Formation de Chercheurs et l'Aide à la Recherche (FCAR) Canada and partially supported by United States Department of Energy grant DEFG0298ER25363.

http://www.siam.org/journals/sima/35-5/41687.html

†Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11794-3600, and Department of Mathematics, Colorado State University, Fort Collins, CO 80523-1874 (laforest@math.colostate.edu).

advocated by, among others, Johnson [13] and Süli [24] involves solving an adjoint problem to characterize the process of error generation and error propagation. The existence of a solution to the adjoint problem has yet to be established except for scalar equations [25], and in practice the problem requires stabilizing the original problem, usually with artificial viscosity. We refer to the monograph edited by Barth and Deconinck [3] for a recent survey.

Our estimate therefore distinguishes itself as the first rigorous a posteriori error estimate for nonlinear systems of conservation laws. In [17], our a posteriori error estimate is used to construct an adaptive version of front-tracking. This error estimate is also of interest for numerical methods that involve discontinuous approximations, like the discontinuous Galerkin finite element method. For those methods, our local error estimators provide a rigorous basis to estimate the error generated by discontinuities. In terms of the stability theory for conservation laws, this paper provides a more explicit form of the error estimate and a proof for a new class of front-tracking approximations. We remark that most of the local estimates needed for this result have already been described in Bressan's proof of L^1-stability [5]. The most important changes are related to our introduction and analysis of waves unique to Risebro's front-tracking approximations but not present in the original ϵ-approximations used in [5, 6].

The main tool in this work is the L^1-equivalent functional $\Phi(\cdot,\cdot)$ of Liu and Yang [19] defined for pairs of front-tracking approximations of small total variation. Extending their already a posteriori estimate, we show that the leading order contribution to the error is a time integral of local error estimators $D(x,t)$ called discrepancies. In [17], discrepancies are further related to entropy production. If $\mathcal{D}\big(u(t)\big)$ denotes the set of positions of discontinuities in $u(\cdot,t)$, then we show that there exists a constant C such that

$$\Phi\big(u(\cdot,t),v(\cdot,t)\big) \le \Phi\big(u(\cdot,0),v(\cdot,0)\big) + C\int_0^t \Big(\sum_{\substack{z\in\mathcal{D}(u(s))\bigcup\\ \mathcal{D}(v(s))}} D(z,s)\Big)\,ds. \tag{1.2}$$

This can be immediately translated into a bound in the L^1 norm. We provide only those details of the proof that differ from either [5] or [6].

Our presentation and our notation follow the paper by Bressan, Liu, and Yang [6] and the monograph of Bressan [5]. In section 2.1 we cover preliminaries and give a brief description of front-tracking approximations. The discrepancy and the Liu–Yang functional are described in sections 2.2 and 2.3. Assuming a local estimate, we state and prove our main results in section 3. The proof of the local estimate is given in section 4 albeit only for the important case of a discontinuity modeling a rarefaction wave. A complete proof of this a posteriori error estimate may be found in [16]. We conclude with some remarks in section 5.

2. Preliminaries.

2.1. Front-tracking.

Consider a system of n conservation laws

$$u_t + f(u)_x = 0, \tag{2.1}$$

where the Jacobian of the smooth function f has n distinct real eigenvalues $\lambda_1(u) < \cdots < \lambda_n(u)$ for u inside some neighborhood of the origin $\Omega \subset \mathbb{R}^n$. Given initial data $\bar{u} : \mathbb{R} \to \Omega$ we shall say that the bounded measurable function $u : \mathbb{R}\times\mathbb{R}^+ \to \Omega$ is a

weak solution of the system (2.1) if $u(x,0)=\bar{u}(x)$ and if

$$\int_0^\infty \int_{-\infty}^\infty \big(\phi_t u + \phi_x f(u)\big)\, dxdt + \int_{-\infty}^\infty \phi(x,0)u(x,0)\, dx = 0 \tag{2.2}$$

for every smooth ϕ with compact support in $t \geq 0$.

Let $r_1, \ldots, r_n$ be the eigenvectors of $Df(u)$, given as smooth functions of $u \in \Omega$. Define the kth rarefaction curve through the point u^- to be the unique solution $R_k(\cdot)(u^-)$ of

$$\frac{d}{d\sigma} R_k(\sigma)(u^-) = r_k\big(R_k(\sigma)(u^-)\big) \qquad \text{and} \qquad R_k(0)(u^-) = u^-.$$

Given two states u^+ and u^-, the matrix

$$A(u^+,u^-) = \int_0^1 Df\Big((1-\theta)u^- + \theta u^+\Big)\, d\theta \tag{2.3}$$

satisfies the relation

$$A(u^+,u^-)(u^+-u^-) = f(u^+) - f(u^-). \tag{2.4}$$

Define the kth shock curve through u^- to be the set of states $u^+ \equiv S_k(\sigma)(u^-)$ satisfying the Rankine–Hugoniot condition

$$f(u^+) - f(u^-) = s(u^+ - u^-),$$

and such that $s \in \mathbb{R}$ is the kth eigenvalue of $A(u^+,u^-)$. Let $s_k(u^+,u^-)$ denote the kth eigenvalue of $A(u^+,u^-)$.

We say that a family $k \in \{1,\ldots,n\}$ is *genuinely nonlinear* if $r_k \cdot \nabla\lambda_k \neq 0$, and *linearly degenerate* if $r_k \cdot \nabla\lambda_k = 0$. If the kth family is genuinely nonlinear, then the shock and rarefaction curves can be parameterized to satisfy

$$\begin{array}{rclcrcl} \frac{d}{d\sigma}\lambda_k\big(S_k(\sigma)(u)\big) & = & 1, & & \frac{d}{d\sigma}\lambda_k\big(R_k(\sigma)(u)\big) & = & 1, \\ \lambda_k\big(S_k(\sigma)(u)\big) - \lambda_k(u) & = & \sigma, & & \lambda_k\big(R_k(\sigma)(u)\big) - \lambda_k(u) & = & \sigma. \end{array} \tag{2.5}$$

We parameterize a linearly degenerate family by arc-length. For a genuinely nonlinear family k define

$$T_k(\sigma)(u^-) = \begin{cases} R_k(\sigma)(u^-), & \sigma \geq 0, \\ S_k(\sigma)(u^-), & \sigma < 0, \end{cases} \tag{2.6}$$

and for a linearly degenerate family let $T_k(\sigma)(u^-) = S_k(\sigma)(u^-)$. The curve T_k is smooth for $\sigma \neq 0$ with two continuous derivatives at $\sigma = 0$. In particular, for small σ the difference between $R_k(\sigma)(u^-)$ and $S_k(\sigma)(u^-)$ is third order in σ.

A Riemann problem is an initial value problem for (2.1) consisting of piecewise constant initial data along the $t=0$ axis formed of two constant states u^- and u^+ separated at the origin. It is well known [18] that if u^- and u^+ belong to a sufficiently small neighborhood of the origin in $\mathbb{R}^n$ then the Riemann problem has a unique self-similar solution composed of $n+1$ constant states $\widetilde{u}_0 = u^-, \widetilde{u}_1, \ldots, \widetilde{u}_n = u^+$ satisfying $\widetilde{u}_k = T_k(p_k)(\widetilde{u}_{k-1})$ for real numbers $p_k = p_k(u^-,u^+)$. The solution is formed of n self-similar regions where it takes on the values $\widetilde{u}_{k-1}, \widetilde{u}_k$ along the boundaries of the kth region. Each region contains either

(i) a discontinuity traveling with the Rankine–Hugoniot speed and separating $\widetilde{u}_{k-1}$ from

$$\widetilde{u}_k = S_k\big(p_k(u^-, u^+)\big)(\widetilde{u}_{k-1}),$$

(ii) or a continuous solution $u\big(x/t\big)$ satisfying $\lambda_k\big(u(x/t)\big) = x/t$ and

$$\widetilde{u}_k = R_k\big(p_k(u^-, u^+)\big)(\widetilde{u}_{k-1}).$$

A similar result also holds with the curves T_k replaced by the shock curves S_k. For any two states u^- and u^+ sufficiently close to the origin, there exists a unique sequence of shocks of strength $q_1(u^-, u^+), \dots, q_n(u^-, u^+)$ satisfying

$$u^+ = S_n\big(q_n(u^-, u^+)\big) \circ \cdots \circ S_1\big(q_1(u^-, u^+)\big)(u^-). \tag{2.7}$$

The parameters $q_1(u^-, u^+), \dots, q_n(u^-, u^+)$ are called the shock coordinates of u^- with respect to u^+. For more information on hyperbolic conservation laws one may consult [5].

We assume that there exists a positive constant d such that

$$d < \inf\Big\{|l_i - l_j| \,\Big|\, i \neq j \text{ and for } k = i, j,\ l_k \in \lambda_k(\Omega), \text{ or } l_k = s_k(u, v)\ \forall u, v \in \Omega\Big\}. \tag{2.8}$$

Such a number exists if the flux f is strictly hyperbolic over a sufficiently small neighborhood Ω.

Front-tracking approximations are piecewise constant approximations containing a finite number of discontinuities. We present an extension of Risebro's version of the front-tracking algorithm [22]. Our form of Risebro's front-tracking algorithm has weaker restrictions on the speed of the waves, the strength of the rarefactions, and on the onset of linear approximations to nonlinear interactions. These new aspects of the front-tracking construction lead to errors that were not seen in the approximate solutions of Bressan, Liu, and Yang [2, 4, 6]. We begin with a few preliminary definitions which will help to clarify the construction. For convenience, given a function v of time t and space x let $v(t)$ denote the function of space $v(x, \cdot)$.

DEFINITION 2.1. *A wave in v at time t is a position and an integer $(x_\alpha, k_\alpha) \in \mathbb{R} \times \{1, \dots, n\}$ such that the states separating the discontinuity at $x = x_\alpha$, say, $v^- = v(x_\alpha-, t), v^+ = v(x_\alpha+, t)$, satisfy $p_{k_\alpha}(v^-, v^+) \neq 0$. In this case, we say that α belongs to the discontinuity at $x = x_\alpha$ and call*

(i) *k_α its characteristic type,*
(ii) *$\sigma_\alpha = p_{k_\alpha}(v^-, v^+)$ its strength, and*
(iii) *x_α its position.*

The set of all waves in $v(t)$ is denoted $\mathcal{W}\big(v(t)\big)$.

This definition distinguishes between discontinuities, bounded by any two arbitrary states, and the components of the solution of the underlying Riemann problem. In contrast to [6], where all discontinuities are called waves, here each nonzero component $p_k(v^-, v^+)$ of a discontinuity separating v^- and v^+ corresponds to a different wave.

DEFINITION 2.2. *A wave α is called isolated if it is the only wave located at $x_\alpha \in \mathbb{R}$. If a wave belonging to a genuinely nonlinear family satisfies $\sigma_\alpha < 0$, it will be called a shock wave; otherwise it will be called a rarefaction wave. A wave will be called linearly degenerate if the family k_α is linearly degenerate.*

Let $\mathcal{D}\big(v(t)\big)$ denote the set of points in space where the approximation $v(t)$ has discontinuities.

The construction of a front-tracking approximation v for all times begins with the choice of piecewise constant initial data $v(\cdot,0)$ with a finite number of discontinuities. We choose the initial data so that the difference $v(\cdot,0)-\bar{v}$ belongs to $L^1(\mathbb{R})$ and that each discontinuity represents one isolated wave α. In this case the states neighboring a discontinuity at x_α, say, v^+ and v^-, are related by

$$v^+ = T_{k_\alpha}\big(p_{k_\alpha}(v^-,v^+)\big)(v^-). \tag{2.9}$$

Each isolated wave α in the initial approximation is propagated forward along $x_\alpha(t)$ in any way that satisfies

$$\dot{x}_\alpha \in \lambda_{k_\alpha}(\Omega).$$

In a first instance, we attempt to construct v using only isolated waves satisfying (2.9). Under this constraint, we may use the following approximate Riemann solver. Consider two isolated waves meeting at time t_1 and separated by three consecutive states of v, namely, v_l, v_m, and v_r. If $\widetilde{v}_0 = v_l, \widetilde{v}_1, \dots, \widetilde{v}_n = v_r$ are the $n+1$ states of the self-similar solution to the Riemann problem for v_l and v_r, then, for $t > t_1$, we set v to be a piecewise constant function composed of n regions separated by the states $\{\widetilde{v}_i\}$. In particular,

(i) if $\widetilde{v}_k = S_k\big(p_k(v_l,v_r)\big)(\widetilde{v}_{k-1})$, then the states $\widetilde{v}_{k-1}$ and $\widetilde{v}_k$ are connected by an isolated shock wave α traveling at a speed

$$\dot{x}_\alpha \in \lambda_{k_\alpha}(\Omega); \tag{2.10}$$

(ii) if $\widetilde{v}_k = R_k\big(p_k(v_l,v_r)\big)(\widetilde{v}_{k-1})$, then $\widetilde{v}_k$ and $\widetilde{v}_{k-1}$ are connected by a finite number of isolated rarefaction waves $\alpha(1),\dots,\alpha(j)$ of strengths

$$\sum_{i=1}^{j} \sigma_{\alpha(i)} = p_k(v_l,v_r), \tag{2.11}$$

and traveling at speeds

$$\dot{x}_{\alpha(i-1)} < \dot{x}_{\alpha(i)} \qquad \text{and} \qquad \dot{x}_{\alpha(i)} \in \lambda_{k_\alpha}(\Omega) \quad \forall i. \tag{2.12}$$

There is considerable flexibility in the application of such a Riemann solver, especially with respect to the speeds $\dot{x}_\alpha$ and the strength of the rarefaction waves. Unfortunately, without further restrictions on the Riemann solver the approximation might develop infinitely many discontinuities in a finite amount of time. For this reason, Risebro proposed that for certain interactions this solver be replaced by a second one. To explain when this must be done, let t_k be the time when the kth interaction occurs and compute

$$P_k = \sum_{\alpha,\beta} |\sigma_\alpha \cdot \sigma_\beta|, \tag{2.13}$$

where the sum occurs over pairs of waves colliding at time t_k. Risebro showed that for any positive γ there exists an integer $N = N(\gamma)$ such that

$$\forall k \geq N, \qquad P_k < \gamma. \tag{2.14}$$

The proof of the existence of the time t_N follows from the total variation boundedness of the approximations.

When condition (2.14) is satisfied, the following new solver is used to limit the number of new discontinuities in the approximation. This second solver does not generally produce isolated waves. Applying the algorithm inductively, we therefore suppose that two discontinuities meet at time t_N and that the discontinuities contain, respectively, waves α and β such that $\dot{x}_\alpha \in \lambda_{k_\alpha}(\Omega)$ and $\dot{x}_\beta \in \lambda_{k_\beta}(\Omega)$. If $k_\alpha < k_\beta$ and $\widetilde{v}_0 = v_l, \widetilde{v}_1, \dots, \widetilde{v}_n = v_r$ are the $n+1$ states of the exact Riemann solver, then we pick any integer i between k_α and k_β and construct v for time $t > t_N$ by using only the three states $\widetilde{v}_0 = v_l$, $\widetilde{v}_i$, and $\widetilde{v}_n = u_r$. The result is two outgoing discontinuities separated by the state $\widetilde{v}_i$. The speeds of the outgoing discontinuities are restricted to the range of the incoming discontinuities $\lambda_{k_\alpha}(\Omega) \bigcup \lambda_{k_\beta}(\Omega)$. If $k_\alpha = k_\beta$, only one outgoing discontinuity is produced with neighboring states v_l and v_r. In effect, this algorithm begins with an approximation defined up to some maximal but finite time t_∞, then computes the time $t_N < t_\infty$ and redefines the approximation at time t_N with the second solver. Using the first solver a new approximation can then be recomputed up to a new maximal time $t'_\infty > t_\infty$, and the process is repeated.

After some time, condition (2.14) will always hold and therefore no new discontinuities will be created [22]. Since the system is strictly hyperbolic, eventually all discontinuities will cross paths and the approximation will consist of a finite number of discontinuities spreading apart. The approximation will therefore be defined for all times. This method satisfies the Courant–Friedrichs–Levy condition

$$|\dot{x}_\alpha| \le \Lambda \quad \forall \alpha \in \mathcal{W}\big(v(t)\big), \tag{2.15}$$

where

$$\Lambda = \sup_{\substack{w \in \Omega \\ k=1,\dots,n}} \big|\lambda_k(w)\big|. \tag{2.16}$$

We summarize the result of Risebro. We say that two waves $\alpha, \beta \in \mathcal{W}\big(v(t)\big)$ are *approaching* if $x_\alpha < x_\beta$ and $k_\alpha > k_\beta$, or if $k_\alpha = k_\beta$ and at least one of the waves is a shock. Define the interaction potential to be

$$Q\big(v(t)\big) = \sum_{\alpha \text{ approaches } \beta} |\sigma_\alpha \cdot \sigma_\beta|, \tag{2.17}$$

and consider the functional

$$V(v(t)) = \sum_{\alpha \in \mathcal{W}(v(t))} |\sigma_\alpha|, \tag{2.18}$$

equivalent to the total variation norm $||v(t)||_{\mathrm{TV}}$. Adapting the existence proof of Glimm [10], Risebro demonstrated the following.

LEMMA 2.3. *For any γ, there exist positive constants δ_1 and K, depending on the system* (2.1) *but independent of γ, such that if v is a front-tracking approximation satisfying*

$$||v(0)||_{L^\infty} + ||v(0)||_{\mathrm{TV}} < \delta_1, \tag{2.19}$$

then it is defined for all t and has values in Ω. Moreover, for this approximation the functional

$$V\big(v(t)\big) + KQ\big(v(t)\big) \tag{2.20}$$

is decreasing in time.

In the original construction of Risebro, a narrower class of approximations was considered so that existence results for the initial value problem (2.1) could be achieved. These so-called ϵ-approximations [2, 4, 22] depended on a parameter ϵ controlling the size of rarefaction waves, errors in wave speeds, the size of the constant γ, and the approximation of the initial data.

2.2. Residuals and discrepancies. In a global a posteriori estimate, the upper bound on the error should be measurable in terms of local and computable quantities. Usually, local error estimators for an approximation v are obtained by evaluating the residual pointwise

$$R(v) = v_t + f(v)_x. \tag{2.21}$$

For front-tracking approximations, the residual might vanish along rarefaction waves, even if (2.21) is considered in a weak sense [17], since this quantity only verifies conservation of v. For this reason, we propose a different error estimator measuring the rate of growth of the local error in L^1. In [17], these error estimators are constructed as measures of entropy production following Cockburn and Gau [7] and Dafermos [8].

For our purposes, it suffices to define the discrepancy as the pointwise change in time of the error in the L^1 norm measured along the curves T_k. Let v be a front-tracking approximation with an arbitrary discontinuity located for convenience at the origin. If the discontinuity is initially traveling at speed $\dot{z}(0)$, then define

$$V(x,t) = \begin{cases} v^- & \text{for } x < \dot{z}(0)t, \\ v^+ & \text{for } x \geq \dot{z}(0)t, \end{cases}$$

and for $\xi = x/t$ the self-similar solution of the Riemann problem $V_r(\xi)$ with initial data $V(\cdot, t)$. Instead of measuring the distance between V and V_r in phase-space, we measure the distance in oriented wave coordinates $(\widetilde{p}_1(\xi), \dots, \widetilde{p}_n(\xi))$, that is, those satisfying

$$\begin{aligned} V_r(\xi) &= T_n\big(\widetilde{p}_n(\xi)\big) \circ \cdots \circ T_1\big(\widetilde{p}_1(\xi)\big)\big(V(\xi)\big) \text{ for } \xi < \dot{z}(0), \\ V(\xi) &= T_n\big(\widetilde{p}_n(\xi)\big) \circ \cdots \circ T_1\big(\widetilde{p}_1(\xi)\big)\big(V_r(\xi)\big) \text{ for } \xi \geq \dot{z}(0). \end{aligned} \tag{2.22}$$

To treat shock waves consistently, their strengths must be measured along the shock curves S_k both before and after the point $\xi = \dot{z}(0)$. This explains why the order in which we measure the distance between V_r and V must be reversed as we cross the line $\xi = \dot{z}(0)$.

DEFINITION 2.4. *Define the discrepancy of the wave $\alpha = (x_\alpha, k_\alpha)$ to be*

$$D_\alpha(t) = \int_{-\infty}^{\infty} |\widetilde{p}_{k_\alpha}(\xi)| \, d\xi. \tag{2.23}$$

Let the discrepancy of a discontinuity at (z,t) be the sum of the discrepancies of the waves $\{\alpha | x_\alpha(t) = z\}$.

Discrepancies were first introduced by Kružkov [15] and defined in terms of the one-parameter family of Kružkov entropies. Note that the discrepancy of a discontinuity is equivalent to the quantity

$$\int_{-\infty}^{\infty} \|V(\xi) - V_r(\xi)\| \, d\xi,$$

measuring the error in phase-space. The discrepancy (2.23) could also have been defined as the rate of change of the L^1 distance between the V and V_r. By their very construction, these local error estimators are necessarily global error estimators. In section 4 we will show that they are also sufficient. The next lemma provides us with an explicit formula for the discrepancies of waves in front-tracking approximations. We begin with a definition.

DEFINITION 2.5. *Let $\alpha = (x_\alpha, k_\alpha) \in \mathcal{W}(v(t))$ be a wave belonging to a discontinuity at x_α, with neighboring states v^- and v^+. We define the left- and right-hand states of α to be, respectively,*

$$v_\alpha^- = T_{k_\alpha - 1}\big(p_{k_\alpha - 1}(v^-, v^+)\big) \circ \cdots \circ T_1\big(p_1(v^-, v^+)\big)(v^-) \qquad \text{and} \qquad v_\alpha^+ = T_{k_\alpha}(\sigma_\alpha)(v_\alpha^-). \tag{2.24}$$

We mention these states because the discrepancy in fact depends only on v_α^-, v_α^+, and the speed of the discontinuity, although Definition 2.4 emphasizes the dependence of D_α on the neighboring states v^- and v^+ of the discontinuity. For a fixed wave α, this can be seen by constructing

$$\widetilde{V}(x,t) = \begin{cases} v_\alpha^- & \text{for } x < \dot{x}_\alpha(0)t, \\ v_\alpha^+ & \text{for } x \geq \dot{x}_\alpha(0)t, \end{cases}$$

and the solution $\widetilde{V}_r$ to the Riemann problem with initial data $\widetilde{V}(0,\cdot)$. It is then straightforward to verify that the integrand $\widetilde{p}_{k_\alpha}$, originally introduced in (2.22), also satisfies

$$\begin{aligned} \widetilde{V}_r(\xi) &= T_{k_\alpha}\big(\widetilde{p}_{k_\alpha}(\xi)\big)\big(\widetilde{V}(\xi)\big) \text{ for } \xi < \dot{x}_\alpha(0), \\ \widetilde{V}(\xi) &= T_{k_\alpha}\big(\widetilde{p}_{k_\alpha}(\xi)\big)\big(\widetilde{V}_r(\xi)\big) \text{ for } \xi \geq \dot{x}_\alpha(0). \end{aligned} \tag{2.25}$$

For this reason, when computing discrepancies we will be able to assume that α is an isolated wave. We now present two lemmas providing us with explicit descriptions of discrepancies for shocks and rarefactions.

LEMMA 2.6. *If $\alpha \in \mathcal{W}(v(t))$ is a shock wave or a linearly degenerate wave, then*

$$D_\alpha(t) = |\sigma_\alpha|\big|\dot{x}_\alpha - s_{k_\alpha}(v_\alpha^-, v_\alpha^+)\big|. \tag{2.26}$$

If α is a rarefaction wave, then

$$D_\alpha(t) = \begin{cases} \frac{1}{2}\big(\lambda_{k_\alpha}(v_\alpha^-) - \dot{x}_\alpha\big)^2 + \frac{1}{2}\big(\dot{x}_\alpha - \lambda_{k_\alpha}(v_\alpha^+)\big)^2 & \text{if } \dot{x}_\alpha \in [\lambda_{k_\alpha}(v_\alpha^-), \lambda_{k_\alpha}(v_\alpha^+)], \\ |\sigma_\alpha|\big|\dot{x}_\alpha - \frac{1}{2}\big(\lambda_{k_\alpha}(v_\alpha^-) + \lambda_{k_\alpha}(v_\alpha^+)\big)\big|, & \text{otherwise.} \end{cases} \tag{2.27}$$

Proof. Suppose that $(x_\alpha(t), t) = (0,0)$ and write $v^- = v_\alpha^-$, $v^+ = v_\alpha^+$. Assume that α is a shock wave or a linearly degenerate wave and that $s_{k_\alpha}(v^-, v^+) < \dot{x}_\alpha$. The parameterization (2.5) implies that

$$\widetilde{p}_{k_\alpha}(\xi) = \begin{cases} \sigma_\alpha & \text{if } \xi \in [s_{k_\alpha}(v^-, v^+), \dot{x}_\alpha], \\ 0, & \text{otherwise.} \end{cases}$$

A short calculation of (2.23) then suffices to verify (2.26).

Suppose now that α is a rarefaction wave. We consider only the case $\dot{x}_\alpha \in [\lambda_{k_\alpha}(v^-), \lambda_{k_\alpha}(v^+)]$. Using again the parameterization (2.5), we find

$$\widetilde{p}_{k_\alpha}(\xi) = \begin{cases} \xi - \lambda_{k_\alpha}(v^-) & \text{if } \xi \in [\lambda_{k_\alpha}(v^-), \dot{x}_\alpha), \\ \lambda_{k_\alpha}(v^+) - \xi & \text{if } \xi \in [\dot{x}_\alpha, \lambda_{k_\alpha}(v^+)), \\ 0, & \text{otherwise.} \end{cases}$$

The discrepancy can then be evaluated

$$\begin{aligned} D_\alpha(t) &= \int_{\lambda_{k_\alpha}(v^-)}^{\dot{x}_\alpha} |\xi - \lambda_{k_\alpha}(v^-)| \, d\xi + \int_{\dot{x}_\alpha}^{\lambda_{k_\alpha}(v^+)} |\lambda_{k_\alpha}(v^+) - \xi| \, d\xi \\ &= \frac{1}{2}(\dot{x}_\alpha - \lambda_{k_\alpha}(v^-))^2 + \frac{1}{2}(\lambda_{k_\alpha}(v^+) - \dot{x}_\alpha)^2. \end{aligned}$$

The other cases are similar. ∎

LEMMA 2.7. *Assume that α is a rarefaction wave. For any speed $\dot{x}_\alpha$ we have*

$$|\sigma_\alpha| \Big| \dot{x}_\alpha - \frac{1}{2}\big(\lambda_{k_\alpha}(v_\alpha^-) + \lambda_{k_\alpha}(v_\alpha^+)\big) \Big| \le D_\alpha(t) \tag{2.28}$$

with equality precisely when $\dot{x}_\alpha \notin \big(\lambda_{k_\alpha}(v_\alpha^-), \lambda_{k_\alpha}(v_\alpha^+)\big)$. Moreover,

$$\inf_{\dot{x}_\alpha} D_\alpha(t) = \frac{|\sigma_\alpha|^2}{4}. \tag{2.29}$$

Proof. The proof is omitted. ∎

2.3. The Liu–Yang functional. Following the succinct presentation of [6], we construct the functional Φ of Liu and Yang. Let u and v be two front-tracking approximations and assume that $u(0) - v(0) \in L^1$. Since the total variation of both solutions remains bounded, $u(t) - v(t) \in L^1$ for all times t. The distance between both approximate solutions $u(t)$ and $v(t)$ will be measured in the shock coordinates (2.7). The shock coordinates will only be used in this context and the time t will be fixed throughout; therefore we will abbreviate

$$q_k\big(u(x,t), v(x,t)\big) = q_k(x). \tag{2.30}$$

The quantities q_k do not represent waves in either approximation, but they do play a similar role, and for clarity we name them virtual waves.

The functional of Liu and Yang

$$\Phi\big(u(t), v(t)\big) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} |q_i(x)| W_i(x) \, dx \tag{2.31}$$

is finite if the W_i are uniformly bounded. It is convenient to require

$$1 \le W_i(x) \le 2 \qquad \forall x \in \mathbb{R}, \tag{2.32}$$

uniformly in time. The weights have the form

$$W_i(x) = 1 + \kappa_1\Big(A_i(x) + B_i(x)\Big) + \kappa_2\Big(Q\big(u(t)\big) + Q\big(v(t)\big)\Big), \tag{2.33}$$

where the terms A_i and B_i are given below and the constants κ_1, κ_2 will later be chosen to satisfy (2.32). Let

$$A_i(x) = \Big(\sum_{\substack{\alpha \in \mathcal{W}(u) \bigcup \mathcal{W}(v) \\ x_\alpha < x,\, k_\alpha > i}} + \sum_{\substack{\alpha \in \mathcal{W}(u) \bigcup \mathcal{W}(v) \\ x_\alpha > x,\, k_\alpha < i}} \Big) |\sigma_\alpha| \tag{2.34}$$

measure the strength of all waves α, $k_\alpha \neq i$, which approach the virtual wave q_i. When the ith family is linearly degenerate, set $B_i = 0$. If the ith family is genuinely nonlinear, then let

$$B_i(x) = \begin{cases} \Big(\sum_{\substack{\alpha \in \mathcal{W}(u) \\ x_\alpha < x,\, k_\alpha = i}} + \sum_{\substack{\alpha \in \mathcal{W}(v) \\ x_\alpha > x,\, k_\alpha = i}} \Big) |\sigma_\alpha| & \text{if } q_i(x) < 0, \\ \Big(\sum_{\substack{\alpha \in \mathcal{W}(v) \\ x_\alpha < x,\, k_\alpha = i}} + \sum_{\substack{\alpha \in \mathcal{W}(u) \\ x_\alpha > x,\, k_\alpha = i}} \Big) |\sigma_\alpha| & \text{if } q_i(x) \geq 0 \end{cases} \tag{2.35}$$

measure the strength of the waves of the family $k_\alpha = i$ which approach q_i. Many of our later arguments will be localized around a specific wave α, and it will therefore be natural, when it is clear from the context, to suppress the symbol α. For a quantity G we write

$$G^{\alpha\pm} = \lim_{x \to x_\alpha \pm} G(x). \tag{2.36}$$

LEMMA 2.8. *If $\alpha \in \mathcal{W}\big(v(t)\big)$ is an isolated wave, then, for $i \neq k_\alpha$,*

$$W_i^{\alpha+} - W_i^{\alpha-} = \begin{cases} \kappa_1 |\sigma_\alpha| & \text{if } i < k_\alpha, \\ -\kappa_1 |\sigma_\alpha| & \text{if } i > k_\alpha. \end{cases} \tag{2.37}$$

If α is a genuinely nonlinear wave, then

$$W_{k_\alpha}^{\alpha+} - W_{k_\alpha}^{\alpha-} = \begin{cases} \kappa_1 |\sigma_\alpha| & \text{if } q_{k_\alpha}^- \text{ and } q_{k_\alpha}^+ \text{ are both positive}, \\ -\kappa_1 |\sigma_\alpha| & \text{if } q_{k_\alpha}^- \text{ and } q_{k_\alpha}^+ \text{ are both negative}. \end{cases} \tag{2.38}$$

Proof. The lemma follows directly from the definition of Φ. ∎

Assuming that κ_1 is some known but large constant, the next lemma states that condition (2.32) can be satisfied. The value of κ_1 will be determined later in the proof, where it will be shown to depend only on the flux f and the domain Ω.

LEMMA 2.9. *For any fixed value of κ_1, there exist δ_2 and κ_2, depending only on κ_1 and the system* (2.1)*, such that if u and v are front-tracking approximations satisfying*

$$\begin{gathered} ||u(0)||_{L^\infty} + ||u(0)||_{\mathrm{TV}},\ ||v(0)||_{L^\infty} + ||v(0)||_{\mathrm{TV}} < \delta_2, \\ u(0) - v(0) \in L^1(\mathbb{R}, \Omega), \end{gathered} \tag{2.39}$$

then they are defined for all t and have values in Ω. Under these conditions, the functional Φ exists, condition (2.32) *holds, and the weights W_i decrease in time. Moreover, we have the bound*

$$\kappa_1 \bigg(V\Big(u(t)\Big) + V\Big(v(t)\Big) \bigg) \leq 1. \tag{2.40}$$

Proof. For δ_1 and K from Lemma 2.3, and two approximations satisfying (2.19), both u and v take values in Ω and can be defined for all time. At the cost of possibly further restricting the size of δ_1, we may assume that

$$\Big(V\big(u(t)\big)+V\big(v(t)\big)\Big)+K\Big(Q\big(u(t)\big)+Q\big(v(t)\big)\Big)\leq\frac{1}{\kappa_1}\quad\forall t. \tag{2.41}$$

Both A_i and B_i are sums over subsets of $V(u(t))+V(v(t))$, and therefore the decay of the functional (2.20) implies the decay of $W_i(x)$. Multiply (2.41) by κ_1 and add 1 to demonstrate (2.32). ∎

At times when no interactions occur, the functional Φ is smooth in time and the weights $W_i(x)$ are constant away from the trajectories of the waves. On the other hand, at a discrete time t_0 when an interaction does occur at $x_0\in\mathbb{R}$, Lemma 2.9 implies that

$$\lim_{t\to t_0+}W_i(x)\leq\lim_{t\to t_0-}W_i(x),\qquad x\neq x_0. \tag{2.42}$$

Since the terms W_i decrease during interactions, so does the functional Φ. The main step will therefore be an analysis of Φ during those times when it is smooth.

3. Main results. In this section we provide the statement and proof of our main results while assuming certain local estimates. The proof of the required local estimates is postponed until section 4.

Fix a time t and let u and v be front-tracking approximations satisfying the conditions of Lemma 2.9. For every x, we define the intermediate states $w_0(x)\equiv u(x,t)$, and

$$w_k(x)\equiv S_k\big(q_k(x)\big)\circ\cdots\circ S_1\big(q_1(x)\big)\big(u(x,t)\big),\qquad k=1,\dots,n; \tag{3.1}$$

and to each virtual wave q_k connecting w_{k-1} to w_k, we associate a wave speed

$$s_k(x)\equiv s_k\big(w_k(x),w_{k-1}(x)\big). \tag{3.2}$$

Note that we have suppressed the dependence on time. Let $\mathcal{D}\big(u(t)\big)$ be the set of all points of discontinuity in the approximation u at time t. As explained in section 2.3, the functional Φ is discontinuous but decreasing when discontinuities cross paths. When no interactions occur, Φ is a differentiable function of time. Using the notation (2.36) we compute the derivative of Φ

$$\begin{aligned}\frac{d}{dt}\Big(\Phi\big(u(t),v(t)\big)\Big)&=\sum_{z\in\mathcal{D}(u(t))\bigcup\limits_{\mathcal{D}(v(t))}}\sum_{k=1}^{n}\Big\{\big|q_k^{z-}\big|W_k^{z-}-\big|q_k^{z+}\big|W_k^{z+}\Big\}\cdot\dot z\\&=\sum_{z\in\mathcal{D}(u(t))\bigcup\limits_{\mathcal{D}(v(t))}}\sum_{k=1}^{n}\Big\{\big|q_k^{z+}\big|W_k^{z+}(s_k^{z+}-\dot z)-\big|q_k^{z-}\big|W_k^{z-}(s_k^{z-}-\dot z)\Big\}.\end{aligned} \tag{3.3}$$

This last identity follows from the observation that $q_k(x)$ vanishes at infinity and that for two successive discontinuities at $z<y$,

$$\big|q_k^{z+}\big|W_k^{z+}s_k^{z+}=\big|q_k^{y-}\big|W_k^{y-}s_k^{y-}. \tag{3.4}$$

Introducing the terms

$$E_k(z) = \left|q_k^{z+}\right| W_k^{z+}\left(s_k^{z+} - \dot{z}\right) - \left|q_k^{z-}\right| W_k^{z-}\left(s_k^{z-} - \dot{z}\right), \tag{3.5}$$

we may rewrite the derivative (3.3) as

$$\frac{d}{dt}\Big(\Phi\big(u(t), v(t)\big)\Big) = \sum_{z \in \mathcal{D}(u(t)) \bigcup \mathcal{D}(v(t))} \sum_{k=1}^{n} E_k(z). \tag{3.6}$$

We now present upper bounds for the local quantities $E_k(z)$ which can be computed a posteriori. In the following, $\mathcal{O}(1)$ will denote a quantity whose absolute value satisfies a uniform bound depending only on the system (2.1) and Ω.

LEMMA 3.1. *Consider the strictly hyperbolic system* (2.1) *with genuinely nonlinear or linearly degenerate families in a neighborhood* Ω *of the origin. There exist positive constants* $\delta, \kappa_1, \kappa_2$, *depending only on the system* (2.1) *and* Ω, *with the following property. For any pair of front-tracking approximations* u *and* v *satisfying*

$$||u(0)||_{L^\infty} + ||u(0)||_{\mathrm{TV}},\ ||v(0)||_{L^\infty} + ||v(0)||_{\mathrm{TV}} < \delta, \tag{3.7}$$

and any position z *and time* t, *we have*

$$\sum_{k=1}^{n} E_k(z) \leq \mathcal{O}(1) \sum_{\{\alpha | x_\alpha = z\}} D_\alpha(t). \tag{3.8}$$

In contrast to [6], where the local estimate for an ϵ-approximation is of order $\mathcal{O}(1)\epsilon|\sigma_\alpha|$, we have identified the approximation error and related it to the discrepancies. The local estimates in Bressan [5], which are an improvement over those in the original paper [6], can be used to demonstrate (3.8) for isolated waves. A new argument is needed for nonisolated waves. The proof of Lemma 3.1 for isolated and nonisolated waves is the topic of sections 4.1 and 4.2. Our global a posteriori error estimate is an immediate consequence of this local estimate.

THEOREM 3.2. *Under the hypothesis of Lemma* 3.1, *there exists a constant* C, *depending only on the system* (2.1) *and on* Ω, *such that at any time* t

$$\Phi\big(u(t), v(t)\big) \leq \Phi\big(u(0), v(0)\big) + C \int_0^t \Big[\sum_{\substack{\alpha \in \mathcal{W}(u(s)) \bigcup \\ \mathcal{W}(v(s))}} D_\alpha(s) \Big]\, ds. \tag{3.9}$$

Proof. The coefficient $\mathcal{O}(1)$ in (3.8) is bounded over all waves by some constant C depending only on (2.1). Replace $\mathcal{O}(1)$ by C in inequality (3.8). Use (3.8) in (3.6) and integrate over time to obtain the final result. ∎

Theorem 3.2 can be translated into a bound in the L^1 norm for the difference between an approximate solution and a weak solution.

THEOREM 3.3. *Consider the strictly hyperbolic system* (2.1) *with genuinely nonlinear or linearly degenerate families in a neighborhood* Ω *of the origin. There exist positive constants* δ *and* C, *depending only on* (2.1) *and* Ω, *with the following property. If* u *is any front-tracking approximation with* $u(0) \in L^1(\mathbb{R}, \Omega)$ *and* v *is any weak solution in* $L^\infty\big([0,t], L^1(\mathbb{R}, \Omega)\big)$ *which is the limit in* L^1 *of a sequence of* ϵ*-approximations and moreover, if both* u *and* v *satisfy*

$$||u(0)||_{\mathrm{TV}},\ ||v(0)||_{\mathrm{TV}} < \delta, \tag{3.10}$$

then we have the estimate

$$||u(t)-v(t)||_{L^1} \le C||u(0)-v(0)||_{L^1} + C\int_0^t \Big[\sum_{z\in\mathcal{D}(u(s))} D_z(s)\Big]\,ds. \tag{3.11}$$

Proof. Use Theorem 3.2 with u and v_ϵ an ϵ-approximation. Then replace Φ by the equivalent L^1 norm. Taking the limit as $\epsilon \to 0$ demonstrates the estimate. ∎

4. Local estimates. The goal of this section is to indicate the portions of the proof of Lemma 3.1 which differ significantly from the proof of the analogous lemmas in [6] and [5]. The ϵ-approximations used in the original proof of L^1-stability involved nonisolated waves whose strength was a priori small. Although this was sufficient for the L^1 bound, our local estimates for those waves require a more delicate analysis. Nonisolated waves are treated in section 4.2. In section 4.1, we demonstrate Lemma 3.1 for isolated waves and essentially repeat the arguments given in [5] with a few minor changes. For the sake of brevity, we focus on demonstrating the estimate for genuinely nonlinear rarefaction waves. A complete proof can be found in [16].

4.1. Isolated waves. In this section we prove Lemma 3.1 under the assumption that α is a isolated rarefaction wave. The proof for this type of wave is sufficient to indicate the main ideas for shocks and linearly degenerate waves. Nonisolated waves will be discussed in section 4.2. We assume that the waves under discussion belong to the approximate solution v, omitting the similar arguments for the waves in u. Time t and the wave α will be fixed throughout the proof and the subscripts and the superscripts involving t and α will usually be removed. In particular, our notation $q_i^{\alpha\pm}$ given in (2.36) will be abbreviated to $q_i^{\pm}$. The left- and right-hand states of an isolated rarefaction wave, as given in Definition 2.5, coincide with the left- and right-sided limits of v at $\dot{x}_\alpha$, here denoted by v^- and v^+.

We begin with two interaction estimates from Bressan [5], where we have already restricted our attention to the case of a single isolated rarefaction wave. These two estimates are slight improvements over the ones presented in [6].

LEMMA 4.1. *For an isolated rarefaction wave $\alpha \in \mathcal{W}(v(t))$, we have*

$$\begin{aligned} &|q^+_{k_\alpha} - q^-_{k_\alpha} - \sigma_\alpha| + \sum_{i\neq k_\alpha} |q_i^+ - q_i^-| \\ &\qquad = \mathcal{O}(1)\Big(|\sigma_\alpha|^2 + |q^-_{k_\alpha}||\sigma_\alpha + q^-_{k_\alpha}| + \sum_{i\neq k_\alpha} |q_i^-|\Big)|\sigma_\alpha|. \end{aligned} \tag{4.1}$$

LEMMA 4.2. *Choose $\overline{w} \in \Omega, \sigma, \sigma' \in \mathbb{R},\ k \in \{1,\dots,n\}$. Define the states and the wave speeds*

$$\begin{array}{rclcrcl} w &=& S_k(\sigma)(\overline{w}), & & s &=& s_k(\overline{w}, w), \\ w' &=& S_k(\sigma')(w), & & s' &=& s_k(w, w'), \\ w'' &=& S_k(\sigma+\sigma')(\overline{w}), & & s'' &=& s_k(\overline{w}, w''). \end{array}$$

If the family k is genuinely nonlinear, then

$$\big|(\sigma+\sigma')(s''-s') - \sigma(s-s')\big| = \mathcal{O}(1)|\sigma\sigma'|\cdot|\sigma+\sigma'|. \tag{4.2}$$

During the proof of Lemma 3.1 the possibly large value of κ_1 will be mitigated by other factors appearing in the expansions.

LEMMA 4.3. *For any positive value of κ_1, there exist a positive constant δ_3, depending only on κ_1, the system* (2.1)*, and Ω such that if u and v are front-tracking approximations satisfying*

$$||u(0)||_{L^\infty} + ||u(0)||_{\mathrm{TV}},\ ||v(0)||_{L^\infty} + ||v(0)||_{\mathrm{TV}} < \delta_3, \tag{4.3}$$

then

$$\kappa_1 \sum_{i=1}^{n} |q_i(x)| \le 1 \qquad \forall x \in \mathbb{R}\,, \tag{4.4}$$

$$\kappa_1 |\sigma_\alpha| \le 1 \qquad \forall \alpha \in \mathcal{W}\big(v(t)\big). \tag{4.5}$$

Proof. The distance $\sum q_i(x)$ is bounded by the sum of the total variation of u and v measured between x and infinity since both approximations coincide and vanish there. Using the bound (2.40) completes the proof. ∎

The proof of Lemma 3.1 for isolated rarefaction waves can be derived from the estimates contained in Chapter 8 of [5]. It suffices to observe that every occurrence in [5] of a factor ϵ corresponds to either an error in wave speeds $|\dot{x}_\alpha - s_{k_\alpha}(v^-, v^+)|$ or a bound on $|\sigma_\alpha|$. Inequalities (2.28) and (2.29) can then be used to bound the resulting quantities by discrepancies. For the reader's convenience we provide the details for the important case of a rarefaction wave, since the verification of these facts would be straightforward but tedious.

LEMMA 4.4. *Under the assumptions of Lemma* 3.1*, there exist positive constants $\delta, \kappa_1, \kappa_2$, depending only on the system* (2.1) *and Ω, such that, for any pair of front-tracking approximations satisfying* (3.7) *and any isolated rarefaction wave $\alpha \in \mathcal{W}(v(t))$, we have*

$$\sum_{i=1}^{n} E_i(\alpha) \le \mathcal{O}(1) D_\alpha(t). \tag{4.6}$$

Proof. Overall, the proof involves three steps. The first step is a decomposition of $E_i(\alpha)$, which permits a simplification to the case of a shock with $\sigma_\alpha > 0$. In the second step, estimates are derived for $E_i(\alpha)$, $i \neq k_\alpha$. Finally, we bound the remaining term $E_{k_\alpha}(\alpha)$. Our goal is to indicate the changes to the original proof in [5] required to obtain (4.6). For this purpose, it is sufficient to assume that both $q^-_{k_\alpha}$ and $q^+_{k_\alpha}$ are positive. Throughout, we also assume that the total variation and the L^∞ norms of the initial data are sufficiently small to guarantee the conclusions of Lemmas 2.9 and 4.2.

We begin by defining the state and the wave speed

$$v^\diamond = S_{k_\alpha}(\sigma_\alpha)(v^-), \qquad \dot{x}^\diamond_\alpha = s_{k_\alpha}(v^-, v^\diamond). \tag{4.7}$$

Let $q_i^\diamond$ denote the virtual waves satisfying

$$v^\diamond = S_n\big(q_n^\diamond\big) \circ \cdots \circ S_1\big(q_1^\diamond\big)\big(u(x_\alpha, t)\big). \tag{4.8}$$

Associated to these virtual waves, we have intermediate states $w_0^\diamond = u(x_\alpha, t)$,

$$w_i^\diamond \equiv S_i\big(q_i^\diamond\big) \circ \cdots \circ S_1\big(q_1^\diamond\big)\big(u(x_\alpha, t)\big), \qquad i = 1, \dots, n, \tag{4.9}$$

with speeds

$$s_i^\diamond = s_i(w_i^\diamond, w_{i-1}^\diamond). \tag{4.10}$$

Shock and rarefaction curves are tangent to second order, and so the following identities hold:

$$\begin{aligned} |v^\diamond - v^+| &= \mathcal{O}(1)|\sigma_\alpha|^3, & |w_i^\diamond - w_i^+| &= \mathcal{O}(1)|\sigma_\alpha|^3, \\ |q_i^\diamond - q_i^+| &= \mathcal{O}(1)|\sigma_\alpha|^3, & |s_i^\diamond - s_i^+| &= \mathcal{O}(1)|\sigma_\alpha|^3. \end{aligned} \tag{4.11}$$

Following Bressan [5], we decompose $E_i(\alpha)$ into one equivalent expression involving only shock curves and two others measuring the error introduced by changing over to shock curves.

$$\begin{aligned} E_i(\alpha) &= W_i^+|q_i^+|(s_i^+ - \dot{x}_\alpha) - W_i^-|q_i^-|(s_i^- - \dot{x}_\alpha) \\ &= W_i^+|q_i^\diamond|(s_i^\diamond - \dot{x}_\alpha^\diamond) - W_i^-|q_i^-|(s_i^- - \dot{x}_\alpha^\diamond) \\ &\quad + \Big\{ W_i^+|q_i^\diamond|(s_i^+ - s_i^\diamond) + W_i^+\big(|q_i^+| - |q_i^\diamond|\big)(s_i^+ - \dot{x}_\alpha^\diamond) \Big\} \\ &\quad + (\dot{x}_\alpha^\diamond - \dot{x}_\alpha)\Big\{ W_i^+\big(|q_i^+| - |q_i^-|\big) + \big(W_i^+ - W_i^-\big)|q_i^-| \Big\} \\ &\doteq E_i'(\alpha) + E_i''(\alpha) + E_i'''(\alpha). \end{aligned} \tag{4.12}$$

Using estimates (4.11), we can verify that

$$E_i''(\alpha) = \mathcal{O}(1)|\sigma_\alpha|^3. \tag{4.13}$$

Furthermore, this quantity is bounded by the discrepancy since the minimum of the discrepancy of a rarefaction is second order (2.29). This is carried out below.

$$E_i''(\alpha) = \mathcal{O}(1)|\sigma_\alpha|^3 = \left[\frac{\mathcal{O}(1)|\sigma_\alpha|^3}{D_\alpha(t)}\right] D_\alpha(t) \le \mathcal{O}(1)|\sigma_\alpha| D_\alpha(t) = \mathcal{O}(1) D_\alpha(t). \tag{4.14}$$

To bound $E_i'''(\alpha)$, notice that Lemma 2.8 yields

$$W_i^+ - W_i^- = \mathcal{O}(1)\kappa_1|\sigma_\alpha|. \tag{4.15}$$

Recall that, to second order, the shock speeds are the average of the characteristic velocities of the states neighboring a shock. We also observe that when $q_i^+ q_i^- > 0$, then

$$\big||q_i^+| - |q_i^-|\big| = |q_i^+ - q_i^-| = \mathcal{O}(1)|\sigma_\alpha|.$$

With the help of the upper bound (2.28), Lemma 4.3, and the two previous relations we can estimate the third term in (4.12) as

$$\begin{aligned} E_i'''(\alpha) &= \mathcal{O}(1)|\dot{x}_\alpha - \dot{x}_\alpha^\diamond||\sigma_\alpha| \\ &\le \mathcal{O}(1)\Big|\dot{x}_\alpha - \frac{1}{2}\big(\lambda_{k_\alpha}(v^-) + \lambda_{k_\alpha}(v^+)\big)\Big||\sigma_\alpha| \\ &\quad + \mathcal{O}(1)\Big|\frac{1}{2}\big(\lambda_{k_\alpha}(v^-) + \lambda_{k_\alpha}(v^+)\big) - s_{k_\alpha}(v^-, v^+)\Big||\sigma_\alpha| \\ &\quad + \mathcal{O}(1)\Big|s_{k_\alpha}(v^-, v^+) - \dot{x}_\alpha^\diamond\Big||\sigma_\alpha| \\ &= \mathcal{O}(1) D_\alpha(t) + \mathcal{O}(1)|\sigma_\alpha|^3 = \mathcal{O}(1) D_\alpha(t). \end{aligned} \tag{4.16}$$

When $q_i^+ q_i^- \le 0$ then both q_i^+ and q_i^- must be of the order of σ_α and the calculation (4.16) is still applicable. It therefore remains to show that

$$E_i'(\alpha) \le \mathcal{O}(1) D_\alpha(t). \tag{4.17}$$

Before we can study (4.17) we need a few preliminary estimates. Define the sets of waves

$$\mathcal{I} = \big\{ i \big| i \neq k_\alpha, q_i^+, q_i^-, q_i^\diamond \text{ have the same sign} \big\},$$
$$\mathcal{I}' = \big\{ i \big| i \neq k_\alpha, q_i^+, q_i^-, q_i^\diamond \text{ are not all of the same sign} \big\}.$$

The virtual waves indexed by $\mathcal{I}'$ are in some sense smaller than those in $\mathcal{I}$. Combining identities (4.11) and Lemma 4.1 produces

$$\begin{aligned} &|q_{k_\alpha}^\diamond - q_{k_\alpha}^- - \sigma_\alpha| + \sum_{i \neq k_\alpha} |q_i^\diamond - q_i^-| \\ &\qquad = \mathcal{O}(1)\Big(|\sigma_\alpha|^2 + |q_{k_\alpha}^-||\sigma_\alpha + q_{k_\alpha}^-| + \sum_{i \neq k_\alpha} |q_i^-| \Big) |\sigma_\alpha|. \end{aligned} \tag{4.18}$$

When $i \in \mathcal{I}'$ and $q_i^- q^\diamond < 0$, this implies that

$$|q_i^\diamond| + |q_i^-| = |q_i^\diamond - q_i^-| = \mathcal{O}(1)\Big(|\sigma_\alpha|^2 + |q_{k_\alpha}^-||\sigma_\alpha + q_{k_\alpha}^-| + \sum_{i \neq k_\alpha} |q_i^-| \Big) |\sigma_\alpha|. \tag{4.19}$$

If $q_i^- q^\diamond \geq 0$, then (4.19) still holds because of (4.11). Apply this identity to each wave $i \in \mathcal{I}'$ and suppose that the total variation bound (3.7) is small enough to ensure

$$\begin{aligned} \sum_{i \in \mathcal{I}'} |q_i^-| &\leq n\mathcal{O}(1)|\sigma_\alpha| \Big(|\sigma_\alpha|^2 + |q_{k_\alpha}^-||\sigma_\alpha + q_{k_\alpha}^-| + \sum_{i \neq k_\alpha} |q_i^-| \Big) \\ &\leq \frac{1}{2} \Big(|\sigma_\alpha|^2 + |q_{k_\alpha}^-||\sigma_\alpha + q_{k_\alpha}^-| + \sum_{i \neq k_\alpha} |q_i^-| \Big). \end{aligned} \tag{4.20}$$

With this last bound, we can check that

$$\sum_{i \neq k_\alpha} |q_i^-| = \sum_{i \in \mathcal{I}'} |q_i^-| + \sum_{i \in \mathcal{I}} |q_i^-| \leq \Big(|\sigma_\alpha|^2 + |q_{k_\alpha}^-||\sigma_\alpha + q_{k_\alpha}^-| \Big) + 2 \sum_{i \in \mathcal{I}} |q_i^-|. \tag{4.21}$$

In the remainder of this proof, we will often use this bound to restrict ourselves to the sum of strengths of waves in $\mathcal{I}$. The next property of the auxiliary waves $q_i^\diamond$ with $i \neq k_\alpha$ can be derived from identities (4.18) and (4.21).

$$\begin{aligned} q_i^\diamond(s_i^\diamond - \dot{x}_\alpha^\diamond) - q_i^-(s_i^- - \dot{x}_\alpha^\diamond) &= (q_i^\diamond - q_i^-)(s_i^\diamond - \dot{x}_\alpha^\diamond) - q_i^-(s_i^- - s_i^\diamond) \\ &= \mathcal{O}(1)\Big(|\sigma_\alpha|^2 + |q_{k_\alpha}^-||\sigma_\alpha + q_{k_\alpha}^-| + \sum_{i \in \mathcal{I}} |q_i^-| \Big) |\sigma_\alpha|. \end{aligned} \tag{4.22}$$

Finally, we make the observation that if the sign of $q_{k_\alpha}^\diamond$ is different from the sign of $q_{k_\alpha}^-$, then (4.18) supplies the bound

$$|q_{k_\alpha}^\diamond| = \mathcal{O}(1)\Big(|\sigma_\alpha|^2 + |q_{k_\alpha}^-||\sigma_\alpha + q_{k_\alpha}^-| + \sum_{i \in \mathcal{I}} |q_i^-| \Big) |\sigma_\alpha|. \tag{4.23}$$

This estimate essentially allows us to replace $q_{k_\alpha}^\diamond$, $q_{k_\alpha}^+$, and $q_{k_\alpha}^-$ by $|\sigma_\alpha|$ when those three waves are not of the same sign. Under those circumstances the remaining work would be greatly simplified but no longer illustrative of these techniques. Henceforth, we assume that the three virtual waves $q_{k_\alpha}^\diamond$, $q_{k_\alpha}^+$, and $q_{k_\alpha}^-$ are positive.

Returning to our proof of (4.17) for an isolated rarefaction wave, we distinguish between two cases, $i \neq k_\alpha$ and $i = k_\alpha$. To examine the first case, we begin by assuming that $i \in \mathcal{I}$. According to Lemma 2.8,

$$\operatorname{sign}(W_i^+ - W_i^-) = \operatorname{sign}(k_\alpha - i) = -\operatorname{sign}(s_i^- - \dot{x}_\alpha^\diamond). \tag{4.24}$$

Using this observation, the definition (2.8) of d, and (4.22), we verify that

$$\begin{aligned} E_i'(\alpha) =& W_i^+|q_i^\diamond|(s_i^\diamond - \dot{x}_\alpha^\diamond) - W_i^-|q_i^-|(s_i^- - \dot{x}_\alpha^\diamond) \\ =& \big(W_i^+ - W_i^-\big)|q_i^-|(s_i^- - \dot{x}_\alpha^\diamond) + W_i^+\Big\{|q_i^\diamond|(s_i^\diamond - \dot{x}_\alpha^\diamond) - |q_i^-|(s_i^- - \dot{x}_\alpha^\diamond)\Big\} \\ =& -\kappa_1 d|\sigma_\alpha||q_i^-| + \mathcal{O}(1)\Big(|\sigma_\alpha|^2 + |q_{k_\alpha}^-||\sigma_\alpha + q_{k_\alpha}^-| + \sum_{i\in\mathcal{I}}|q_i^-|\Big)|\sigma_\alpha|. \end{aligned} \tag{4.25}$$

When $i \in \mathcal{I}'$, expression (4.20) states that estimates for the virtual waves in $\mathcal{I}'$ involve only interaction terms, and therefore, from the definition of $E_i'(\alpha)$ in (4.12), it follows that

$$E_i'(\alpha) = \mathcal{O}(1)\Big(|\sigma_\alpha|^2 + |q_{k_\alpha}^-||\sigma_\alpha + q_{k_\alpha}^-| + \sum_{i\in\mathcal{I}}|q_i^-|\Big)|\sigma_\alpha|. \tag{4.26}$$

The third order terms in $|\sigma_\alpha|$ are bounded by the discrepancy, as seen in (4.14), and therefore (4.25) and (4.26) can be combined to obtain

$$\begin{aligned} \sum_{i\neq k_\alpha} E_i'(\alpha) =& -\kappa_1 d|\sigma_\alpha|\sum_{i\in\mathcal{I}}|q_i^-| + \mathcal{O}(1)D_\alpha(t) \\ &+ \mathcal{O}(1)\Big(|q_{k_\alpha}^-||\sigma_\alpha + q_{k_\alpha}^-| + \sum_{i\in\mathcal{I}}|q_i^-|\Big)|\sigma_\alpha|. \end{aligned} \tag{4.27}$$

We complete the proof by considering the last case in (4.17) with $i = k_\alpha$. Rewrite $E_{k_\alpha}'(\alpha)$ as follows:

$$\begin{aligned} E_{k_\alpha}'(\alpha) &= W_{k_\alpha}^+|q_{k_\alpha}^\diamond|(s_{k_\alpha}^\diamond - \dot{x}_\alpha^\diamond) - W_{k_\alpha}^-|q_{k_\alpha}^-|(s_{k_\alpha}^- - \dot{x}_\alpha^\diamond) \\ &= \big(W_{k_\alpha}^+ - W_{k_\alpha}^-\big)|q_{k_\alpha}^-|(s_{k_\alpha}^- - \dot{x}_\alpha^\diamond) + W_{k_\alpha}^+\Big\{|q_{k_\alpha}^\diamond|(s_{k_\alpha}^\diamond - \dot{x}_\alpha^\diamond) - |q_{k_\alpha}^-|(s_{k_\alpha}^- - \dot{x}_\alpha^\diamond)\Big\}. \end{aligned} \tag{4.28}$$

Lemma (2.8) states that when the signs of the virtual waves are all positive then the change in the weight W_{k_α} across the discontinuity is

$$W_{k_\alpha}^+ - W_{k_\alpha}^- = \kappa_1|\sigma_\alpha|. \tag{4.29}$$

Introduce the state and the wave speed

$$\widetilde{w} = S_{k_\alpha}(\sigma_\alpha)(w_{k_\alpha}^-) \qquad \text{and} \qquad \widetilde{s} = s_{k_\alpha}(w_{k_\alpha}^-, \widetilde{w}). \tag{4.30}$$

The waves centered at v^- and $w_{k_\alpha}^-$ and ending, respectively, at $v^\diamond$ and $\widetilde{w}$ have the same strength. Their associated shock speeds must therefore be proportional to the distance separating v^- and $w_{k_\alpha}^-$, namely,

$$|\widetilde{s} - \dot{x}_\alpha^\diamond| = \mathcal{O}(1)\sum_{i\neq k_\alpha}|q_i^-|. \tag{4.31}$$

Approximating the shock speeds by the average of the characteristic velocities generates the second-order errors

$$s^-_{k_\alpha} = \frac{1}{2}\big(\lambda_{k_\alpha}(w^-_{k_\alpha-1}) + \lambda_{k_\alpha}(w^-_{k_\alpha})\big) + \mathcal{O}(1)|q^-_{k_\alpha}|^2, \tag{4.32}$$

$$\widetilde{s} = \frac{1}{2}\big(\lambda_{k_\alpha}(w^-_{k_\alpha}) + \lambda_{k_\alpha}(\widetilde{w})\big) + \mathcal{O}(1)|\sigma_\alpha|^2. \tag{4.33}$$

To the first term in our decomposition (4.28) of $E'_{k_\alpha}(\alpha)$, we apply the bounds (4.29), (4.31), and the previous two estimates

$$\begin{aligned}
&(W^+_{k_\alpha} - W^-_{k_\alpha})|q^-_{k_\alpha}|(s^-_{k_\alpha} - \dot{x}^\diamond_\alpha)\\
&= \kappa_1|\sigma_\alpha||q^-_{k_\alpha}|(s^-_{k_\alpha} - \widetilde{s}) + \kappa_1|\sigma_\alpha||q^-_{k_\alpha}|(\widetilde{s} - \dot{x}^\diamond_\alpha)\\
&= \kappa_1|\sigma_\alpha||q^-_{k_\alpha}|\frac{1}{2}\Big(\lambda_{k_\alpha}(w^-_{k_\alpha-1}) - \lambda_{k_\alpha}(\widetilde{w})\Big)\\
&\quad + \mathcal{O}(1)\kappa_1|\sigma_\alpha||q^-_{k_\alpha}|\Big(|\sigma_\alpha|^2 + |q^-_{k_\alpha}|^2\Big) + \mathcal{O}(1)|\sigma_\alpha|\sum_{i\neq k_\alpha}|q^-_i|.
\end{aligned} \tag{4.34}$$

With the help of the parameterization (2.5) for shock curves, we compute that for a rarefaction α and positive virtual waves

$$\lambda_{k_\alpha}(w^-_{k_\alpha-1}) - \lambda_{k_\alpha}(\widetilde{w}) = -(\sigma_\alpha + q^-_{k_\alpha}). \tag{4.35}$$

Substituting this identity into (4.34) and using (4.21) result in the estimate

$$\begin{aligned}
(W^+_{k_\alpha} - W^-_{k_\alpha})|q^-_{k_\alpha}|(s^-_{k_\alpha} - \dot{x}^\diamond_\alpha) \le &-\frac{1}{2}\kappa_1|\sigma_\alpha||q^-_{k_\alpha}||\sigma_\alpha + q^-_{k_\alpha}| + \mathcal{O}(1)D_\alpha(t)\\
&+ \mathcal{O}(1)\Big(|q^-_{k_\alpha}||\sigma_\alpha + q^-_{k_\alpha}| + \sum_{i\in\mathcal{I}}|q^-_i|\Big)|\sigma_\alpha|.
\end{aligned} \tag{4.36}$$

The second term in the decomposition (4.28) of $E'_{k_\alpha}(\alpha)$ requires the use of an additional intermediate state. Recall the definition (4.30) of $\widetilde{w}$ and introduce

$$w^* = S_{k_\alpha}(q^-_{k_\alpha} + \sigma_\alpha)(w^-_{k_\alpha-1}) \qquad \text{and} \qquad s^* = s_{k_\alpha}(w^-_{k_\alpha-1}, w^*). \tag{4.37}$$

Using the positivity of the virtual waves, we reorganize the second term in (4.28) in the following manner:

$$\begin{aligned}
&W^+_{k_\alpha}\Big\{|q^\diamond_{k_\alpha}|(s^\diamond_{k_\alpha} - \dot{x}^\diamond_\alpha) - |q^-_{k_\alpha}|(s^-_{k_\alpha} - \dot{x}^\diamond_\alpha)\Big\}\\
&= W^+_{k_\alpha}\Big\{|q^\diamond_{k_\alpha} - q^-_{k_\alpha} - \sigma_\alpha|(s^\diamond_{k_\alpha} - \dot{x}^\diamond_\alpha) + |q^-_{k_\alpha} + \sigma_\alpha|(s^\diamond_{k_\alpha} - s^*)\\
&\qquad + |q^-_{k_\alpha} + \sigma_\alpha|(s^* - \widetilde{s}) + |q^-_{k_\alpha} + \sigma_\alpha|(\widetilde{s} - \dot{x}^\diamond_\alpha)\\
&\qquad - |q^-_{k_\alpha}|(s^-_{k_\alpha} - \widetilde{s}) - |q^-_{k_\alpha}|(\widetilde{s} - \dot{x}^\diamond_\alpha)\Big\}\\
&= W^+_{k_\alpha}\Big\{|q^-_{k_\alpha} + \sigma_\alpha|(s^* - \widetilde{s}) - |q^-_{k_\alpha}|(s^-_{k_\alpha} - \widetilde{s})\\
&\qquad + |q^\diamond_{k_\alpha} - q^-_{k_\alpha} - \sigma_\alpha|(s^\diamond_{k_\alpha} - \dot{x}^\diamond_\alpha)\\
&\qquad + |q^-_{k_\alpha} + \sigma_\alpha|(s^\diamond_{k_\alpha} - s^*) - |\sigma_\alpha|(\widetilde{s} - \dot{x}^\diamond_\alpha)\Big\}.
\end{aligned} \tag{4.38}$$

We may ignore the factor $W^+_{k_\alpha}$ because of the bounds (2.32) on W_i. We are then left with five terms. The expression presented in Lemma 4.2 with $\bar{w} = w^-_{k_\alpha - 1}$, $\sigma = q^-_{k_\alpha}$, and $\sigma' = \sigma_\alpha$ is actually the first two terms in (4.38). Therefore, these terms are of order

$$\mathcal{O}(1)|\sigma_\alpha||q^-_{k_\alpha}||\sigma_\alpha + q^-_{k_\alpha}|. \tag{4.39}$$

The third term in (4.38) is bounded by an interaction term of order

$$\mathcal{O}(1)\Big(|q^-_{k_\alpha}||\sigma_\alpha + q^-_{k_\alpha}| + \sum_{i\in\mathcal{I}} |q^-_i|\Big)|\sigma_\alpha| + \mathcal{O}(1)D_\alpha(t). \tag{4.40}$$

The fourth term in (4.38) involves the factor

$$\begin{aligned} s^\diamond_{k_\alpha} - s^* =& \mathcal{O}(1)\big|w^\diamond_{k_\alpha - 1} - w^-_{k_\alpha - 1}\big| \\ =& \mathcal{O}(1)\Big(|q^-_{k_\alpha}||\sigma_\alpha + q^-_{k_\alpha}| + \sum_{i\in\mathcal{I}} |q^-_i|\Big)|\sigma_\alpha|. \end{aligned} \tag{4.41}$$

From definitions (4.7) and (4.30), it is clear that

$$\begin{aligned} |\sigma_\alpha||\widetilde{s} - \dot{x}^\diamond_\alpha| =& \mathcal{O}(1)|\sigma_\alpha|\big|w^-_{k_\alpha} - v^-\big| \\ =& \mathcal{O}(1)|\sigma_\alpha| \sum_{i\neq k_\alpha} |q^-_i|. \end{aligned} \tag{4.42}$$

A bound on the second term in (4.28) can be found be summing the estimates (4.39)–(4.42). After an additional application of (2.29) and (4.21) to this sum, we conclude that

$$\begin{aligned} & W^+_{k_\alpha}\Big\{|q^\diamond_{k_\alpha}|(s^\diamond_{k_\alpha} - \dot{x}^\diamond_\alpha) - |q^-_{k_\alpha}|(s^-_{k_\alpha} - \dot{x}^\diamond_\alpha)\Big\} \\ & \qquad \le \mathcal{O}(1)D_\alpha(t) + \mathcal{O}(1)\Big(|q^-_{k_\alpha}||\sigma_\alpha + q^-_{k_\alpha}| + \sum_{i\in\mathcal{I}} |q^-_i|\Big)|\sigma_\alpha|. \end{aligned} \tag{4.43}$$

Combining the local estimates (4.36) and (4.43), we have thus demonstrated that

$$\begin{aligned} E'_{k_\alpha}(\alpha) \le & -\frac{1}{2}\kappa_1|\sigma_\alpha||q^-_{k_\alpha}||\sigma_\alpha + q^-_{k_\alpha}| + \mathcal{O}(1)D_\alpha(t) \\ & + \mathcal{O}(1)\Big(|q^-_{k_\alpha}||\sigma_\alpha + q^-_{k_\alpha}| + \sum_{i\in\mathcal{I}} |q^-_i|\Big)|\sigma_\alpha|. \end{aligned} \tag{4.44}$$

So far, estimates (4.27) and (4.44) provide us with

$$\begin{aligned} \sum_{i=1}^n E'_i(\alpha) \le & \mathcal{O}(1)D_\alpha(t) + \big(\mathcal{O}(1) - \kappa_1 d\big)|\sigma_\alpha| \sum_{i\in\mathcal{I}} |q^-_i| \\ & + \left(\mathcal{O}(1) - \frac{1}{2}\kappa_1\right)|\sigma_\alpha||q^-_{k_\alpha}||\sigma_\alpha + q^-_{k_\alpha}|, \end{aligned} \tag{4.45}$$

which is close to our goal of (4.6) since the remaining terms E''_i and E'''_i in (4.12) have already been shown to be of order $\mathcal{O}(1)|\sigma_\alpha|$. Recall that $\mathcal{O}(1)$ depends on f, Ω, and the initial data but is independent of u and v. We can now take κ_1 sufficiently large to make the last two terms negative, possibly at the cost of decreasing the total variation bound (2.39) in Lemma 2.9. The values for δ, κ_1, and κ_2 in Lemma 2.9 can therefore be chosen a priori. The final estimate is a posteriori since the discrepancies depend only on the approximate solutions. ∎

4.2. Nonisolated waves. In this section, we show that the local estimates of Lemma 3.1 for discontinuities containing more than one wave can be deduced from our earlier estimates for isolated waves. In the ϵ-approximations of [5, 6], nonisolated waves were weak and did not interact in a consistent manner with other waves. The estimates of Lemma 2.8 therefore did not hold for these waves and their local estimates were a consequence of the fact these waves were a priori weak. For our a posteriori error estimates, we must show that the discrepancy measures the local error even when it is small. We therefore present a new argument for the nonisolated waves in Risebro's front-tracking approximations.

The front-tracking approximations of Risebro give rise to discontinuities containing more than one wave when wave interactions satisfy condition (2.14). In this case the approximate Riemann solver limits the number of outgoing discontinuities by merging n discontinuities, each of which can be identified with an isolated wave, into two discontinuities, each composed of several waves. To demonstrate the estimate of Lemma 3.1 for one of these discontinuities, we construct an auxiliary front-tracking approximation where the discontinuity of the original approximation has been replaced locally by n isolated waves. In this way, the original approximation can be obtained as the limit of a sequence of these auxiliary approximations. The estimates for the isolated waves of the auxiliary approximations are then also shown to pass to the limit.

Consider a front-tracking approximation v with a single discontinuity located at $z(t)$ and containing n waves in $\mathcal{W}(v(t))$ denoted $\alpha(1)$, $\alpha(2), \dots,$ $\alpha(n)$. Assume that these waves are ordered by the requirement that $k_{\alpha(i)} < k_{\alpha(i+1)}$. It suffices to demonstrate Lemma 3.1 at a time t_0 when discontinuities are not colliding. Therefore, for any point $(z(t_0), t_0)$ on the discontinuity but away from a collision we can find an $\epsilon > 0$ such that there is only one discontinuity within the rectangle

$$B_\epsilon = [z(t_0) - \epsilon, z(t_0) + \epsilon] \times [t_0 - \epsilon, t_0 + \epsilon].$$

If v^- and v^+ are the neighboring states at $z(t)$, let $\widetilde{v}_0, \widetilde{v}_1, \dots, \widetilde{v}_n$ be the $n+1$ states of the exact Riemann solution of section 2.1. For every $\epsilon \geq 0$ construct a family of smooth and disjoint trajectories $\tau_{i,\epsilon}$, $i = 1, \dots, n$, inside of B_ϵ that satisfy

$$\tau_{i,\epsilon}(t) < \tau_{i+1,\epsilon}(t), \tag{4.46}$$

$$\tau_{i,\epsilon}(t) = z(t), \qquad \text{when } |t - t_0| \geq \epsilon, \text{ and} \tag{4.47}$$

$$\lim_{\epsilon \to 0} \dot{\tau}_{i,\epsilon}(t) = \dot{z}(t). \tag{4.48}$$

A family of auxiliary approximations can then be constructed as

$$V_\epsilon(x,t) = \begin{cases} \widetilde{v}_i & \text{if } (x,t) \in B_\epsilon \text{ and } \tau_{i,\epsilon}(t) \leq x \leq \tau_{i+1,\epsilon}(t),\ i = 1, \dots, n-1, \\ v(x,t), & \text{otherwise.} \end{cases} \tag{4.49}$$

In the approximation V_ϵ, the single discontinuity along z originally in v is split inside B_ϵ into n isolated waves. It is easy to see that V_ϵ is still a front-tracking approximation and that the results of Lemmas 2.3 and 2.9 continue to hold. In particular, the Liu–Yang functional is well defined for V_ϵ

$$\Phi\big(u(t), V_\epsilon(t)\big) = \sum_{k=1}^{n} \int_{-\infty}^{\infty} |q_{k,\epsilon}(x)| W_{k,\epsilon}(x)\, dx, \tag{4.50}$$

where we have used ϵ to distinguish it from the integrand in the functional

$$\Phi\big(u(t),v(t)\big)=\sum_{k=1}^{n}\int_{-\infty}^{\infty}|q_k(x)|W_k(x)\,dx.$$

In general, we will use an additional subscript ϵ to denote any quantity derived from (4.50). Recall the notation

$$q_k^{\beta\pm}=\lim_{x\to x_\beta\pm}q_k(x), \tag{4.51}$$

which will also be used for any function of x.

LEMMA 4.5.

$$\lim_{\epsilon\to0}\sum_{i=1}^{n}\sum_{k=1}^{n}E_{k,\epsilon}\big(\alpha(i)\big)=\sum_{k=1}^{n}E_k(z(t_0)). \tag{4.52}$$

Proof. For $\alpha(i)$ and $\alpha(i+1)$ compare the two contributions from the time derivative of $\Phi\big(u(t),V_\epsilon(t)\big)$:

$$\begin{aligned}E_{k,\epsilon}\big(\alpha(i)\big)=&\big|q_{k,\epsilon}^{\alpha(i)+}\big|W_{k,\epsilon}^{\alpha(i)+}\big(s_{k,\epsilon}^{\alpha(i)+}-\dot\tau_i\big)\\&-\big|q_{k,\epsilon}^{\alpha(i)-}\big|W_{k,\epsilon}^{\alpha(i)-}\big(s_{k,\epsilon}^{\alpha(i)-}-\dot\tau_i\big),\end{aligned} \tag{4.53}$$

$$\begin{aligned}E_{k,\epsilon}\big(\alpha(i+1)\big)=&\big|q_{k,\epsilon}^{\alpha(i+1)+}\big|W_{k,\epsilon}^{\alpha(i+1)+}\big(s_{k,\epsilon}^{\alpha(i+1)+}-\dot\tau_{i+1}\big)\\&-\big|q_{k,\epsilon}^{\alpha(i+1)-}\big|W_{k,\epsilon}^{\alpha(i+1)-}\big(s_{k,\epsilon}^{\alpha(i+1)-}-\dot\tau_{i+1}\big).\end{aligned} \tag{4.54}$$

Using Definition 2.5,

$$q_{k,\epsilon}^{\alpha(i)+}=q_k(u,v^+_{\alpha(i)})=q_k(u,v^-_{\alpha(i+1)})=q_{k,\epsilon}^{\alpha(i+1)-}. \tag{4.55}$$

The same procedure also shows that

$$s_{k,\epsilon}^{\alpha(i)+}=s_{k,\epsilon}^{\alpha(i+1)-}\text{ and }W_{k,\epsilon}^{\alpha(i)+}=W_{k,\epsilon}^{\alpha(i+1)-}. \tag{4.56}$$

Moreover, notice that as $\epsilon\to0$ both $\dot\tau_i$ and $\dot\tau_{i+1}$ approach $\dot z$. For fixed k, in the limit as $\epsilon\to0$, the terms on the left-hand side of (4.52) form a telescoping sum and only the first and last terms remain. The limit can therefore be reduced to

$$\begin{aligned}\lim_{\epsilon\to0}\sum_{i=1}^{n}\sum_{k=1}^{n}E_{k,\epsilon}\big(\alpha(i)\big)=\lim_{\epsilon\to0}\sum_{k=1}^{n}&\big|q_{k,\epsilon}^{\alpha(n)+}\big|W_{k,\epsilon}^{\alpha(n)+}\big(s_{k,\epsilon}^{\alpha(n)+}-\dot z\big)\\&-\big|q_{k,\epsilon}^{\alpha(1)-}\big|W_{k,\epsilon}^{\alpha(1)-}\big(s_{k,\epsilon}^{\alpha(1)-}-\dot z\big).\end{aligned} \tag{4.57}$$

For the waves $\alpha(1)$ and $\alpha(n)$ we have

$$\begin{aligned}q_{k,\epsilon}^{\alpha(1)-}&=q_k(u,v^-_{\alpha(1)})=q_k(u,v^-)=q_k^{z-},\\q_{k,\epsilon}^{\alpha(n)+}&=q_k(u,v^+_{\alpha(n)})=q_k(u,v^+)=q_k^{z+},\end{aligned} \tag{4.58}$$

and

$$s_{k,\epsilon}^{\alpha(1)-}=s_k^{z-}\text{ and }s_{k,\epsilon}^{\alpha(n)+}=s_k^{z+}. \tag{4.59}$$

We now examine the weight

$$W_{k,\epsilon}^{\alpha(1)-} = 1 + \kappa_1\big(A_{k,\epsilon}^{\alpha(1)-} + B_{k,\epsilon}^{\alpha(1)-}\big) + \kappa_2\big(Q(u) + Q(V_\epsilon)\big). \tag{4.60}$$

We know that $Q(V_\epsilon) = Q(v)$ since both approximations contain the same waves. From the definition of A_i, (2.34), we find

$$\begin{aligned}\lim_{\epsilon\to 0} A_{k,\epsilon}^{\alpha(1)-} =& \lim_{\epsilon\to 0}\Big(\sum_{\substack{\beta\in\mathcal{W}(u)\bigcup\mathcal{W}(V_\epsilon)\\ x_\beta<\tau_1,\, k_\beta>k}} + \sum_{\substack{\beta\in\mathcal{W}(u)\bigcup\mathcal{W}(V_\epsilon)\\ x_\beta\geq\tau_1,\, k_\beta<k}}\Big)|\sigma_\beta| \\ =& A_k^{z-}. \end{aligned} \tag{4.61}$$

One can also show that

$$\lim_{\epsilon\to 0} A_{k,\epsilon}^{\alpha(n)+} = A_k^{z+}. \tag{4.62}$$

If k is a genuinely nonlinear family, then

$$\begin{aligned}\lim_{\epsilon\to 0} B_{k,\epsilon}^{\alpha(1)-} =& \lim_{\epsilon\to 0}\begin{cases}\Big(\sum\limits_{\substack{\beta\in\mathcal{W}(u)\\ x_\beta<\tau_1,\, k_\alpha=k}} + \sum\limits_{\substack{\beta\in\mathcal{W}(V_\epsilon)\\ x_\beta\geq\tau_1,\, k_\alpha=k}}\Big)|\sigma_\beta| & \text{if } q_{k,\epsilon}^{\alpha(1)-} < 0,\\ \Big(\sum\limits_{\substack{\beta\in\mathcal{W}(V_\epsilon)\\ x_\beta<\tau_1,\, k_\beta=k}} + \sum\limits_{\substack{\beta\in\mathcal{W}(u)\\ x_\beta\geq\tau_1,\, k_\beta=k}}\Big)|\sigma_\beta| & \text{if } q_{k,\epsilon}^{\alpha(1)-} > 0\end{cases} \\ =& B_k^{z-}. \end{aligned} \tag{4.63}$$

In the same manner, one verifies that

$$\lim_{\epsilon\to 0} B_{k,\epsilon}^{\alpha(n)+} = B_k^{z+}. \tag{4.64}$$

Applying identities (4.58)–(4.64) to (4.57) completes the demonstration of the lemma. ∎

There is a similar identity relating the sum of the discrepancies of the isolated waves in V_ϵ to the discrepancies in v.

LEMMA 4.6.

$$\lim_{\epsilon\to 0} D_{\alpha(i),\epsilon}(t_0) = D_{\alpha(i)}(t_0). \tag{4.65}$$

Proof. If $\alpha(i)$ is a shock wave or a linearly degenerate wave, then use formula (2.26) to compute

$$\begin{aligned}\lim_{\epsilon\to 0} D_{\alpha(i),\epsilon}(t_0) =& \lim_{\epsilon\to 0} |\sigma_{\alpha(i)}|\big|\dot\tau_i - s_i(v_{\alpha(i)}^-, v_{\alpha(i)}^+)\big| \\ =& |\sigma_{\alpha(i)}|\big|\dot z - s_i(v_{\alpha(i)}^-, v_{\alpha(i)}^+)\big| \\ =& D_{\alpha(i)}(t_0). \end{aligned} \tag{4.66}$$

If α is a rarefaction wave, repeat the calculation with formula (2.27). ∎

We now provide an outline of the proof of Lemma 3.1 for nonisolated waves.

Proof. It suffices to demonstrate this lemma at a point (z_0, t_0). As noted earlier, V_ϵ satisfies the same bounds on the total variation as v. Summing the estimates of Lemma 3.1 for isolated waves, near (z_0, t_0) we find that

$$\sum_{i=1}^{n}\sum_{k=1}^{n} E_{k,\epsilon}\big(\alpha(i)\big) \leq \mathcal{O}(1)\sum_{i=1}^{n} D_{\alpha(i),\epsilon}(t_0). \tag{4.67}$$

Using the two previous lemmas in the limit as $\epsilon\to 0$ completes the proof. ∎

5. Concluding remarks. We have demonstrated an a posteriori error estimate for hyperbolic systems of nonlinear conservation laws which improves on the L^1 stability estimate of Bressan, Liu, and Yang [6]. The unknown factor $\mathcal{O}(1)$ in the local estimate (3.8) of Lemma 3.1 makes our discrepancy an unreliable estimate of the true error, but it may still be used for mesh refinement [17]. On the other hand, this a posteriori estimate does show that the entropy produced along discontinuities in general approximations is an appropriate measure of the error. These error processes for nonlinear discontinuities were not present in either linear systems [24] or nonlinear systems with dissipation [12, 13]. In systems of conservation laws with smooth solutions the residual of the approximation has sufficed as a practical error indicator. For numerical approximations with discontinuities, discrepancies suggest that entropy production might be an appropriate error indicator. A numerical study of entropy production in central schemes also supports its potential use as an error indicator [21].

We mention two avenues for future research. In the same way that Schatzman extended Glimm's functional to piecewise smooth approximations [23], it might be possible to use Dafermos' [8] version of the Liu–Yang functional for such approximations to demonstrate an a posteriori error estimate. This might lead to an error estimate involving both the residual in the smooth regions and the discrepancy along discontinuities. Another interesting possibility might be to demonstrate existence and stability results for the adjoint problem to nonlinear systems of conservation laws, thus answering a question raised by Houston and Süli [12]. The existence of the adjoint problem for a single conservation law has already been examined by Gimse and Risebro [9].

Acknowledgments. The author thanks James Glimm for having proposed this problem. His numerous editorial comments during early drafts were appreciated. The author is also indebted to Tai-Ping Liu for his advice and for his insightful exposition of his joint work with Tong Yang.

REFERENCES

[1] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, John Wiley & Sons, New York, 2000.

[2] P. Baiti and H. K. Jenssen, *On the front tracking algorithm*, J. Math. Anal. Appl., 217 (1998), pp. 395–404.

[3] T. J. Barth and H. Deconinck, *Error Estimation and Adaptive Discretization Methods in Computational Fluid Dynamics*, Lecture Notes in Comput. Sci. Engrg. 25, Springer-Verlag, Berlin, 2003.

[4] A. Bressan, *Global solutions of systems of conservation laws by wave-front tracking*, J. Math. Anal. Appl., 170 (1992), pp. 414–432.

[5] A. Bressan, *Hyperbolic Systems of Conservation Laws: The One-Dimensional Cauchy Problem*, Oxford Lecture Ser. Math. Appl. 20, Oxford University Press, New York, 2000.

[6] A. Bressan, T. P. Liu, and T. Yang, *L^1 stability estimates for $n \times n$ conservation laws*, Arch. Ration. Mech. Anal., 149 (1999), pp. 1–22.

[7] B. Cockburn and H. Gau, *A posteriori error estimates of general numerical methods for scalar conservation laws*, Mat. Apl. Comput., 14 (1995), pp. 37–47.

[8] C. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Grundlehren Math. Wiss. 325, Springer-Verlag, New York, 2000.

[9] T. Gimse and N. H. Risebro, *Solution of the Cauchy problem for a conservation law with a discontinuous flux function*, SIAM. J. Math. Anal., 23 (1992), pp. 635–648.

[10] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.

[11] L. Gosse and C. Makridakis, *Two a posteriori error estimates for one-dimensional scalar conservation laws*, SIAM J. Numer. Anal., 38 (2000), pp. 964–988.

[12] P. Houston and E. Süli, *Adaptive finite element approximation of hyperbolic problems*, in Error Estimation and Adaptive Discretization in Computational Fluid Dynamics, Lect. Notes Comput. Sci. Eng. 25, T. J. Barth and H. Deconick, eds., Springer-Verlag, Berlin, 2003, pp. 267–344.

[13] C. Johnson and A. Szepessy, *Adaptive finite element methods for conservation laws based on a posteriori error estimates*, Comm. Pure Appl. Math., 48 (1995), pp. 199–234.

[14] B. Keyfitz, *Solutions with shocks, an example of an L^1-contractive semigroup*, Comm. Pure Appl. Math., 24 (1971), pp. 165–170.

[15] S. Kružkov, *Generalized solutions of the Cauchy problem in the large for first order nonlinear equations*, Dokl. Akad. Nauk SSSR, 187 (1969), pp. 29–32 (in Russian).

[16] M. Laforest, *A Posteriori Error Estimate for Front-Tracking*, Ph.D. thesis, State University of New York at Stony Brook, Stony Brook, NY, 2001.

[17] M. Laforest, *Entropy and A Posteriori Error Estimate for Conservation Laws*, in preparation.

[18] P. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[19] T.-P. Liu and T. Yang, *L_1 stability for 2×2 systems of hyperbolic conservation laws*, J. Amer. Math. Soc., 12 (1999), pp. 729–774.

[20] H. Nessyahu and E. Tadmor, *The convergence rate of approximate solutions for nonlinear scalar conservation laws*, SIAM J. Numer. Anal., 29 (1992), pp. 1505–1519.

[21] G. Puppo, *Numerical entropy production on shocks and smooth transitions*, J. Sci. Comput., 17 (2002), pp. 287–296.

[22] N. H. Risebro, *A front-tracking alternative to the random choice method*, Proc. Amer. Math. Soc., 117 (1993), pp. 1125–1139.

[23] M. Schatzman, *Continuous Glimm functionals and uniqueness of solutions of the Riemann problem*, Indiana Univ. Math. J., 34 (1985), pp. 533–589.

[24] E. Süli, *A-posteriori error analysis and adaptivity for finite element approximations of hyperbolic problems*, in An Introduction to Recent Developments in Theory and Numerics for Conservation Laws, M. O. D. Kröner and C. Rohde, eds., Lecture Notes in Comput. Sci. Engrg. 5, Springer-Verlag, Berlin, 1999, pp. 123–194.

[25] E. Tadmor, *Local error estimates for discontinuous solutions of nonlinear hyperbolic equations*, SIAM J. Numer. Anal., 28 (1991), pp. 891–906.

SIAM J. MATH. ANAL.
Vol. 35, No. 6, pp. 1371–1393

DETERMINATION OF TWO CONVECTION COEFFICIENTS FROM DIRICHLET TO NEUMANN MAP IN THE TWO-DIMENSIONAL CASE*

JIN CHENG† AND MASAHIRO YAMAMOTO‡

Abstract. In the two-dimensional case, we consider the problem of determining two convection coefficients from the Dirichlet to Neumann map. With the theory of generalized analytic functions which was developed by Bers and Vekua, we can formulate the problem as an inverse problem for a first order elliptic system. By using the inverse scattering method for the first order elliptic system, we prove that, in two dimensions, the Dirichlet to Neumann map uniquely determines two convection coefficients without any smallness assumption of unknown coefficients.

Key words. Dirichlet to Neumann map, convection coefficients, global uniqueness, inverse scattering method, first order elliptic system

AMS subject classifications. 35R30, 35J45

DOI. 10.1137/S0036141003422497

1. Introduction.

In recent years, the problems of determining coefficients in elliptic equations by using all possible boundary measurements, i.e., the Dirichlet to Neumann map, have been studied extensively. An important reason for this current interest is that these kinds of formulations are reasonable for detection or identification problems in many applied fields such as geophysics, medicine, biology, etc.

Since the mathematical problems for electric impedance tomography was first proposed and studied by Calderón in [7], gratifying progress has been made in this field [8], [10], [13], [14], [15], [17], [22], [23], [24]. In particular, in [22] it was first shown that, when the dimension is greater than two, the conductivity can be uniquely determined by all voltage and current flux at the boundary. However, the problem is more difficult in the two-dimensional case. Nachman has proved that the global uniqueness theorem is true also in two dimensions [16]. Here, by the global uniqueness we mean the uniqueness in determining coefficients without any smallness assumptions on coefficients. For the global uniqueness for the inverse conductivity problem, Nachman used an inverse scattering theory for a first order elliptic system. Recently, Brown and Uhlmann [5] improved Nachman's result to nonsmooth conductivity coefficients by treating the elliptic equation as a first order elliptic system in the complex plane. For other results in two dimensions, we refer the reader to [12], [18]. It should be noted here that all the mentioned papers treat determination of a single function from the Dirichlet to Neumann map. For determining multiple coefficients from the Dirichlet to Neumann map, there are only partial answers in the two-

*Received by the editors February 8, 2003; accepted for publication (in revised form) August 1, 2003; published electronically February 18, 2004.
http://www.siam.org/journals/sima/35-6/42249.html

†Department of Mathematics, Fudan University, Shanghai 200433, China and Division of Computational Science, E-Institute of Shanghai Universities, Shanghai, China (jcheng@fudan.edu.cn). This author was supported in part by the NSF of China (10271032), the Nonlinear Mathematical Models and Methods Laboratory at Fudan University, and the E-Institutes of Shanghai Municipal Education Commission (N. E03004).

‡Department of Mathematical Sciences, University of Tokyo, Komaba, Meguro, Tokyo 153, Japan (myama@ms.u-tokyo.ac.jp). This author was supported in part by grants 15340027 and 15654015 from the Japan Society for the Promotion of Science and Sanwa Systems Development Co., Ltd. (Tokyo, Japan).

dimensional case (see [11]). In the three-dimensional case, some global results are obtained by Nakamura, Sun, and Uhlmann [17]. The global uniqueness of multiple coefficients in the two-dimensional case is open. The purpose of this paper is to answer to a global uniqueness problem in determining two coefficients in the two-dimensional case.

Since we want to determine two functions from boundary measurements, we cannot transform our equation into the Schrödinger equation with one potential. Therefore Nachman's approach to the inverse conductivity problem cannot work well. Our method is similar to Brown and Uhlmann's paper [5] in the following ways:

(i) We treat second order elliptic equations in terms of first order elliptic systems in the complex plane and apply the theory of generalized analytic functions (Vekua [25] and Bers [4]).

(ii) We apply the inverse scattering method for first order systems which was developed by Beals and Coifman [3].

In [19], [20], [21] Sung gives more detailed treatments for that method. In this paper, we will mainly use Sung's notation and formulations. Moreover, a Carleman estimate by Bukhgeim [6] is used for our argument.

We note that in spite of the similarity to [5] our method allows us to prove that, in the two-dimensional case, the Dirichlet to Neumann map can uniquely determine two unknown functions of a convection term. Moreover, we can show that at most two coefficients are uniquely determined in two dimensions. Our proof is constructive in a sense similar to [16].

Our paper is organized as follows:

- Section 2: Formulation of the inverse problem and the main result.
- Section 3: Some lemmata.
- Section 4: Proof of the main result.
- Section 5: Concluding remarks.

2. Formulation of the inverse problem and the main result. Throughout this paper, we identify $\mathcal{R}^2$ with the complex plane $\mathcal{C}$ and let $W^{s,p}(\partial\Omega)$ be the Sobolev space ([1, pp. 214–217]).

Suppose that Ω is a simply connected bounded domain in $\mathcal{R}^2$ with Lipschitz boundary $\partial\Omega$. We consider the following elliptic equation in Ω:

$$\Delta u(x) + b_1(x)\frac{\partial u}{\partial x_1}(x) + b_2(x)\frac{\partial u}{\partial x_2}(x) = 0, \qquad x \in \Omega, \tag{2.1}$$

where $x = (x_1, x_2)$ and $b(x) = (b_1(x), b_2(x)) \in L^p(\Omega) \times L^p(\Omega)$ $(p > 2)$, b_j, $j = 1, 2$, are real-valued. Henceforth $\overline{\Omega}$ denotes the closure of Ω.

By [25], we know that (2.1) with Dirichlet boundary condition

$$u(x) = f(x), \qquad x \in \partial\Omega, \tag{2.2}$$

is uniquely solvable in $W^{2,p}(\Omega)$ for all real-valued $f \in W^{2-\frac{1}{p},p}(\partial\Omega)$.

In terms of this solution, we can define the Dirichlet to Neumann map $\Lambda_b : W^{2-\frac{1}{p},p}(\partial\Omega) \longrightarrow W^{1-\frac{1}{p},p}(\partial\Omega)$ by

$$\Lambda_b f = \frac{\partial u}{\partial \nu}.$$

Our inverse problem is to determine $b(x) = (b_1(x), b_2(x))$ from Λ_b.

Remark 2.1. Let $b(x) = \nabla \ln \gamma(x)$, $x \in \Omega$, where $\gamma > 0$ on $\overline{\Omega}$. Then our problem is the inverse conductivity problem which has been studied more extensively (see [5], [16] for the two-dimensional case).

We establish our main result, the following global uniqueness theorem for our inverse problem.

THEOREM 2.1. *Suppose that* $p > 2$ *and* $b^{(j)} = (b_1^{(j)}, b_2^{(j)}) \in L^p(\Omega) \times L^p(\Omega)$, $j = 1, 2$, *are real-valued. If*

$$\Lambda_{b^{(1)}} = \Lambda_{b^{(2)}}, \tag{2.3}$$

then we have

$$b_1^{(1)}(x) = b_1^{(2)}(x), \quad b_2^{(1)}(x) = b_2^{(2)}(x), \qquad x \in \Omega. \tag{2.4}$$

Remark 2.2. In [17], it is shown that the Dirichlet to Neumann map is invariant under a gauge transform so that the difficulty of unique determination of multiple coefficients is mentioned. However, in our case we assume that (2.1) has no zeroth order term in u, which cancels such nonuniqueness.

3. Some lemmata.

3.1. Transform of (2.1) to a first order elliptic equation in the complex plane $\mathcal{C}$. Henceforth we identify $x = (x_1, x_2) \in \mathcal{R}^2$ with the complex variable $z = x_1 + ix_2$, and we set $\bar{z} = x_1 - ix_2$, $\Re z = x_1$, and $\Im z = x_2$.

Let $w(z) = \partial_z u(z) = \frac{1}{2}(\frac{\partial}{\partial x_1} - i\frac{\partial}{\partial x_2})u(z)$.

From (2.1), we have that $w(z)$ satisfies

$$\partial_{\bar{z}} w(z) + \frac{1}{4}(b_1 + ib_2)(z)w(z) + \frac{1}{4}(b_1 - ib_2)(z)\overline{w(z)} = 0, \quad z \in \Omega, \tag{3.1}$$

where $\partial_{\bar{z}} = \frac{1}{2}(\frac{\partial}{\partial x_1} + i\frac{\partial}{\partial x_2})$.

Let $B(z) = \frac{1}{4}(b_1(z) + ib_2(z))$. Then (3.1) can be written as

$$\partial_{\bar{z}} w(z) + B(z)w(z) + \overline{B(z)w(z)} = 0, \quad z \in \Omega. \tag{3.2}$$

LEMMA 3.1. *If* $u \in W^{2,p}(\Omega)$ *satisfies* (2.1), *then* $w = \partial_z u \in W^{1,p}(\Omega)$ *satisfies* (3.2). *Conversely, if* $w \in W^{1,p}(\Omega)$ *satisfies* (3.2), *then there exists* $u \in W^{2,p}(\Omega)$ *(not necessarily unique) such that*

$$\Delta u + b_1 \frac{\partial u}{\partial x_1} + b_2 \frac{\partial u}{\partial x_2} = 0 \qquad in \quad \Omega$$

and

$$w = \partial_z u \qquad in \quad \Omega.$$

Proof. We need only to prove that, from the solution $w = \Re w + i\Im w$ of (3.2), we can obtain a solution u of (2.1).

From the imaginary part of (3.2), we obtain

$$\frac{\partial \Im w}{\partial x_1} + \frac{\partial \Re w}{\partial x_2} = 0.$$

Since Ω is a simply connected domain in $\mathcal{C}$, there exists a real function u of class $W^{1,p}(\Omega)$ such that

$$w = \partial_z u.$$

By (3.2), we can directly verify that u satisfies

$$\Delta u(x) + b_1(x)\frac{\partial u}{\partial x_1}(x) + b_2(x)\frac{\partial u}{\partial x_2}(x) = 0, \qquad x \in \Omega.$$

The proof is complete. □

Let us define an operator T by

$$(Tv)(z) = -\frac{1}{\pi}\int_\Omega \frac{v(\zeta)}{\zeta - z}d\zeta, \qquad z \in \Omega.$$

Then the following is known (see, e.g., Vekua [25]):

(i) $T : L^p(\Omega) \longrightarrow C^\beta(\overline{\Omega})$ is a bounded linear operator, where $p > 2$ and $\beta = \frac{p-2}{p}$.

(ii) $\lim_{|z|\to\infty}(Tv)(z) = 0$ if $v \in L^p(\Omega)$, $p > 2$.

Moreover, we have the following lemma.

LEMMA 3.2. *If $w(z)$ satisfies (3.2), then*

$$w(z) + T(Bw)(z) + T(\overline{Bw})(z) = \Phi(z), \quad z \in \Omega, \tag{3.3}$$

where $\Phi = \Phi(z)$ is a holomorphic function in Ω and $\Phi(z) = \frac{1}{2\pi i}\int_{\partial\Omega}\frac{w(\zeta)}{\zeta - z}d\zeta$, $z \in \Omega$. In particular, Φ depends only on boundary values of w.

Conversely, for any given holomorphic function Φ, if w satisfies (3.3), then w satisfies (3.2) and $\Phi(z) = \frac{1}{2\pi i}\int_{\partial\Omega}\frac{w(\zeta)}{\zeta - z}d\zeta$, $z \in \Omega$.

Since the fundamental solution of $\partial_{\bar z}$ is $\frac{1}{\pi z}$ (see, e.g., [2]), this lemma can be proved by the Green formula. See also Chapter I in [25].

We conclude this subsection with the following lemma.

LEMMA 3.3. *The integral equation*

$$w(z) + T(Bw)(z) + T(e^{-\frac{i}{2}(\bar k z + k\bar z)}\overline{Bw})(z) = \Phi(z), \quad z \in \Omega,$$

possesses a unique solution $w \in W^{1,p}(\Omega) \cap C(\overline{\Omega})$ for any given holomorphic function $\Phi \in C(\overline{\Omega})$. Here k is a complex number.

Proof. Since the operator $(\mathcal{P}w)(z) = T(Bw)(z) + T(e^{-\frac{i}{2}(\bar k z + k\bar z)}\overline{Bw})(z)$ is compact from $C(\overline{\Omega})$ to $C(\overline{\Omega})$ (see [25]), it is sufficient to prove that there is only a trivial solution for the homogeneous equation.

Assume that $w_0(z)$ is the solution of

$$w_0 + \mathcal{P}w_0 = 0 \qquad \text{in} \quad \Omega.$$

We take the 0-extension of B outside of Ω. Then we extend $w_0(z)$ to the whole complex plane by defining $w_0(z) = (\mathcal{P}w_0)(z)$, $z \in \Omega^c$. Then w_0 satisfies

$$\partial_{\bar z} w_0 + Bw_0 + e^{-\frac{i}{2}(\bar k z + k\bar z)}\overline{Bw_0} = 0 \qquad \text{in} \quad \mathcal{C}$$

and

$$\lim_{|z|\to\infty} w_0(z) = 0.$$

By the Liouville theorem for generalized analytic functions (p. 154 in [25]), we have $w_0(z) = 0$, $z \in \mathcal{C}$. The proof is complete. □

3.2. Introduction of a complex parameter k. Let $w = w(z)$ satisfy (3.2). With a complex parameter k, we set

$$\alpha(z,k) = w(z)e^{-\frac{1}{2}i\bar{k}z}.$$

Then, noticing that $\partial_{\bar{z}} g = 0$ for a holomorphic function g, we see that $\alpha(z,k)$ satisfies

$$\partial_{\bar{z}}\alpha(z,k) + B(z)\alpha(z,k) + e^{-\frac{1}{2}i(\bar{k}z+k\bar{z})}\overline{B(z)\alpha(z,k)} = 0, \quad z \in \Omega. \tag{3.4}$$

We set

$$e_k(z) = \exp\left(-\frac{i}{2}(\bar{k}z + k\bar{z})\right).$$

We note that $|e_k(z)| = 1$ for all $k, z \in \mathcal{C}$.

LEMMA 3.4. *Let $v \in L^p(\Omega)$ be complex-valued with $p > 2$. Then*

$$\lim_{|k|\to\infty} \max_{z\in\overline{\Omega}} |(Te_k v)(z)| = 0.$$

Proof. Contrarily, assume that the conclusion is not true. Then there exists a constant $\varepsilon_0 > 0$ and two sequences $\{k_n\}_{n=1}^{\infty} \subset \mathcal{C}$, $\{z_n\}_{n=1}^{\infty} \subset \overline{\Omega}$ such that

$$\lim_{n\to\infty} |k_n| = \infty$$

and

$$|(Te_{k_n} v)(z_n)| > \varepsilon_0, \quad n = 1, 2, 3, \ldots. \tag{3.5}$$

Since $\overline{\Omega}$ is compact, there exists $z_0 \in \overline{\Omega}$ such that $\lim_{n\to\infty} z_n = z_0$, by choosing a subsequence if necessary.

By Theorem I-1.19 (p. 38) in [25], we have

$$\begin{aligned} &|(Te_{k_n} v)(z_n) - (Te_{k_n} v)(z_0)| \\ &\le M_1 \|v\|_{L^p(\Omega)} |z_n - z_0|^{\beta}, \quad k \in \mathcal{C}, \end{aligned}$$

where $M_1 > 0$ is independent of k.

Therefore there exists $N_1 \in \mathcal{N}$ such that

$$|(Te_{k_n} v)(z_n) - (Te_{k_n} v)(z_0)| \le \frac{\varepsilon_0}{4}$$

for $n \ge N_1$.

Hence it follows from (3.5) that

$$|(Te_{k_n} v)(z_0)| > \frac{3\varepsilon_0}{4}, \quad n > N_1 + 1. \tag{3.6}$$

We set $\zeta = \xi_1 + i\xi_2 = (\xi_1, \xi_2) \in \mathcal{R}^2$, $k = k_1 + ik_2 = (k_1, k_2) \in \mathcal{R}^2$, and $(k, \zeta) = k_1\xi_1 + k_2\xi_2$.

Then $e^{-\frac{i}{2}(\bar{k}\zeta + k\bar{\zeta})} = e^{-i(k,\zeta)}$ and

$$(Te_k v)(z_0) = -\frac{1}{\pi}\int_{\Omega} e^{-i(k,\zeta)} \frac{v(\zeta)}{\zeta - z_0} d\zeta.$$

Since $h(\zeta) = \frac{v(\zeta)}{\zeta - z_0} \in L^1(\Omega)$ by $v \in L^p(\Omega)$, $p > 2$, and the Hölder inequality, we can apply the Riemann–Lebesgue lemma. Therefore there exists $N_2 \in \mathcal{N}$ such that

$$|(Te_k v)(z_0)| < \frac{\varepsilon_0}{4}$$

for $|k| > N_2$.

This contradicts (3.6). Thus the proof of the lemma is complete. □

By Lemma 3.2, equation (3.4) is equivalent to the following integral equation:

$$\alpha(z,k) + T(B\alpha)(z) + T(e_k\overline{B\alpha})(z) = \Phi(z), \quad z \in \Omega, \tag{3.7}$$

where Φ is a holomorphic function with respect to z in Ω.

Henceforth we consider a solution of (3.7) for $\Phi(z) = 1$, and for simplicity we denote this solution by the same notation $\alpha(z,k)$:

$$\alpha(z,k) + T(B\alpha)(z) + T(e_k\overline{B\alpha})(z) = 1, \quad z \in \Omega. \tag{3.8}$$

By Lemma 3.3, we note that there exists a unique solution to (3.8) in $C(\overline{\Omega})$.

Then we have the following result about the asymptotic property of $\alpha(z,k)$ as $|k| \to \infty$.

LEMMA 3.5. *For $z \in \overline{\Omega}$, we have*

$$\alpha(z,k) \to \alpha_0(z) \qquad as \quad |k| \to \infty,$$

where $\alpha_0(z)$ is the solution of

$$\alpha_0(z) + T(B\alpha_0)(z) = 1, \quad z \in \overline{\Omega}. \tag{3.9}$$

Proof. By Lemma 3.3, we know that

$$\alpha_0 \in L^\infty(\Omega).$$

From (3.8) and (3.9), we see that $\rho(z,k) = \alpha(z,k) - \alpha_0(z)$ satisfies

$$\rho + T(B\rho) = -T(e_k\overline{B\rho}) - T(e_k\overline{B\alpha_0}) \qquad \text{in} \quad \Omega. \tag{3.10}$$

We introduce a new function $\lambda = \lambda(z,k)$:

$$\lambda(z,k) = \rho(z,k) + T(B\rho)(z), \quad z \in \Omega. \tag{3.11}$$

If we regard (3.11) as an integral equation with respect to ρ, then this integral equation is uniquely solvable (see, e.g., [25] or Lemma 3.3) because $\lambda, \rho \in L^\infty(\Omega)$ and the solution can be expressed as

$$\rho(z,k) = \lambda(z,k) + \int_\Omega L(\zeta,z)\lambda(\zeta,k)d\zeta \equiv \lambda(z,k) + L\lambda(z,k), \quad z \in \Omega. \tag{3.12}$$

Moreover,

$$|L(\zeta,z)| \le M_2 \frac{1}{|\zeta - z|}, \qquad z, \zeta \in \Omega.$$

Here $M_2 > 0$ depends only on Ω and B. In particular, L is a bounded linear operator from $L^\infty(\Omega)$ to itself.

Therefore (3.10) can be written as

$$\lambda(z,k)+T(e_k\overline{B(\lambda+L\lambda)})(z)=-T(e_k\overline{B\alpha_0})(z),\quad z\in\Omega. \tag{3.13}$$

From (3.13), we have

$$\begin{aligned}\lambda &= -T\left(e_k\overline{B(\lambda+L\lambda)}\right)-T(e_k\overline{B\alpha_0})\\ &= T\left(e_k\overline{B}\left[\overline{T(e_k\overline{B(\lambda+L\lambda)})+T(e_k\overline{B\alpha_0})}\right]\right)\\ &\quad+T\Big(e_k\overline{BL}\left[\overline{T(e_k\overline{B(\lambda+L\lambda)})+T(e_k\overline{B\alpha_0})}\right]\Big)-T(e_k\overline{B\alpha_0})\\ &= T\left(e_k\overline{B}\left[\overline{T(e_k\overline{B\lambda})}\right]\right)+T\left(e_k\overline{B}\left[\overline{T(e_k\overline{BL\lambda})}\right]\right)\\ &\quad+T\left(e_k\overline{BL}\left[\overline{T(e_k\overline{B\lambda})}\right]\right)+T\left(e_k\overline{BL}\left[\overline{T(e_k\overline{B\overline{L}\lambda})}\right]\right)+\eta_k(z)\\ &\equiv A_1^k\lambda+A_2^k\lambda+A_3^k\lambda+A_4^k\lambda+\eta_k.\end{aligned} \tag{3.14}$$

Here we note that

$$\eta_k(z)=T\left(e_k\overline{BT(e_k\overline{B\alpha_0})}\right)(z)+T\left(e_k\overline{BLT(e_k\overline{B\alpha_0})}\right)(z)-T(e_k\overline{B\alpha_0})(z).$$

Next, we will show that, for sufficiently large $|k|$, A_j^k $(j=1,2,3,4)$ are contraction operators on $L^\infty(\Omega)$.

(i) For the operator A_1^k,

$$(A_1^k\lambda)(z)=T\left(e_k\overline{B}\left[\overline{T(e_k\overline{B\lambda})}\right]\right)(z)=\int_\Omega A_1^k(\zeta,z)\lambda(\zeta)d\zeta,\quad z\in\Omega,$$

where

$$A_1^k(\zeta,z)=\frac{1}{\pi^2}\int_\Omega\frac{e^{-\frac{1}{2}i(\bar k\zeta_1+k\bar\zeta_1)}\overline{B(\zeta_1)}}{\zeta_1-z}\frac{e^{\frac{1}{2}i(\bar k\zeta+k\bar\zeta)}B(\zeta)}{\bar\zeta-\bar\zeta_1}d\zeta_1.$$

Here and henceforth we denote the kernels of the integral operators A_j^k by the same letters A_j^k $(j=1,2,3,4)$. Therefore we have

$$\begin{aligned}A_1^k(\zeta,z) &= -\frac{e^{\frac{1}{2}i(\bar k\zeta+k\bar\zeta)}B(\zeta)}{\pi^2(\zeta-z)}\int_\Omega e^{-\frac{1}{2}i(\bar k\zeta_1+k\bar\zeta_1)}\overline{B(\zeta_1)}\frac{\zeta_1-\zeta}{\bar\zeta_1-\bar\zeta}\left[\frac{1}{\zeta_1-\zeta}-\frac{1}{\zeta_1-z}\right]d\zeta_1\\ &= -\frac{1}{\pi^2}\frac{e^{\frac{1}{2}i(\bar k\zeta+k\bar\zeta)}B(\zeta)}{\zeta-z}\int_\Omega e^{-\frac{1}{2}i(\bar k\zeta_1+k\bar\zeta_1)}\overline{B(\zeta_1)}\frac{1}{\bar\zeta_1-\bar\zeta}d\zeta_1\\ &\quad+\frac{1}{\pi^2}\frac{e^{\frac{1}{2}i(\bar k\zeta+k\bar\zeta)}B(\zeta)}{\zeta-z}\int_\Omega e^{-\frac{1}{2}i(\bar k\zeta_1+k\bar\zeta_1)}\overline{B(\zeta_1)}\frac{\zeta_1-\zeta}{\bar\zeta_1-\bar\zeta}\frac{1}{\zeta_1-z}d\zeta_1\\ &= \frac{1}{\pi}\frac{e^{\frac{i}{2}(\bar k\zeta+k\bar\zeta)}B(\zeta)}{\zeta-z}(\overline{Te_kB})(\zeta)\\ &\quad-\frac{1}{\pi}\frac{e^{\frac{i}{2}(\bar k\zeta+k\bar\zeta)}B(\zeta)}{\zeta-z}\left(Te_k(\zeta_1)\overline{B(\zeta_1)}\frac{\zeta_1-\zeta}{\bar\zeta_1-\bar\zeta}\right)(z).\end{aligned}$$

At the last equality we have used

$$\overline{\int_\Omega v(\zeta)d\zeta}=\int_\Omega\overline{v(\zeta)}d\zeta.$$

By Lemma 3.4, we have

$$\sigma_k \equiv \max_{\zeta\in\overline{\Omega}} \left|\overline{(Te_kB)(\zeta)}\right| \to 0 \qquad \text{as} \quad |k|\to\infty \tag{3.15}$$

and

$$\widehat{\sigma}_k \equiv \max_{z,\zeta\in\overline{\Omega}} \left| \int_\Omega e^{-\frac{1}{2}i(\bar{k}\zeta_1+k\bar{\zeta}_1)} \overline{B(\zeta_1)} \frac{\zeta_1-\zeta}{\bar{\zeta}_1-\bar{\zeta}} \frac{1}{\zeta_1-z} d\zeta_1 \right| \to 0 \quad \text{as} \quad |k|\to\infty. \tag{3.16}$$

In fact, (3.15) is straightforward from Lemma 3.4. For (3.16), we assume contrarily that there exist $\varepsilon_0 > 0$, $\{k_n\}_{n=1}^\infty \subset \mathcal{C}$, $\{z_n\}_{n=1}^\infty \subset \overline{\Omega}$, $\{\zeta_n\}_{n=1}^\infty \subset \overline{\Omega}$ such that $\lim_{n\to\infty}|k_n| = \infty$ and

$$\left| \left(Te_{k_n} B \frac{\zeta_1-\zeta_n}{\bar{\zeta}_1-\bar{\zeta}_n} \right)(z_n) \right| > \varepsilon_0, \quad n = 2,3,\dots. \tag{3.17}$$

We can choose subsequences $\{z_n\}_{n=1}^\infty$, $\{\zeta_n\}_{n=1}^\infty$ such that $\lim_{n\to\infty} z_n = z_0$ and $\lim_{n\to\infty}\zeta_n = \zeta_0$ by the compactness of $\overline{\Omega}$. Then we have

$$\begin{aligned}
&\left| \left(Te_{k_n} B \frac{\zeta_1-\zeta_n}{\bar{\zeta}_1-\bar{\zeta}_n} \right)(z_n) - \left(Te_{k_n} B \frac{\zeta_1-\zeta_0}{\bar{\zeta}_1-\bar{\zeta}_0} \right)(z_0) \right| \\
&\le \left| \left(Te_{k_n} B \frac{\zeta_1-\zeta_n}{\bar{\zeta}_1-\bar{\zeta}_n} \right)(z_n) - \left(Te_{k_n} B \frac{\zeta_1-\zeta_n}{\bar{\zeta}_1-\bar{\zeta}_n} \right)(z_0) \right| \\
&+ \left| \left(Te_{k_n} B \left(\frac{\zeta_1-\zeta_n}{\bar{\zeta}_1-\bar{\zeta}_n} - \frac{\zeta_1-\zeta_0}{\bar{\zeta}_1-\bar{\zeta}_0} \right) \right)(z_0) \right| \\
&\le M_3 \left\| e_{k_n} B \frac{\zeta_1-\zeta_n}{\bar{\zeta}_1-\bar{\zeta}_n} \right\|_{L^p_{\zeta_1}(\Omega)} |z_n - z_0|^\beta \\
&+ M_3 \left\| e_{k_n} B \left(\frac{\zeta_1-\zeta_n}{\bar{\zeta}_1-\bar{\zeta}_n} - \frac{\zeta_1-\zeta_0}{\bar{\zeta}_1-\bar{\zeta}_0} \right) \right\|_{L^p_{\zeta_1}(\Omega)}
\end{aligned}$$

by Theorem I-1.19 (p. 38) in [25] and $TL^p(\Omega) \subset C^\beta(\overline{\Omega})$.

Since

$$\lim_{n\to\infty} \left\| \frac{\zeta_1-\zeta_n}{\bar{\zeta}_1-\bar{\zeta}_n} - \frac{\zeta_1-\zeta_0}{\bar{\zeta}_1-\bar{\zeta}_0} \right\|_{L^p_{\zeta_1}(\Omega)} = 0$$

by the Lebesgue convergence theorem, we see that

$$\begin{aligned}
&\left| \left(Te_{k_n} B \frac{\zeta_1-\zeta_n}{\bar{\zeta}_1-\bar{\zeta}_n} \right)(z_n) - \left(Te_{k_n} B \frac{\zeta_1-\zeta_0}{\bar{\zeta}_1-\bar{\zeta}_0} \right)(z_0) \right| \\
&\le \frac{\varepsilon_0}{4}
\end{aligned}$$

for $n > N_1$ which is some large natural number.

Therefore (3.17) implies

$$\left| \left(Te_{k_n} B \frac{\zeta_1-\zeta_0}{\bar{\zeta}_1-\bar{\zeta}_0} \right)(z_0) \right| > \frac{3\varepsilon_0}{4}$$

if $n > N_1$. This is a contradiction by Lemma 3.4. Thus the proof of (3.16) is complete.

Therefore, noticing that $\max_{z\in\mathcal{C}}\int_\Omega \frac{1}{|\zeta-z|}d\zeta<\infty$, we can obtain

$$\begin{aligned}\|A_1^k\lambda\|_{L^\infty(\Omega)} &\le \sup_{z\in\Omega}\int_\Omega |A_1^k(\zeta,z)|d\zeta\|\lambda\|_{L^\infty(\Omega)}\\ &\le \frac{M_4}{\pi^2}\|B\|_{L^p(\Omega)}(\sigma_k+\widehat{\sigma}_k)\|\lambda\|_{L^\infty(\Omega)}.\end{aligned}\tag{3.18}$$

(ii) For the operator A_2^k,

$$(A_2^k\lambda)(z)=T\left(e_k\overline{B}\left[\overline{T(e_k\overline{BL\lambda})}\right]\right)(z)=(A_1^kL\lambda)(z),\quad z\in\Omega.$$

Therefore by (3.18) we see

$$\|A_2^k\lambda\|_{L^\infty(\Omega)}\le\frac{M_4}{\pi^2}\|B\|_{L^p(\Omega)}(\sigma_k+\widehat{\sigma}_k)\|L\|_{L^\infty(\Omega)}\|\lambda\|_{L^\infty(\Omega)}.$$

Here we recall that L is bounded from $L^\infty(\Omega)$ to $L^\infty(\Omega)$.

For the operators A_3^k and A_4^k, we can similarly prove that

$$\|A_j^k\|_{L^\infty}\to 0,\qquad |k|\to\infty,\quad j=3,4.$$

Therefore there exists a positive constant R such that, for $|k|>R$,

$$\|A_j^k\|_{L^\infty}\le\frac{1}{8},\qquad j=1,2,3,4.$$

From (3.14), we can have that, for $|k|>R$,

$$\|\lambda\|_{L^\infty(\Omega)}\le\frac{1}{2}\|\lambda\|_{L^\infty(\Omega)}+\|\eta_k\|_{L^\infty(\Omega)}.$$

Therefore for $|k|>R$ we can obtain

$$\|\lambda\|_{L^\infty(\Omega)}\le 2\|\eta_k\|_{L^\infty(\Omega)}.$$

From Lemma 3.4, we can verify that

$$\lim_{|k|\to\infty}\|\eta_k\|_{L^\infty(\Omega)}=0.$$

Consequently,

$$\lim_{|k|\to\infty}\|\lambda(\cdot,k)\|_{L^\infty(\Omega)}=0.$$

Therefore by (3.12) we can obtain

$$\lim_{|k|\to\infty}\|\rho(\cdot,k)\|_{L^\infty(\Omega)}=0.$$

The proof is complete. □

We conclude this subsection with the following lemma.

LEMMA 3.6. *The solution of* (3.9) *can be written as*

$$\alpha_0(z)=\Phi_0(z)e^{-(TB)(z)},\qquad z\in\Omega,\tag{3.19}$$

where Φ_0 *is a holomorphic function in* Ω *and is continuous on* $\overline{\Omega}$.

Moreover,

$$\Phi_0(z) \neq 0, \qquad z \in \Omega. \tag{3.20}$$

Proof. It is easy to verify that α_0 satisfies the following first order elliptic complex equation:

$$\partial_{\bar{z}}\alpha_0 + B\alpha_0 = 0 \qquad \text{in} \quad \Omega. \tag{3.21}$$

Since $\partial_{\bar{z}}(TB) = B$ (see, e.g., [25]), we have $\partial_{\bar{z}}(\alpha_0 e^{TB}) = (\partial_{\bar{z}}\alpha_0)e^{TB} + \alpha_0(\partial_{\bar{z}}e^{TB}) = -B\alpha_0 e^{TB} + \alpha_0 B e^{TB} = 0$ in Ω by (3.21). Therefore $\Phi_0 \equiv \alpha_0 e^{TB}$ is holomorphic in Ω. Hence (3.19) follows.

Next, we have to prove (3.20). If there exists $z_0 \in \Omega$ such that $\Phi_0(z_0) = 0$, then we can see that Φ_0 can be written as

$$\Phi_0(z) = (z - z_0)\Phi_1(z), \qquad z \in \Omega, \tag{3.22}$$

where Φ_1 is a holomorphic function in Ω.

Substituting (3.22) and (3.19) in (3.9), we have

$$(z - z_0)\Phi_1(z)e^{-(TB)(z)} - \frac{1}{\pi}\int_\Omega \frac{B(\zeta)(\zeta - z_0)\Phi_1(\zeta)e^{-(TB)(\zeta)}}{\zeta - z}d\zeta = 1. \tag{3.23}$$

Therefore, letting $z \to z_0$, we can obtain

$$-\frac{1}{\pi}\int_\Omega B(\zeta)\Phi_1(\zeta)e^{-(TB)(\zeta)}d\zeta = 1. \tag{3.24}$$

By (3.23), we have

$$\begin{aligned} 1 &= (z - z_0)\Phi_1(z)e^{-(TB)(z)} \\ &\quad -\frac{1}{\pi}\int_\Omega \frac{B(\zeta)(\zeta - z + z - z_0)\Phi_1(\zeta)e^{-(TB)(z)}}{\zeta - z}d\zeta \\ &= (z - z_0)\left(\Phi_1(z)e^{-(TB)(z)} - \frac{1}{\pi}\int_\Omega \frac{B(\zeta)\Phi_1(\zeta)e^{-(TB)(\zeta)}}{\zeta - z}d\zeta\right) \\ &\quad -\frac{1}{\pi}\int_\Omega B(\zeta)\Phi_1(\zeta)e^{-(TB)(\zeta)}d\zeta. \end{aligned}$$

Therefore (3.24) implies

$$(z - z_0)(\Phi_1(z)e^{-(TB)(z)} - \frac{1}{\pi}\int_\Omega \frac{B(\zeta)\Phi_1(\zeta)e^{-(TB)(\zeta)}}{\zeta - z}d\zeta) = 0, \quad z \in \Omega;$$

that is,

$$\Phi_1 e^{-TB} + T(B\Phi_1 e^{-TB}) = 0 \qquad \text{in} \quad \Omega.$$

By Lemma 3.3, we can see that $\Phi_1 = 0$ in Ω. Therefore $\alpha_0 = 0$. This contradicts (3.9). The proof is complete. □

4. Proof of the main result. For $j = 1, 2$, let us consider

$$\Delta u(x) + b_1^{(j)}(x)\frac{\partial u}{\partial x_1}(x) + b_2^{(j)}(x)\frac{\partial u}{\partial x_2}(x) = 0, \qquad x \in \Omega,$$

and

$$u(x) = f(x), \qquad x \in \partial\Omega,$$

with $f \in W^{2-\frac{1}{p},p}(\partial\Omega)$, $(b_1^{(j)}, b_2^{(j)}) \in L^p(\Omega) \times L^p(\Omega)$, $p > 2$.

By [25], there exists a unique solution in $W^{2,p}(\Omega)$, and we denote it by $u^{(j)}(x, f)$, $j = 1, 2$.

We set $b^{(j)} = (b_1^{(j)}, b_2^{(j)}) \in L^p(\Omega) \times L^p(\Omega)$ and

$$B^{(j)}(z) = \frac{1}{4}(b_1^{(j)}(z) + ib_2^{(j)}(z)), \qquad z \in \Omega.$$

By Lemma 3.3, the integral equation

$$\alpha + T(B^{(j)}\alpha) + T(e_k\overline{B^{(j)}\alpha}) = 1 \qquad \text{in} \quad \Omega \tag{4.1}$$

is uniquely solvable in $L^\infty(\Omega)$, and henceforth we denote the unique solution to (4.1) by $\alpha^{(j)}(z, k)$, $j = 1, 2$, $k \in \mathcal{C}$.

We state a relation between $\alpha^{(1)}$ and $\alpha^{(2)}$, provided that $\Lambda_{b^{(1)}} = \Lambda_{b^{(2)}}$.

LEMMA 4.1. *Let $\Lambda_{b^{(1)}} = \Lambda_{b^{(2)}}$. Then*

$$\alpha^{(1)}(z, k) = \alpha^{(2)}(z, k), \qquad z \in \partial\Omega, \quad k \in \mathcal{C}.$$

Proof. First it follows from Lemma 3.2 that

$$\frac{1}{2\pi i}\int_{\partial\Omega}\frac{\alpha^{(1)}(\zeta, k)d\zeta}{\zeta - z} = 1, \qquad z \in \Omega. \tag{4.2}$$

By Lemma 3.3, we see $\alpha^{(1)}(\cdot, k) \in W^{1,p}(\Omega)$, and so

$$w^{(1)}(z, k) \equiv \alpha^{(1)}(z, k)e^{\frac{1}{2}i\bar{k}z} \in W^{1,p}(\Omega) \tag{4.3}$$

and

$$\partial_{\bar{z}}w^{(1)} + B^{(1)}w^{(1)} + \overline{B^{(1)}w^{(1)}} = 0 \qquad \text{in} \quad \Omega.$$

Hence Lemma 3.1 implies the existence of $u^{(1)} \in W^{2,p}(\Omega)$ such that

$$w^{(1)} = \partial_z u^{(1)} \qquad \text{in} \quad \Omega$$

and

$$\Delta u^{(1)}(x) + b_1^{(1)}(x)\frac{\partial u^{(1)}}{\partial x_1}(x) + b_2^{(1)}(x)\frac{\partial u^{(1)}}{\partial x_2}(x) = 0, \qquad x \in \Omega.$$

We set

$$f = u^{(1)}|_{\partial\Omega}.$$

Then we can see that $f \in W^{2-\frac{1}{p},p}(\partial\Omega)$.

By Vekua [25], there exists a unique solution $u^{(2)} = u^{(2)}(\cdot, f) \in W^{2,p}(\Omega)$ to

$$\Delta u^{(2)}(x) + b_1^{(2)}(x)\frac{\partial u^{(2)}}{\partial x_1}(x) + b_2^{(2)}(x)\frac{\partial u^{(2)}}{\partial x_2}(x) = 0, \qquad x \in \Omega,$$

and

$$u^{(2)}|_{\partial\Omega} = f.$$

Since $\Lambda_{b^{(1)}} = \Lambda_{b^{(2)}}$, it follows that $\frac{\partial u^{(1)}}{\partial \nu} = \frac{\partial u^{(2)}}{\partial \nu}$ on $\partial\Omega$.

With $u^{(1)} = u^{(2)} = f$ on $\partial\Omega$, we obtain

$$\partial_z u^{(1)} = \partial_z u^{(2)} \qquad \text{on} \quad \partial\Omega.$$

We set

$$w^{(2)} = \partial_z u^{(2)} \qquad \text{in} \quad \Omega.$$

By Lemma 3.1, $w^{(2)}$ satisfies

$$\partial_{\bar z} w^{(2)} + B^{(2)} w^{(2)} + \overline{B^{(2)} w^{(2)}} = 0 \qquad \text{in} \quad \Omega$$

and

$$w^{(1)} = w^{(2)} \qquad \text{on} \quad \partial\Omega.$$

Setting

$$\widetilde{\alpha}^{(2)}(z,k) = w^{(2)}(z) e^{-\frac{1}{2}i\bar k z}, \qquad z \in \Omega, \quad k \in \mathcal{C},$$

from (4.3), we see that

$$\alpha^{(1)}(z,k) = \widetilde{\alpha}^{(2)}(z,k), \qquad z \in \partial\Omega, \quad k \in \mathcal{C}, \tag{4.4}$$

and

$$\partial_{\bar z}\widetilde{\alpha}^{(2)} + B^{(2)}\widetilde{\alpha}^{(2)} + e_k\overline{B^{(2)}\widetilde{\alpha}^{(2)}} = 0 \qquad \text{in} \quad \Omega.$$

Hence Lemma 3.2 yields

$$\widetilde{\alpha}^{(2)} + T(B^{(2)}\widetilde{\alpha}^{(2)}) + T(e_k\overline{B^{(2)}\widetilde{\alpha}^{(2)}}) = \Phi_2(\cdot, k) \qquad \text{in} \quad \Omega$$

with a holomorphic function Φ_2 in Ω given by

$$\Phi_2(z,k) = \frac{1}{2\pi i}\int_{\partial\Omega} \frac{\widetilde{\alpha}^{(2)}(\zeta,k)}{\zeta - z} d\zeta, \qquad z \in \Omega, \quad k \in \mathcal{C}.$$

In terms of (4.2) and (4.4), we have

$$\Phi_2(z,k) = 1, \qquad z \in \Omega, \quad k \in \mathcal{C},$$

and from (4.1) we conclude that $\alpha^{(2)}(z,k) = \widetilde{\alpha}^{(2)}(z,k)$, $z \in \Omega$, $k \in \mathcal{C}$. Thus the equality (4.4) completes the proof of the lemma. □

Henceforth we assume that

$$\Lambda_{b^{(1)}} = \Lambda_{b^{(2)}}.$$

Next, we will prove the following lemma.

LEMMA 4.2.

$$(TB^{(1)})(z) = (TB^{(2)})(z), \qquad z \in \mathcal{C} \setminus \Omega,$$

where $B^{(j)} = \frac{1}{4}(b_1^{(j)} + ib_2^{(j)})$, $j = 1, 2$.

Proof. By Lemma 3.5, we see that

$$\alpha^{(j)}(z, k) \to \alpha_0^{(j)}(z), \qquad z \in \overline{\Omega}, \quad |k| \to \infty,$$

where $\alpha_0^{(j)} = \alpha_0^{(j)}(z)$, $j = 1, 2$, satisfy

$$\alpha_0^{(j)} + T(B^{(j)}\alpha_0^{(j)}) = 1 \qquad \text{in} \quad \Omega.$$

Therefore from Lemma 4.1 we see that

$$\alpha_0^{(1)}(z) = \alpha_0^{(2)}(z), \qquad z \in \partial\Omega. \tag{4.5}$$

From Lemma 3.6, we obtain

$$\alpha_0^{(j)}(z) = \Phi_0^{(j)}(z)e^{-(TB^{(j)})(z)}, \qquad z \in \overline{\Omega},$$

where $\Phi_0^{(j)}$ is holomorphic in Ω and $\Phi_0^{(j)}(z) \neq 0$, $z \in \Omega$ for $j = 1, 2$.

By (4.5), we have

$$\Phi_0^{(1)}(z)e^{-(TB^{(1)})(z)} = \Phi_0^{(2)}(z)e^{-(TB^{(2)})(z)}, \qquad z \in \partial\Omega.$$

We define $\Psi = \Psi(z)$ by

$$\begin{aligned} &\Psi(z) = e^{(TB^{(1)} - TB^{(2)})(z)}, \qquad z \in \mathcal{C} \setminus \overline{\Omega}, \\ &\Psi(z) = \tfrac{\Phi_0^{(1)}(z)}{\Phi_0^{(2)}(z)}, \qquad z \in \Omega. \end{aligned} \tag{4.6}$$

By the form of T and $\Phi_0^{(2)}(z) \neq 0$ for all $z \in \Omega$, it is easy to verify that Ψ is holomorphic in $\mathcal{C} \setminus \partial\Omega$. Therefore by (4.5) we can apply the Painlevé theorem on analytic continuation, and we can see that Ψ is holomorphic in $\mathcal{C}$. Moreover, we can directly see that

$$(TB^{(1)})(z) - (TB^{(2)})(z) \longrightarrow 0 \qquad \text{as} \quad |z| \to \infty$$

and

$$\Psi(z) \to 1 \qquad \text{as} \quad |z| \to \infty.$$

Therefore by the Liouville theorem for holomorphic functions we have $\Psi(z) = 1, z \in \mathcal{C}$, and so (4.6) implies

$$(TB^{(1)} - TB^{(2)})(z) = 2\pi ni \qquad \text{for} \quad z \in \mathcal{C} \setminus \Omega,$$

where $n \in \mathcal{N}$ is independent of z. Since $\lim_{|z|\to\infty}(TB^{(1)} - TB^{(2)})(z) = 0$, we obtain $n = 0$. Thus the proof is complete. □

The next step is to apply the inverse scattering method by [19], [20], [21].

For $j = 1, 2$, we set

$$C^{(j)} = e^{TB^{(j)}} e^{-\overline{TB^{(j)}}} \overline{B^{(j)}} \quad \text{in} \quad \Omega, \tag{4.7}$$

$$Q^{(j)} = \begin{pmatrix} 0 & -C^{(j)} \\ -C^{(j)} & 0 \end{pmatrix} \quad \text{in} \quad \Omega,$$

and

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

We consider an integral equation for 2×2 matrix function $\mu^{(j)} = \begin{pmatrix} \mu_{11}^{(j)} & \mu_{12}^{(j)} \\ \mu_{21}^{(j)} & \mu_{22}^{(j)} \end{pmatrix}$, $j = 1, 2$, which is a key in [19], [20], [21]:

$$\mu^{(j)}(z,k) = I - \frac{1}{\pi}\int_\Omega \frac{e_k(\zeta)Q^{(j)}(\zeta)}{\zeta - z}\overline{\mu^{(j)}(\zeta,k)}d\zeta, \qquad z \in \Omega, \quad k \in \mathcal{C}, \quad j = 1, 2. \tag{4.8}$$

By [25], there exists a unique solution $\mu^{(j)}(\cdot, k) \in W^{1,p}(\Omega)$ to (4.8) for $j = 1, 2$.

Then, in terms of Lemma 4.2, an argument similar to the proof of Lemma 4.1 leads us to the following lemma.

LEMMA 4.3. *Let* $\Lambda_{b^{(1)}} = \Lambda_{b^{(2)}}$. *Then*

$$\mu^{(1)}(z,k) = \mu^{(2)}(z,k), \qquad z \in \partial\Omega, \quad k \in \mathcal{C}. \tag{4.9}$$

Proof. By [25], we have

$$I = \frac{1}{2\pi i}\int_{\partial\Omega} \frac{\mu^{(j)}(\zeta,k)}{\zeta - z}d\zeta, \qquad z \in \Omega, \quad k \in \mathcal{C}, \quad j = 1, 2. \tag{4.10}$$

We set

$$\omega^{(j)}(z,k) = \mu^{(j)}(z,k)e^{\frac{i}{2}\bar{k}z}, \qquad z \in \Omega, \quad k \in \mathcal{C}, \quad j = 1, 2. \tag{4.11}$$

Again, by [25], noticing that $\partial_{\bar z}(e^{\frac{i}{2}\bar{k}z}) = 0$, we can verify

$$\partial_{\bar z}\omega^{(j)} = Q^{(1)}\overline{\omega^{(j)}} \quad \text{in} \quad \Omega \tag{4.12}$$

for $j = 1, 2$.

We set

$$\omega^{(j)}(z,k) = \begin{pmatrix} \phi_{11}^{(j)}(z,k) & \phi_{12}^{(j)}(z,k) \\ \phi_{21}^{(j)}(z,k) & \phi_{22}^{(j)}(z,k) \end{pmatrix}, \qquad z \in \Omega, \quad k \in \mathcal{C}, \quad j = 1, 2. \tag{4.13}$$

We can rewrite (4.12) as

$$\begin{aligned} &\partial_{\bar z}\phi_{11}^{(1)} = -C^{(1)}\overline{\phi_{21}^{(1)}}, \quad \partial_{\bar z}\phi_{21}^{(1)} = -C^{(1)}\overline{\phi_{11}^{(1)}}, \\ &\partial_{\bar z}\phi_{12}^{(1)} = -C^{(1)}\overline{\phi_{22}^{(1)}}, \quad \partial_{\bar z}\phi_{22}^{(1)} = -C^{(1)}\overline{\phi_{12}^{(1)}} \end{aligned} \quad \text{in} \quad \Omega.$$

Setting

$$\begin{aligned} v_{11}^{(1)} &= \phi_{11}^{(1)} + \phi_{21}^{(1)}, \quad v_{21}^{(1)} = i(\phi_{11}^{(1)} - \phi_{21}^{(1)}), \\ v_{12}^{(1)} &= \phi_{12}^{(1)} + \phi_{22}^{(1)}, \quad v_{22}^{(1)} = i(\phi_{12}^{(1)} - \phi_{22}^{(1)}) \end{aligned} \qquad \text{in} \quad \Omega,$$

we can further rewrite (4.12) as

$$\partial_{\bar{z}} v_{lm}^{(1)} = -C^{(1)} \overline{v_{lm}^{(1)}} \qquad \text{in} \quad \Omega, \quad 1 \le l, \quad m \le 2. \tag{4.14}$$

Next, by setting

$$w_{lm}^{(1)}(z,k) = e^{-(TB^{(1)})(z)} v_{lm}^{(1)}(z,k), \qquad z \in \Omega, \quad k \in \mathcal{C}, \quad 1 \le l, \quad m \le 2, \tag{4.15}$$

the definition (4.7) and the equalities (4.14) yield

$$\partial_{\bar{z}} w_{lm}^{(1)} + B^{(1)} w_{lm}^{(1)} + \overline{B^{(1)} w_{lm}^{(1)}} = 0, \qquad z \in \Omega, \quad k \in \mathcal{C}, \quad 1 \le l, \quad m \le 2.$$

Therefore, for $1 \le l, m \le 2$, there exist $u_{lm}^{(1)} \in W^{2,p}(\Omega)$ such that $\partial_z u_{lm}^{(1)} = w_{lm}^{(1)}$ and

$$\Delta u_{lm}^{(1)} + b_1^{(1)} \frac{\partial u_{lm}^{(1)}}{\partial x_1} + b_2^{(1)} \frac{\partial u_{lm}^{(1)}}{\partial x_2} = 0 \qquad \text{in} \quad \Omega.$$

We put

$$f_{lm} = u_{lm}^{(1)}|_{\partial\Omega}.$$

By $f_{lm} \in W^{2-\frac{1}{p},p}(\partial\Omega)$, we can uniquely solve

$$\begin{aligned} \Delta u_{lm}^{(2)} + b_1^{(2)} \frac{\partial u_{lm}^{(2)}}{\partial x_1} + b_2^{(2)} \frac{\partial u_{lm}^{(2)}}{\partial x_2} &= 0 \qquad \text{in} \quad \Omega, \\ u_{lm}^{(2)}|_{\partial\Omega} &= f_{lm}, \qquad 1 \le l, \quad m \le 2, \end{aligned}$$

and

$$u_{lm}^{(2)} \in W^{2,p}(\Omega), \qquad 1 \le l, \quad m \le 2.$$

Then by $\Lambda_{b^{(1)}} = \Lambda_{b^{(2)}}$ we see that

$$\partial_z u_{lm}^{(2)} = \partial_z u_{lm}^{(1)} \qquad \text{on} \quad \partial\Omega, \quad 1 \le l, \quad m \le 2. \tag{4.16}$$

Conversely to the derivation for $v_{lm}^{(1)}$, we set

$$v_{lm}^{(2)} = e^{TB^{(2)}} \partial_z u_{lm}^{(2)} \qquad \text{in} \quad \Omega, \quad 1 \le l, \quad m \le 2.$$

Noticing Lemma 3.1 and (4.7), we see by direct calculations that Lemma 4.2 and (4.16) imply

$$\partial_{\bar{z}} v_{lm}^{(2)} + C^{(2)} \overline{v_{lm}^{(2)}} = 0 \qquad \text{in} \quad \Omega, \quad 1 \le l, \quad m \le 2, \tag{4.17}$$

and

$$v_{lm}^{(1)} = v_{lm}^{(2)} \qquad \text{on} \quad \partial\Omega. \tag{4.18}$$

Setting

$$\widetilde{\mu} = \begin{pmatrix} \widetilde{\phi}_{11} & \widetilde{\phi}_{12} \\ \widetilde{\phi}_{21} & \widetilde{\phi}_{22} \end{pmatrix} e^{-\frac{1}{2}i\bar{k}z}$$

and

$$\begin{aligned} &\widetilde{\phi}_{11} = \tfrac{1}{2}v_{11}^{(2)} - \tfrac{i}{2}v_{21}^{(2)}, \quad \widetilde{\phi}_{21} = \tfrac{1}{2}v_{11}^{(2)} + \tfrac{i}{2}v_{21}^{(2)}, \\ &\widetilde{\phi}_{12} = \tfrac{1}{2}v_{12}^{(2)} - \tfrac{i}{2}v_{22}^{(2)}, \quad \widetilde{\phi}_{22} = \tfrac{1}{2}v_{12}^{(2)} + \tfrac{i}{2}v_{22}^{(2)} \end{aligned} \qquad \text{in} \quad \Omega,$$

in terms of (4.11), (4.13), (4.15), (4.17), and (4.18), we can directly see that

$$\partial_{\bar{z}}\widetilde{\mu} = e_k Q^{(2)}\overline{\widetilde{\mu}} \qquad \text{in} \quad \Omega \tag{4.19}$$

and

$$\mu^{(1)} = \widetilde{\mu} \qquad \text{on} \quad \partial\Omega. \tag{4.20}$$

Equation (4.19) is rewritten as

$$\widetilde{\mu}(z) = \Phi_2(z) - \frac{1}{\pi}\int_\Omega \frac{e_k(\zeta)Q^{(2)}(\zeta)}{\zeta - z}\overline{\widetilde{\mu}(\zeta)}d\zeta, \qquad z \in \Omega, \tag{4.21}$$

where

$$\Phi_2(z) = \frac{1}{2\pi i}\int_{\partial\Omega}\frac{\widetilde{\mu}(\zeta)}{\zeta - z}d\zeta, \qquad z \in \Omega.$$

Here (4.20) and (4.10) imply

$$\Phi_2(z) = \frac{1}{2\pi i}\int_{\partial\Omega}\frac{\widetilde{\mu}(\zeta)}{\zeta - z}d\zeta = I, \qquad z \in \Omega.$$

The uniqueness of solutions of (4.21) with $\Phi_2 = I$ yields

$$\widetilde{\mu} = \mu^{(2)} \qquad \text{in} \quad \Omega,$$

with which (4.20) completes the proof of the lemma. □

Henceforth we take the zero extension of $B^{(j)}$, $j = 1, 2$, outside Ω:

$$B^{(j)}(z) = 0, \qquad z \in \mathcal{C} \setminus \Omega, \quad j = 1, 2.$$

We note that $C^{(j)}(z) = 0$, $Q^{(j)}(z) = 0$, $z \in \mathcal{C} \setminus \Omega$, $j = 1, 2$.

For $z \in \mathcal{C} \setminus \Omega$, we set

$$\mu^{(j)}(z, k) = I - \frac{1}{\pi}\int_\Omega \frac{e_k(\zeta)Q^{(j)}(\zeta)\overline{\mu^{(j)}(\zeta, k)}}{\zeta - z}d\zeta, \qquad k \in \mathcal{C}, \quad j = 1, 2.$$

Then we can rewrite (4.8) as

$$\mu^{(j)}(z, k) = I - \frac{1}{\pi}\int_{\mathcal{C}} \frac{e_k(\zeta)Q^{(j)}(\zeta)}{\zeta - z}\overline{\mu^{(j)}(\zeta, k)}d\zeta, \qquad z, k \in \mathcal{C}, \quad j = 1, 2. \tag{4.22}$$

Consequently, by [25] we can easily verify

$$\partial_{\bar{z}}\mu^{(j)}(z,k) = e_k(z)Q^{(j)}(z)\overline{\mu^{(j)}(z,k)}, \qquad z \in \mathcal{C}, \quad k \in \mathcal{C}, \quad j = 1,2,$$

and

$$\lim_{|z|\to\infty} \mu^{(j)}(z,k) = I, \qquad k \in \mathcal{C}, \quad j = 1,2,$$

On the basis of Lemma 4.3, we can apply the method in [19], [20], [21].

LEMMA 4.4. *If* $\Lambda_{b^{(1)}} = \Lambda_{b^{(2)}}$, *then*

$$C^{(1)}(z) = C^{(2)}(z), \qquad z \in \Omega.$$

Proof. We define $\nu^{(j)} = \nu^{(j)}(z,k)$, $z \in \Omega$, $k \in \mathcal{C}$, $j = 1,2$, by

$$\begin{aligned}\nu^{(j)}(z,k) &= \begin{pmatrix} \nu_{11}^{(j)}(z,k) & \nu_{12}^{(j)}(z,k) \\ \nu_{21}^{(j)}(z,k) & \nu_{22}^{(j)}(z,k) \end{pmatrix} \\ &= \begin{pmatrix} \overline{\mu_{11}^{(j)}(z,k)} & \mu_{12}^{(j)}(z,k)e_{-k}(z) \\ \mu_{21}^{(j)}(z,k)e_{-k}(z) & \overline{\mu_{22}^{(j)}(z,k)} \end{pmatrix},\end{aligned}$$

where we recall $e_{-k}(z) = e^{\frac{1}{2}i(\bar{k}z + k\bar{z})}$.

It is verified in [19], [20], [21] that $\nu^{(j)}$ satisfies the following first order elliptic equation with respect to k:

$$\frac{\partial \nu^{(j)}}{\partial \bar{k}}(z,k) = e_{-k}(z)T^{(j)}(k)\overline{\nu^{(j)}(z,k)}, \qquad k \in \mathcal{C}, \quad z \in \mathcal{C}, \quad j = 1,2,$$

where

$$T^{(j)}(k) = \begin{pmatrix} 0 & T_{12}^{(j)}(k) \\ T_{21}^{(j)}(k) & 0 \end{pmatrix}, \qquad k \in \mathcal{C}, \quad j = 1,2,$$

$$T_{12}^{(j)}(k) = \frac{i}{2\pi}\int_\Omega e_k(\zeta)C^{(j)}(\zeta)\overline{\mu_{22}^{(j)}(\zeta,k)}d\zeta, \tag{4.23}$$

$$T_{21}^{(j)}(k) = \frac{i}{2\pi}\int_\Omega e_k(\zeta)C^{(j)}(\zeta)\overline{\mu_{11}^{(j)}(\zeta,k)}d\zeta, \tag{4.24}$$

$$k \in \mathcal{C}, \quad j = 1,2.$$

By the symmetry of the matrix function $Q^{(j)}$, from (4.22) we can see that

$$\mu_{11}^{(j)}(z,k) = \mu_{22}^{(j)}(z,k), \qquad z \in \mathcal{C}, \quad k \in \mathcal{C}, \quad j = 1,2.$$

Therefore (4.23) and (4.24) imply

$$T_{12}^{(j)}(k) = T_{21}^{(j)}(k), \qquad k \in \mathcal{C}, \quad j = 1,2.$$

Next, we can see

$$T^{(1)}(k) = T^{(2)}(k), \qquad k \in \mathcal{C}. \tag{4.25}$$

In fact, for an arbitrarily fixed $k \in \mathcal{C}$, setting

$$\Psi^{(j)}(z) = \int_\Omega \frac{e_k(\zeta)Q^{(j)}(\zeta)\overline{\mu^{(j)}(\zeta,k)}}{\zeta - z}d\zeta, \quad z \in \mathcal{C}\setminus\Omega, \quad j = 1,2,$$

we see from (4.8) and (4.9) that

$$\Psi^{(1)}(z) = \Psi^{(2)}(z), \qquad z \in \partial\Omega.$$

Since $Q^{(j)} \in L^p(\Omega)$ and $\mu^{(j)} \in L^\infty(\Omega)$ for $j = 1,2$, we see by the Hölder inequality that there exist constant $M_5 > 0$ and $R > 0$ such that

$$|\Psi^{(1)}(z) - \Psi^{(2)}(z)| \le \frac{M_5}{|z|}, \qquad |z| > R.$$

Consequently, since $\Psi^{(1)} - \Psi^{(2)}$ is holomorphic in $\mathcal{C}\setminus\overline{\Omega}$, we apply the Cauchy integration formula on $\partial\Omega \cup \{z||z| = R_1\}$ with sufficiently large $R_1 > 0$ and let R_1 tend to ∞ so that we obtain

$$\Psi^{(1)}(z) = \Psi^{(2)}(z), \qquad z \in \mathcal{C}\setminus\Omega. \tag{4.26}$$

On the other hand, taking $z \in \mathcal{C}\setminus\Omega$ such that $|z|$ is so large that $\left|\frac{\zeta}{z}\right| < \frac{1}{2}$ for all $\zeta \in \overline{\Omega}$, we have

$$\begin{aligned}\Psi^{(j)}(z) &= \frac{1}{z}\int_\Omega \frac{e_k(\zeta)Q^{(j)}(\zeta)\overline{\mu^{(j)}(\zeta,k)}}{\frac{\zeta}{z} - 1}d\zeta \\ &= -\frac{1}{z}\sum_{n=0}^{\infty}\frac{1}{z^n}\int_\Omega e_k(\zeta)Q^{(j)}(\zeta)\overline{\mu^{(j)}(\zeta,k)}\zeta^n d\zeta, \quad j = 1,2.\end{aligned}$$

In terms of (4.26), comparing the coefficients of $n = 0$ of $\Psi^{(1)}(z)$ and $\Psi^{(2)}(z)$, we obtain

$$\int_\Omega e_k(\zeta)Q^{(1)}(\zeta)\overline{\mu^{(1)}(\zeta,k)}d\zeta = \int_\Omega e_k(\zeta)Q^{(2)}(\zeta)\overline{\mu^{(2)}(\zeta,k)}d\zeta.$$

Therefore from (4.23) and (4.24) we can obtain (4.25).

Hence $d(z,k) = \nu^{(1)}(z,k) - \nu^{(2)}(z,k)$ satisfies

$$\frac{\partial d}{\partial\overline{k}}(z,k) = e_{-k}(z)T^{(1)}(k)\overline{d(z,k)}, \qquad k \in \mathcal{C}, \quad z \in \mathcal{C}. \tag{4.27}$$

It is shown that

$$T^{(j)}(k) \in L^2(\mathcal{C}) \tag{4.28}$$

(Theorem B in [5]) and there exists a constant $q > \frac{2p}{p-2}$ such that

$$\sup_{z\in\mathcal{C}}\|\nu^{(j)}(z,\cdot) - I\|_{L^q(\mathcal{C})} < \infty \tag{4.29}$$

(Theorem 2.3 in [5]).

Since $T_{12}^{(j)} = T_{21}^{(j)}$, we can reduce (4.27) to four independent complex equations. Therefore in view of (4.28) and (4.29) a sharp version of the Liouville theorem (Theorem 3.1 in [5]) gives

$$d(z,k) = \nu^{(1)}(z,k) - \nu^{(2)}(z,k) = 0, \qquad z \in \Omega, \quad k \in \mathcal{C}.$$

On the other hand, by [19], [20], [21] we obtain

$$C^{(j)}(z) = \lim_{|k|\to\infty} \frac{ik}{2} \nu_{12}^{(j)}(z,k), \qquad z \in \mathcal{C}, \quad j = 1,2.$$

Consequently,

$$C^{(1)}(z) = C^{(2)}(z), \qquad z \in \Omega.$$

The proof is complete. □

In order to finish the proof of the main result, we have to prove

$$B^{(1)}(z) = B^{(2)}(z), \qquad z \in \Omega.$$

From Lemma 4.4 and (4.7), we can see that

$$B^{(1)}(z) e^{(\overline{TB^{(1)}} - TB^{(1)})(z)} = B^{(2)}(z) e^{(\overline{TB^{(2)}} - TB^{(2)})(z)}, \qquad z \in \Omega. \tag{4.30}$$

Since we set $B^{(j)}(z) = 0$, $j = 1,2$, for $z \in \mathcal{C} \setminus \Omega$, equality (4.30) holds for all $z \in \mathcal{C}$.

Let $\Xi(z) = T(B^{(2)} - B^{(1)})(z)$ and $\kappa(z) = e^{\overline{T(B^{(2)} - B^{(1)})(z)} - T(B^{(2)} - B^{(1)})(z)}$. From (4.30), we can see that

$$B^{(1)}(z) = \kappa(z) B^{(2)}(z), \qquad z \in \mathcal{C}.$$

Therefore we can obtain that

$$\partial_{\bar{z}} \Xi = B^{(2)} - B^{(1)} = (1 - \kappa) B^{(2)} \qquad \text{in} \quad \mathcal{C}.$$

Let $i\theta = \overline{T(B^{(2)} - B^{(1)})} - T(B^{(2)} - B^{(1)})$. Then θ is real. Therefore we can obtain

$$|\kappa - 1| = \left| 2 \sin \frac{\theta}{2} \left(-\sin \frac{\theta}{2} + i \cos \frac{\theta}{2} \right) \right| \le 2 \left| \sin \frac{\theta}{2} \right| \le |\theta|.$$

It is easy to verify directly that

$$|\theta| = |\overline{T(B^{(2)} - B^{(1)})} - T(B^{(2)} - B^{(1)})| \le 2|T(B^{(2)} - B^{(1)})| = 2|\Xi| \tag{4.31}$$

so that

$$|\partial_{\bar{z}} \Xi| \le |B^{(2)}||1 - \kappa| \le 2|B^{(2)}||\Xi| \qquad \text{in} \quad \mathcal{C}. \tag{4.32}$$

From Lemma 4.2, we have

$$\Xi(z) = 0, \quad z \in \mathcal{C} \setminus \Omega. \tag{4.33}$$

Let us take a bounded domain Ω_1 with smooth boundary $\partial\Omega_1$ such that $\overline{\Omega} \subset \Omega_1$. We directly see that $\Xi \in W^{1,p}(\Omega_1)$. Then, since $\mathrm{supp}\Xi \subset \overline{\Omega}$, by (4.33) we have the following Carleman estimate:

$$\int_{\Omega_1} \Delta\varphi |\Xi|^2 e^{\varphi} dx \le 4 \int_{\Omega_1} |\partial_{\bar{z}} \Xi|^2 e^{\varphi} dx, \tag{4.34}$$

where φ is a real-valued function and $\varphi \in W^{2,\frac{p}{2}}(\Omega_1)$.

In fact, by Lemma 2.2 in [6] or Lemma 6.1 in the appendix we directly see (4.34) for $\Xi \in C^1(\overline{\Omega}_1)$ and $\varphi \in C^2(\overline{\Omega}_1)$. By $p > 2$ and Sobolev's embedding, we can take approximation sequences, and we obtain (4.34) for $\Xi \in W^{1,p}(\Omega_1)$ and $\varphi \in W^{2,\frac{p}{2}}(\Omega_1)$.

From (4.32), we obtain that

$$\int_{\Omega_1} \Delta\varphi |\Xi|^2 e^{\varphi} dx \le 16 \int_{\Omega_1} |B^{(2)}|^2 |\Xi|^2 e^{\varphi} dx.$$

Since $|B^{(2)}| \in L^p(\Omega)$ $(p > 2)$, we can choose

$$\varphi(x) = \frac{16}{\pi} \int_{\Omega_1} \ln|x-y| |B|^2 dy \in W^{2,\frac{p}{2}}(\Omega),$$

where $B = \max(|B^{(2)}|, 1)$.

Therefore, noticing $\Delta\varphi = 32|B|^2$ in Ω_1, we obtain that

$$\int_{\Omega_1} |B|^2 |\Xi|^2 e^{\varphi} dx = 0,$$

i.e.,

$$\Xi(z) = 0, \qquad z \in \Omega. \tag{4.35}$$

This means that $B^{(1)}(z) = B^{(2)}(z)$, $z \in \Omega$. Therefore we have that

$$b_1^{(1)}(z) = b_1^{(2)}(z), \quad b_2^{(1)}(z) = b_2^{(2)}(z), \qquad z \in \Omega.$$

The proof of the main result is complete.

Remark 4.1. We note that $B^{(2)}$ is in $L^p(\Omega)$ and thus is not necessarily bounded. Therefore we need special cares in concluding (4.35) from (4.32) by a Carleman estimate. For that, we used the Carleman estimate by [6], which is different from the one in [9].

5. Concluding remarks. I. We can realize the previous arguments for the uniqueness in order to establish the reconstruction algorithm. For higher-dimensional cases, we refer the reader to [16]. Correspondingly to the uniqueness arguments, the algorithm is composed of three steps:

First step. Reconstruct boundary values of TB. This is based on a uniquely solvable linear integral equation on $\partial\Omega$.

Second step. Reconstruct C in Ω. This step is carried out by the inverse scattering method (see, e.g., [19], [20], [21]). See also [16].

Third step. Solve $C = \overline{B} e^{TB - \overline{TB}}$ in Ω. This solution is obtained from a Dirichlet problem for a nonlinear $\partial_{\bar{z}}$-equation.

Every step requires verification, and we have to guarantee the unique solvability, especially in the first and third steps. In a succeeding paper, we give the details.

II. We will show that, in the two-dimensional case, at most two functions can be determined from the Dirichlet to Neumann map, as the following example shows.

Consider two elliptic equations in a simply connected bounded domain Ω whose boundary $\partial\Omega$ is sufficiently smooth:

$$\operatorname{div}(\gamma(x)\nabla u^{(1)}(x)) + b_1^{(1)}(x)\frac{\partial u^{(1)}}{\partial x_1}(x) + b_2^{(1)}(x)\frac{\partial u^{(1)}}{\partial x_2}(x) = 0, \quad x \in \Omega, \tag{5.1}$$

and

$$\Delta u^{(2)}(x) + b_1^{(2)}(x)\frac{\partial u^{(2)}}{\partial x_1}(x) + b_2^{(2)}(x)\frac{\partial u^{(2)}}{\partial x_2}(x) = 0, \quad x \in \Omega. \tag{5.2}$$

Here γ, $b_1^{(j)}$, $b_2^{(j)}$, $j = 1, 2$, are in $C^\infty(\overline{\Omega})$ and $\gamma(x) > \gamma_0 > 0$, $x \in \Omega$, where γ_0 is a constant.

It is well known that (5.1) and (5.2) with Dirichlet boundary condition

$$u^{(j)}(x) = f(x), \qquad x \in \partial\Omega,$$

are uniquely solvable in $C^2(\Omega) \cap C^1(\overline{\Omega})$ for all $f \in C^3(\partial\Omega)$.

Then we can define the Dirichlet to Neumann maps Λ_1 and Λ_2 for (5.1) and (5.2) by

$$(\Lambda_1 f)(x) = \gamma(x)\frac{\partial u^{(1)}}{\partial \nu}(x), \qquad x \in \partial\Omega,$$

and

$$(\Lambda_2 f)(x) = \frac{\partial u^{(2)}}{\partial \nu}(x), \qquad x \in \partial\Omega,$$

respectively.

For γ such that $\gamma|_{\partial\Omega} = 1$, we set

$$\begin{aligned} b_1^{(2)}(x) &= \frac{1}{\gamma(x)}\left(b_1^{(1)}(x) + \frac{\partial \gamma}{\partial x_1}(x)\right), \\ b_2^{(2)}(x) &= \frac{1}{\gamma(x)}\left(b_2^{(1)}(x) + \frac{\partial \gamma}{\partial x_2}(x)\right), \qquad x \in \Omega. \end{aligned}$$

Then (5.1) coincides with (5.2). Therefore we see that

$$\Lambda_1 = \Lambda_2.$$

This means that we cannot determine all of $b_1^{(1)}$, $b_2^{(1)}$, and γ.

Moreover, we cannot determine all b_1, b_2, and E in the elliptic equation

$$\Delta u(x) + b_1(x)\frac{\partial u}{\partial x_1} + b_2(x)\frac{\partial u}{\partial x_2} + E(x)u(x) = 0, \quad x \in \Omega$$

(pp. 121–122 in [11]).

6. Appendix. For the reader's convenience, we present an outline of the proof for Carleman estimate (4.34) which was given in [6].

LEMMA 6.1. *Suppose that $\varphi \in C^2(\overline{\Omega})$ is a real function. Then, for any $u \in W_0^{1,p}(\Omega)$, it holds that*

$$\int_\Omega \Delta\varphi |u|^2 e^\varphi dxdy + 4\int_\Omega |(\partial_z + \partial_z\varphi)u|^2 e^\varphi dxdy = 4\int_\Omega |\partial_{\overline{z}} u|^2 e^\varphi dxdy, \tag{6.1}$$

where $\nu = (\nu_1, \nu_2)$ is the outer unit normal vector to $\partial\Omega$ and $\nu^\perp = (-\nu_2, \nu_1)$.

Proof. By the Green formula and $u|_{\partial\Omega} = 0$, we have

$$\int_\Omega |\partial_{\overline{z}} u|^2 e^\varphi dxdy = \int_\Omega \partial_{\overline{z}} u \partial_z \overline{u} e^\varphi dxdy = -\int_\Omega u \partial_{\overline{z}}[\partial_z \overline{u} e^\varphi] dxdy$$

and

$$\int_\Omega |(\partial_z + \partial_z\varphi)u|^2 e^\varphi dxdy = \int_\Omega (\partial_z + \partial_z\varphi)u \cdot (\partial_{\overline{z}} + \partial_{\overline{z}}\varphi)\overline{u}e^\varphi dxdy$$
$$= -\int_\Omega u\partial_z[(\partial_{\overline{z}}\overline{u} + (\partial_{\overline{z}}\varphi)\overline{u})e^\varphi]dxdy$$
$$+ \int_\Omega (\partial_z\varphi)u(\partial_{\overline{z}}\overline{u} + (\partial_{\overline{z}}\varphi)\overline{u})e^\varphi dxdy.$$

By these two equalities, direct calculation implies that

$$\int_\Omega |\partial_{\overline{z}}u|^2 e^\varphi dxdy - \int_\Omega |(\partial_z + \partial_z\varphi)u|^2 e^\varphi dxdy = \frac{1}{4}\int_\Omega \Delta\varphi|u|^2 e^\varphi dxdy.$$

The proof is completed. □

Acknowledgments. The authors thank the referees' suggestions and comments. The authors thank Professor V. Isakov (Wichita State University) and Professor G. Nakamura (Hokkaido University) for valuable discussions and comments.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] J. Barros-Neto, *An Introduction to the Theory of Distributions*, Marcel Dekker, New York, 1973.
[3] R. Beals and R. R. Coifman, *Multidimensional inverse scatterings and nonlinear partial differential equations*, in Pseudodifferential Operators and Applications, Proc. Sympos. Pure Math. 43, F. Trèves, ed., AMS, Providence, RI, 1985, pp. 45–70.
[4] L. Bers, *Theory of Pseudo-Analytic Functions*, Institute for Mathematics and Mechanics, New York University, New York, 1953.
[5] R. M. Brown and G. A. Uhlmann, *Uniqueness in the inverse conductivity problem for nonsmooth conductivities in two dimensions*, Comm. Partial Differential Equations, 22 (1997), pp. 1009–1027.
[6] A. L. Bukhgeim, *Extension of solutions of elliptic equations from discrete sets*, J. Inverse Ill-posed Problems, 1 (1993), pp. 17–32.
[7] A. P. Calderón, *On an inverse boundary value problem*, in Seminar on Numerical Analysis and its Applications to Continuum Physics, Soc. Brasileira de Matematica, Rio de Janeiro, Brazil, 1980, pp. 65–73.
[8] M. Cheney, *A review of multidimensional inverse potential scattering*, in Inverse Problems in Partial Differential Equations, Proc. Appl. Math. 42, D. Colton, R. Ewing, and W. Rundell, eds., SIAM, Philadelphia, 1990, pp. 37–49.
[9] L. Hörmander, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, 1963.
[10] M. Ikehata and G. Nakamura, *Inverse boundary value problem—15 years since Calderón raised the problem*, Sugaku Expositions, 12 (1999), pp. 57–84.
[11] V. Isakov, *Inverse Problems for Partial Differential Equations*, Springer-Verlag, New York, 1998.
[12] V. Isakov and A. Nachman, *Global uniqueness for a two-dimensional semilinear elliptic inverse problem*, Trans. Amer. Math. Soc., 347 (1995), pp. 3375–3390.
[13] R. Kohn and M. Vogelius, *Determining conductivity by boundary measurements*, Comm. Pure Appl. Math., 37 (1984), pp. 289–298.
[14] R. Kohn and M. Vogelius, *Determining conductivity by boundary measurements,* II*, Interior results*, Comm. Pure Appl. Math., 38 (1985), pp. 643–667.
[15] A. I. Nachman, *Reconstructions from boundary measurements*, Ann. of Math. (2), 128 (1988), pp. 531–576.
[16] A. I. Nachman, *Global uniqueness for a two-dimensional inverse boundary value problem*, Ann. of Math. (2), 143 (1996), pp. 71–96.
[17] G. Nakamura, Z. Sun, and G. Uhlmann, *Global identifiability for an inverse problem for the Schrödinger equation in a magnetic field*, Math. Ann., 303 (1995), pp. 377–388.

[18] Z. Sun, *The inverse conductivity problem in two dimensions*, J. Differential Equations, 87 (1990), pp. 227–255.
[19] L. Y. Sung, *An inverse scattering transform for the Davey-Stewartson* II *equations,* I, J. Math. Anal. Appl., 183 (1994), pp. 121–154.
[20] L. Y. Sung, *An inverse scattering transform for the Davey-Stewartson* II *equations,* II, J. Math. Anal. Appl., 183 (1994), pp. 289–325.
[21] L. Y. Sung, *An inverse scattering transform for the Davey-Stewartson* II *equations,* III, J. Math. Anal. Appl., 183 (1994), pp. 477–494.
[22] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem*, Ann. of Math. (2), 125 (1987), pp. 153–169.
[23] J. Sylvester and G. Uhlmann, *Inverse boundary value problems at the boundary – continuous dependence*, Comm. Pure Appl. Math., 41 (1988), pp. 197–219.
[24] J. Sylvester and G. Uhlmann, *The Dirichlet to Neumann map and its applications*, in Inverse Problems in Partial Differential Equations, Proc. Appl. Math. 42, D. Colton, R. Ewing, and W. Rundell, eds., SIAM, Philadelphia, 1990, pp. 101–139.
[25] I. N. Vekua, *Generalized Analytic Functions*. Pergamon Press, London, 1962.

SIAM J. MATH. ANAL.
Vol. 35, No. 6, pp. 1394–1419

NEURONAL OSCILLATIONS IN THE VISUAL CORTEX: Γ-CONVERGENCE TO THE RIEMANNIAN MUMFORD–SHAH FUNCTIONAL*

GIOVANNA CITTI†, MARIA MANFREDINI†, AND ALESSANDRO SARTI‡

Abstract. The aim of this paper is to provide a formal link between an oscillatory neural model, whose phase is represented by a difference equation, and the Mumford and Shah functional. A Riemannian metric is induced by the pattern of neural connections, and in this setting the difference equation is studied. Its Euler–Lagrange operator Γ-converges as the dimension of the grid tends to 0 to the Mumford and Shah functional in the same Riemannian space. Correspondingly, the solutions of the phase equation converge to a BV function, which is interpreted as the flow associated with the Mumford and Shah functional. In this way we provide a biological motivation to this celebrated functional.

Key words. neural oscillators, Cauchy problem for a difference equation, variational problems, Riemannian metrics, Γ-convergence, Mumford and Shah functional

AMS subject classifications. 49J45, 65K10, 39A70, 92C20

DOI. 10.1137/S0036141002398673

1. Introduction.

An intriguing issue that has to be dealt with in the mammalian visual system is how the information distributed in the visual cortex gets bound together into coherent object representations. Along the path going from the physical object to the observer, radiations are completely independent one of the other. The retina is constituted in its turn by a mosaic of histologically separated elements. At the end of this chain, during which the unity of the original object is completely lost, the object shows up again at the perceptual level as a unit. In which way is it possible to reconstruct at the perceptual level the unity of the physical object? This process is known as "binding" or "perceptual grouping," and it has been extensively studied at least from two different points of view: From one side it has been the subject of research in the experimental psychology of Gestalt, oriented to infer the phenomenological laws of perceptual organization [33]. On the other side, neurophysiological studies have been focused on the determination of biological functionalities underlying grouping. In this paper we prove a formal relation between two of these models: a difference equation describing the phase of neuronal oscillators in the visual cortex, and the celebrated Mumford and Shah functional, first introduced as a phenomenological model. The family of discrete Euler–Lagrange functionals associated with the phase equation Γ-converges as the length of the grid tends to 0 to the Mumford and Shah functional in a BV space related to a Riemannian metric.

1.1. A phenomenological model.

Mumford and Shah in their celebrated paper [36] proposed to obtain the segmentation of a given image u_0 as a minimum of

*Received by the editors March 26, 2002; accepted for publication (in revised form) July 4, 2003; published electronically February 18, 2004. This work was supported by the University of Bologna: founds for selected research topics.

http://www.siam.org/journals/sima/35-6/39867.html

†Dipartimento di Matematica, Universita' di Bologna, Piazza di Porta S. Donato 5, 40127 Bologna, Italia (citti@dm.unibo.it, manfredi@dm.unibo.it).

‡DEIS, Universita' di Bologna, Via Risorgimento 2, Bologna, Italia (asarti@deis.unibo.it).

the following functional:

$$E(u,K)=\alpha\int_{\mathbb{R}^n\setminus K}|\nabla u|^2dx+\beta dH^{n-1}(K)+\int_{\mathbb{R}^n}|u-u_0|^2dx,$$

where K is closed, and $u\in W^{1,2}(\Omega\setminus K)$. This functional has been deeply studied in the weak formulation, provided by De Giorgi, Carriero, and Leaci in [20], who allow u to be a SBV function and K its jump set:

$$MS(u)=\alpha\int_{\mathbb{R}^n}|\nabla u|^2dx+\beta dH^{n-1}(S(u))+\int_{\mathbb{R}^n}|u-u_0|^2dx. \tag{1.1}$$

In the same paper [20] the existence of minima has been proved; their lower semicontinuity has been proved by Ambrosio [1]. The main properties of the minima have been established by Ambrosio and Pallara [3], Ambrosio, Fusco, and Pallara [4], Bonnet [8], David [18], and Bonnet and David [9].

It has also been deeply studied in the problem of Γ-approximation of the functional MS, with elliptic functionals. Different families of approximating functionals have been proposed by Ambrosio and Tortorelli [5], Braides and Dal Maso [13], and Gobbino [28], who proved a conjecture of De Giorgi. We are interested in this last result, since it is an approximation of the MS functional with discrete functionals:

$$\frac{1}{\epsilon^{n+1}}\int_{\Omega\times\Omega}arctan\left(\frac{(u(x+\xi)-u(x))^2}{|\xi|}\right)e^{-\frac{|\xi|^2}{\epsilon}}dxd\xi. \tag{1.2}$$

Similar approximation problems have also been studied in [10, 11, 12, 14] in order to investigate the relation between the finite difference expression of the energy of elastic media and its continuous counterpart.

Here we will study the difference equation satisfied by the phase of neural oscillators with a technique similar to the one introduced in [28] and prove that it naturally leads to a nonisotropic version of the Mumford and Shah functional. A different approximation of nonisotropic functionals, analogous to [13], had already been provided by Cortesani [17]. Properties of minima of general anisotropic functionals of MS type have been established by Fonseca and Fusco [26], Trombetti [40], and Fusco, Mingione, and Trombetti [27]. We also refer the reader to Baldi for a degenerate functional of this type [6].

1.2. A neurophysiological model. From the neurological point of view there is a large amount of experimental evidence that grouping is represented in the brain with a temporal coding, meaning that semantically homogeneous areas in the image would be encoded in the synchronization (phase locking) of oscillatory neural responses [22]. Shuster and Wagner [38, 39] described the emergence of oscillations in the visual cortex by modelling every cortical column by densely connected Wilson–Cowan neurons [41]. The appropriate mean field equations for the cluster of neurons show that every column can be interpreted as an oscillator. The visual cortex is then modelled as a collection of oscillators coupled with long range sparse interactions, represented by the reduced phase equation, on a grid of length 1:

$$\partial_t u(t)=\Delta_{-\xi}\Big(\phi\big(\Delta_\xi u\big)\Big)(x), \tag{1.3}$$

where Δ_ξ is the difference operator which acts as follows on each function f:

$$\Delta_\xi f(x)=f(x+\xi)-f(x).$$

The function ϕ is continuous, odd, and periodic of period 2π so that $\phi(\pi) = \phi(-\pi) = 0$.

The same equation can be adapted to a grid of arbitrary length. Since the function u represents the phases of the oscillators, we can assume that $\Delta_\xi u$ takes its values in the interval $[-\pi, \pi]$ and $\phi = 0$ in $\mathbb{R}\backslash[-\pi, \pi]$. If $p > 0$, we call

$$\phi_{|\xi|}(z) = \frac{1}{|\xi|^{1-1/p}} \phi\big(|\xi|^{1/p} z\big), \quad |\xi| \neq 0, \tag{1.4}$$

and a suitable rescaling of the function u is a solution of the equation

$$\partial_t u(t) = \frac{1}{|\xi|} \Delta_{-\xi} \left(\phi_{|\xi|} \left(\frac{\Delta_\xi u}{|\xi|} \right) \right)(x).$$

This finite difference degenerate parabolic equation has been extensively studied in one dimension in [34, 35]. Its ability to reach phase locking solutions and to present phase discontinuities has been outlined.

In higher dimension Shuster and Wagner also proposed to convolve with a Gaussian kernel, which expresses the probability that an oscillator is connected to another. They obtain the equation

$$\partial_t u(t) = \int_{\mathbb{R}^n} e^{-\frac{|\xi|}{\epsilon}} \frac{1}{|\xi|} \Delta_{-\xi} \left(\phi_{|\xi|} \left(\frac{\Delta_\xi u}{|\xi|} \right) \right) \frac{d\xi}{\epsilon^n} \tag{1.5}$$

with the change of variable $\eta = \xi/\epsilon$

$$= \int_{\mathbb{R}^n} \frac{1}{\epsilon|\eta|} \Delta_{-\epsilon\eta} \left(e^{-|\eta|} \phi_{\epsilon|\eta|} \left(\frac{\Delta_{\epsilon\eta} u}{\epsilon|\eta|} \right) \right) d\eta.$$

In this study we consider (1.5) in the n-dimensional space and with space variant anisotropic connections. Indeed, several neurophysiological studies show that the association field between cortical columns are space variant and strongly anisotropic [25]. Riemannian metric is directly induced by the coupling strength between cortical columns.

A Riemannian metric is defined in $\mathbb{R}^n$ if at every point there is defined a matrix g_{ij} positive defined and continuous. In this case we call the Riemannian norm $|\eta|_g = g_{ij}\eta_i\eta_j$ and the Riemannian difference quotient

$$D^\epsilon_{g\eta} u(x) := \begin{cases} \frac{(\Delta_{\epsilon\eta} u(x)) mod(2\pi)}{\epsilon|\eta|_g} & \text{if } \epsilon|\eta|_g \neq 0, \\ 0 & \text{if } \epsilon|\eta|_g = 0. \end{cases} \tag{1.6}$$

If g is the identity, this difference quotient reduces to the standard one, and we denote it $D^\epsilon_\eta u$.

The resulting equation is then

$$\partial_t u(t) = \int_{\mathbb{R}^n} D^{-\epsilon}_\eta \left(\frac{|\eta|}{|\eta|_g} e^{-|\eta|_g} \phi_{\epsilon|\eta|_g} \left(D^{-\epsilon}_{g\eta} u \right) h \right) d\eta, \tag{1.7}$$

for a continuous function h, where $\phi_{\epsilon|\eta|_g}$ is defined in (1.4).

1.3. Relation between the stated models. In this paper we prove a first relation between the stated models, and we provide a biological motivation for the Mumford and Shah functional. We prove the existence of a solution u_ϵ of the Cauchy problem associated with (1.7), defined for all $t \geq 0$, and we prove that it Γ-converges as ϵ goes to 0 to the gradient flow relative to the Mumford and Shah functional in the Riemannian space with metric g_{ij}.

Precisely the Euler–Lagrange functional associated with (1.7) is

$$F_\epsilon(u) = \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} e^{-|\eta|_g} \varphi_{\epsilon|\eta|_g} \left(D^\epsilon_{g\eta} u \right) h(x) dx \right) d\eta, \tag{1.8}$$

where $\varphi_{\epsilon|\xi|}$ is a primitive of the function $\phi_{\epsilon|\xi|}$ defined in (1.4) and the following theorem holds.

THEOREM 1.1. *Assume as before that ϕ is continuous, it is odd, $\phi > 0$ in $[0, \pi[$, and $\phi = 0$ on $[\pi, \infty[$. Let us call β the constant value assumed by the primitive φ of ϕ on the interval $[\pi, \infty[$, and assume that there exist constants $\alpha > 0$ and $p > 1$ such that*

$$\frac{\varphi(z)}{z^p} \to \alpha \neq 0 \quad as \quad z \to 0^+. \tag{1.9}$$

Then the family F_ϵ defined in (1.8) *Γ-converges in $L^1_{loc}(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z})$ to the Mumford and Shah functional*

$$MS(u, \mathbb{R}^n) = \alpha\, c_{np} \int_{\mathbb{R}^n} |\nabla_g u|^p_g \frac{h(x)}{\sqrt{g(x)}} dx + \beta\, c_{n1} \int_{S(u)} |\nu_g|_g \frac{h(x)}{\sqrt{g(x)}} dH^{n-1} \tag{1.10}$$

if $u \in SBV$, $MS(u, \mathbb{R}^n) = +\infty$ otherwise. $S(u)$ is the jump set of u, ν_g is the normal to $S(u)$ in the Riemannian metrics, $g = \det(g_{ij})$, and c_{np} and c_{n1} are dimensional constants, defined in (2.2). *(We refer the reader to section* 2, *where the formal definitions of the jump set and the metric are recalled).*

REMARK 1.1. *The Riemannian Mumford and Shah functional is obtained for $h = g$ and $p = 2$:*

$$MS(u, \mathbb{R}^n) = \alpha\, c_{n2} \int_{\mathbb{R}^n} |\nabla_g u|^2_g \sqrt{g} dx + \beta\, c_{n1} \int_{S(u)} |\nu_g|_g \sqrt{g} dH^{n-1}.$$

In the limit case $p = 1$, the functional MS becomes the total variation functional, and an approximation result can be obtained with a modification of the technique used here as in [30].

The functional F_ϵ is a generalization of a Riemannian setting of the functional studied in [28]. The proof in this last paper is based on the slicing method and uses in full strength the isotropy of the functional. The main idea of our proof is the adaptation of the known technique to an anisotropic setting. Indeed, we first note that any Riemannian metric admits a representation of the form

$$g^{ij} \xi_i \xi_j = c_{n2} \int_{\mathbb{R}^n} e^{-|\eta|_g} \frac{(\langle \xi, \eta \rangle)^2}{|\eta|^2_g} \sqrt{g} d\eta, \tag{1.11}$$

where c_{n2} is a constant, depending on the dimension of the space (see Proposition 2.5). This representation allows us to write g in terms of an isotropic scalar product

and to extend to an anisotropic situation a convergence result known in the isotropic case.

As an application of the Γ-convergence Theorem 1.1, we prove an approximation result for minima of the *MS* functional.

THEOREM 1.2. *Let $1 \leq q < +\infty$, and let $g \in L^q(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z})$. Then for every $\epsilon > 0$ there exists a solution (u_ϵ) of the minimum problem*

$$m_\epsilon = \min\left\{ F_\epsilon(u) + \int_{\mathbb{R}^n} |u-g|^q dx : u \in BV(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z}), |Du|(\mathbb{R}^n) \leq \frac{1}{\epsilon} \right\}.$$

Moreover, for every sequence (ϵ_j) with $\epsilon_j \to 0$ the family (u_{ϵ_j}) has a subsequence converging in L^1_{loc} to a solution of the minimum problem

$$m_0 = \min\left\{ MS(u, \mathbb{R}^n) + \int_{\mathbb{R}^n} |u-g|^q dx, u \in SBV(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z}) \right\}. \tag{1.12}$$

Finally, $m_\epsilon \to m_0$ as $\epsilon \to 0$.

The proof is mainly based on a compactness result for a family of functions u_ϵ such that $F_\epsilon(u_\epsilon)$ is bounded. Indeed, since for every ϵ the functional F_ϵ has a minimum, by the compactness result, all the minima belong to the same compact subset of BV. Once this is established, the existence of the minimum point for MS follows from a general property of the Γ-convergence.

Then we apply the Γ-convergence result to the difference equation (1.7). For a fixed function $u_0 \in BV(\mathbb{R}, \mathbb{R}/2\pi\mathbb{Z})$, we consider a piecewise constant approximating family $(u_{0\epsilon})$ and for every $\epsilon > 0$ the problem

$$\begin{cases} \partial_t u_\epsilon(t) = -\nabla F_\epsilon(u_\epsilon(t)), \quad t \geq 0, \\ \\ \qquad u_\epsilon(0) = u_{0\epsilon}. \end{cases} \tag{1.13}$$

We prove that the solution (u_ϵ) is defined for every $t > 0$ and belongs to $C([0, +\infty[; L^p_{loc}(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z}))$.

It converges in BV to a function $u \in C([0, +\infty[; L^p_{loc}(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z}))$, which will then be interpreted as the flow associated with the Mumford and Shah functional, with initial datum u_0. This function u is a natural candidate for the flow associated with the Mumford and Shah functional. By now we can give only a characterization for u under the additional assumption that $p = 2$ and out of the jump set (see Corollary 5.5 below). The problem of the behavior of the jump set is still open, even in the Euclidean situation.

This paper is organized as follows. In section 2 we give some preliminary definitions of Γ-convergence and of Riemannian manifold. In sections 3 and 4, respectively, we prove Theorems 1.1 and 1.2. Finally, in section 5 we describe the behavior of the flow.

2. Preliminary definitions and notations.

2.1. *BV* functions and Γ-convergence. In this section we recall the definitions of functions of bounded variation and of Γ-convergence of functionals.

The class of BV functions is a class of functions whose distributional derivative is a nonnegative measure. We recall here the definition and refer the reader to [24] or

[23], where these notions are presented in full details. See also [21], where the set of BV functions with values in $\mathbb{R}/2\pi\mathbb{Z}$ is studied.

DEFINITION 2.1. *Let $\Omega \subset \mathbb{R}^n$ be an open set. We denote $M(\Omega)$ the set of all signed Radon measures on Ω with bounded total variation. We say that a function $u \in L^1(\Omega, \mathbb{R}/2\pi\mathbb{Z})$ is a function of bounded variation, and we write $u \in BV(\Omega, \mathbb{R}/2\pi\mathbb{Z})$ if all its distributional derivatives $D_i u$, $i = 1, \dots, n$, belong to $M(\Omega)$. It is well known that the following relation is satisfied almost everywhere:*

$$\lim_{\rho \to 0} \rho^{-n} \int_{B_\rho(x)} |u(y) - z| dy = 0$$

for some $z \in \mathbb{R}$, and all points x satisfying this relation are called Lebesgue points. The jump set $S(u)$ is the complementary of the set of Lebesgue points of u. If $u \in BV(\Omega, \mathbb{R}/2\pi\mathbb{Z})$, then the set $S(u)$ has Hausdorff measure at most $n-1$. Moreover, for H^{n-1} in almost every $x \in S(u)$ it is possible to find a, $b \in \mathbb{R}/2\pi\mathbb{Z}$ and a unitary vector ν such that

$$\lim_{\rho \to 0} \rho^{-n} \int_{B_\rho^\nu(x)} |u(y) - a| dy = 0, \quad \lim_{\rho \to 0} \rho^{-n} \int_{B_\rho^{-\nu}(x)} |u(y) - b| dy = 0,$$

where $B_r^\nu(x)$ is the half sphere $\{y \in B_r(x) : \langle y - x, \nu \rangle > 0\}$. The triplet (a, b, ν) is uniquely determined up to a change of sign, and it will be denoted $(u^+(x), u^-(x), \nu_u(x))$.

The distributional derivative Du admits the following decomposition:

$$Du = D^a u + D^j u + D^c u,$$

where $D^a u = \nabla u \mathcal{L}_n$ is absolutely continuous with respect to the Lebesgue measure $\mathcal{L}_n$,

$$D^j u = \big(u^+(x) - u^-(x)\big) \nu_u H^{n-1} \lfloor S(u)$$

is the jump part, and $D^c u$ is the Cantor part of Du.

A BV function u is a special function of bounded variation if $D^c u = 0$ and the set of these functions is denoted $SBV(\Omega)$. A function u belongs to $SBV_{loc}(\Omega, \mathbb{R}/2\pi\mathbb{Z})$ if $u \in SBV(A, \mathbb{R}/2\pi\mathbb{Z})$ for all $A \subset\subset \Omega$.

Let us now recall the De Giorgi definition of Γ-convergence.

DEFINITION 2.2. *If (X, d) is a metric space, a family $F_j : X \to \mathbb{R}$ of functionals Γ-converges to F as $j \to \infty$ if the following two conditions are satisfied:*

(i) *for every u in X and any sequence (u_j) converging to u in X,*

$$F(u) \leq \liminf_j F_j(u_j);$$

(ii) *for every $u \in X$ there exists a sequence (u_j) converging to u in X such that*

$$F(u) \geq \limsup_j F_j(u_j).$$

This notion of convergence captures the behavior of minimizers in the sense of the following theorem.

THEOREM 2.3. *Let us suppose that the family F_j of functionals Γ-converges to F as $j \to +\infty$ and that there exists a compact set K such that F_j takes its minimum on K for every $j \in \mathbb{N}$. Then F has a minimum.*

We also refer the reader to [19], where these notions are introduced and described.

2.2. Riemannian metrics. In this subsection we recall the definition of Riemannian metric and refer the reader to [32] for a detailed presentation.

DEFINITION 2.4. *A Riemannian metric on a differentiable manifold M is given by a scalar product on each tangent space T_qM, $q \in M$, which depends smoothly on the point q.*

Thus, if M has dimension n and $x = (x^1, \dots, x^n)$ are local coordinates of M, then a metric can be represented by a positive definite, symmetric matrix $G(x) = (g_{ij}(x))_{i,j}$ whose coefficients depend smoothly on x. Besides, the scalar product of two tangent vectors $v, w \in T_qM$ is $\langle v, w\rangle_g = g_{ij}(x)v^iw^j$, and the norm is $|v|_g^2 = g_{ij}(x)v^iv^j$. We remark that a Riemannian metric induces a metric on the cotangent bundle $T^*M = \cup_{q\in M} T_q^*M$ defined as follows: if $\zeta, \eta \in T_q^*M$, then

$$\langle \zeta, \eta\rangle_g = g^{ij}(x(q))\eta_i\zeta_j,$$

where $G^{-1} = (g^{ij})_{ij}$ is the inverse matrix of G. If a metric g_{ij} is defined on an open set Ω in $\mathbb{R}^n$ and $u \in BV(\Omega, \mathbb{R}/2\pi\mathbb{Z})$, the Riemannian gradient is the vector

$$\nabla_g u = G^{-1}\nabla u,$$

and its norm in the metric (g_{ij}) is

$$|\nabla_g u|_g = (g^{ij}\partial_i u \partial_j u)^{1/2}.$$

Analogously, if ν_u is the normal to the set $S(u)$, defined at the end of Definition 2.1, the normal vector with respect to the metric g is

$$\nu_g = G^{-1}\nu_u, \tag{2.1}$$

and its norm is $|\nu_g|_g = (g^{ij}(\nu_u)_i(\nu_u)_j)^{1/2}$ (see [7]).

Finally, we prove a duality relation between the norm on the tangent space and the cotangent.

PROPOSITION 2.5. *Let $v \in \mathbb{R}^n$, and let us call $v_g = G^{-1}v$, as in the definition of the Riemannian gradient or Riemannian normal vector. Then*

$$(|v_g|_g)^p = c_{np}\int_{\mathbb{R}^n} e^{-|\eta|_g}\frac{|\langle v, \eta\rangle|^p}{|\eta|_g^p}\sqrt{g}d\eta \tag{2.2}$$

and

$$\langle v_g, w_g\rangle_g = c_{n2}\int_{\mathbb{R}^n} e^{-|\eta|_g}\frac{\langle v, \eta\rangle\langle w, \eta\rangle}{|\eta|_g^2}\sqrt{g}d\eta \tag{2.3}$$

for suitable constants c_{np}, depending on the dimension of the space and p.

Proof. We fix a vector w of Euclidean length 1 and note that

$$\int_{\mathbb{R}^n} e^{-|\xi|}\frac{|\langle w, \xi\rangle|^p}{|\xi|^p}d\xi = \frac{1}{c_{np}}$$

is a constant independent of w. Denoting $A = (a_{ij})_{ij}$ the square root of G, with the change of variable $\xi = \eta A$ we have $\sum_s(\xi_s)^2 = \eta_k\eta_h g_{kh} = |\eta|_g^2$. Then the second member of (2.2) can be computed:

$$\int_{\mathbb{R}^n} e^{-|\eta|_g}\frac{|\langle v, \eta\rangle|^p}{|\eta|_g^p}\sqrt{g}d\eta = \int_{\mathbb{R}^n} e^{-|\xi|}\frac{|\langle vA^{-1}, \xi\rangle|^p}{|\xi|^p}d\xi$$

$$= \frac{1}{c_{np}} |vA^{-1}|^p = \frac{1}{c_{np}} |v_g|_g^p.$$

The first assertion is proved.

In order to prove the second one, we first note that

$$\delta_{ij} = c_{n2} \int_{\mathbb{R}^n} e^{-|\xi|} \frac{\xi_i \xi_j}{|\xi|^2} d\xi,$$

where δ_{ij} is the Kronecker delta. On the other hand, if we denote $A^{-1} = (a^{ij})_{ij}$,

$$\langle v_g, w_g \rangle_g = g^{hk} v_h w_k = a^{hi} \delta_{ij} a^{jk} v_h w_k = c_{n2} \int_{\mathbb{R}^n} e^{-|\xi|} \frac{v_h a^{hi} \xi_i \xi_j a^{jk} w_k}{|\xi|^2} d\xi$$

$$= c_{n2} \int_{\mathbb{R}^n} e^{-|\xi|} \frac{\langle \xi A^{-1}, v \rangle \langle \xi A^{-1}, w \rangle}{|\xi|^2} d\xi.$$

Then, with the same change of variable as before, $\eta = \xi A^{-1}$, we get the thesis. ∎

3. Γ-convergence results: Proof of Theorem 1.1. In this section we first recover formally the expression of the Euler–Lagrange functionals F_ϵ; then we prove the Γ-convergence of the family F_ϵ to the Mumford and Shah functional

$$MS(u, \mathbb{R}^n) = \begin{cases} \alpha\, c_{np} \int_{\mathbb{R}^n} |\nabla_g u|_g^p \frac{h(x)}{\sqrt{g(x)}} dx + \beta\, c_{n1} \int_{S(u)} |\nu_g|_g \frac{h(x)}{\sqrt{g(x)}} dH^{n-1} & \text{if } u \in SBV, \\ \\ +\infty & \text{otherwise.} \end{cases}$$

The proof of Theorem 1.1 is based on the slicing method, a general integral-geometric technique which allows us to represent the functional $F_\epsilon(u)$ in terms of its one-dimensional sections. In this way it is possible to reduce the dimension of the problem to one and recover the Γ-limit result through the study of the one-dimensional problem. The method we use is a combination of the techniques in [28] and [10], where similar convergence results are provided.

3.1. An approximating family of discrete functionals. Let us first formally write the expression of the Euler–Lagrange functional for (1.7), giving the definition of the space where the problem will be studied.

The equation is defined in terms of a metric $(g_{ij})_{ij}$ such that g_{ij} are continuous functions on R^n and that there are two positive constants λ and Λ such that

$$\lambda |\eta|^2 \le g_{ij}(x) \eta^i \eta^j \le \Lambda |\eta|^2 \quad \forall x, \eta \in \mathbb{R}^n. \tag{3.1}$$

Let us call $h : \mathbb{R}^n \to \mathbb{R}$ a continuous function such that

$$\lambda \le h(x) \le \Lambda \quad \forall x \in \mathbb{R}^n.$$

Let us recall here the assumptions required in Theorem 1.1. The function $\phi : \mathbb{R} \to \mathbb{R}$ is continuous, it is odd, $\phi > 0$ in $[0, \pi[$, and $\phi = 0$ on $[\pi, +\infty[$. If $\varphi : \mathbb{R} \to \mathbb{R}$ is a primitive function of ϕ null in 0, φ is obviously of class $C^1([0, +\infty[)$ and constantly assumes a value β in $[\pi, +\infty[$. Moreover, we require that (1.9) holds. A primitive φ_ϵ of the rescaled function ϕ_ϵ defined in (1.4) is $\varphi_\epsilon(t) = \frac{1}{\epsilon} \varphi(\epsilon^{1/p} t)$.

We consider the following functional:

$$F_\epsilon : L^p \to \mathbb{R} \quad F_\epsilon(u) = \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} e^{-|\eta|_g}\varphi_{\epsilon|\eta|_g}\big(D^\epsilon_{g\eta}u(x)\big)h(x)dx\,d\eta, \tag{3.2}$$

where $D^\epsilon_{g\eta}u$ is defined in (1.6).

REMARK 3.1. *$F_\epsilon(u) < +\infty$ for every $u \in L^p$. Indeed, by the assumption* (1.9) *on φ there exists $\delta > 0$ such that*

$$\varphi(z) \le c_1 z^p \quad \forall z \in [0,\delta] \tag{3.3}$$

for a suitable constant c_1. Here and in what follows we will denote c_i any constant depending only on the data of the problem. On the other hand, since ϕ is nonnegative, φ is increasing and takes its maximum at π. It then follows that

$$\varphi(z) \le \beta \le c_1 z^p \quad \forall z \ge \delta. \tag{3.4}$$

Analogous inequalities hold for φ_ϵ, with the same constant, so that

$$F_\epsilon(u) \le c_\epsilon \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} e^{-|\eta|_g}\big|D^\epsilon_{g\eta}u(x)\big|^p dx\,d\eta \le c_{\epsilon p}\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} e^{-|\eta|_g}\big|u(x)\big|^p dx\,d\eta,$$

and this is finite if $u \in L^p(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z})$. In particular, due to (3.3) *and* (3.4) *we also have the following: there exist positive constants c_1, c_2 such that*

$$c_1 \min\{\alpha z^p, \beta\} \le \varphi(z) \le c_2 \min\{\alpha z^p, \beta\}. \tag{3.5}$$

In order to recognize that F_ϵ is the Euler–Lagrange functional of the discrete phase equation, we will work in the following set of piecewise constant functions:

$$PC^p_\epsilon = \{u \in L^p(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z}) : u \text{ is constant on the cube } \epsilon z + [0,\epsilon]^n \ \forall z \in \mathbb{Z}^n\}.$$

PROPOSITION 3.1. *Let $\epsilon > 0$. Then we have the following:*

(i) *for every $u \in PC^p_\epsilon$ the gradient of F_ϵ in u is given by*

$$(\nabla F_\epsilon(u))(x) = -\int_{\mathbb{R}^n} D^{-\epsilon}_\eta \left(h e^{-|\eta|_g}\frac{|\eta|}{|\eta|_g}\phi_{\epsilon|\eta|_g}\big(D^\epsilon_{g\eta}u\big)\right)(x)d\eta,$$

where we simply denote D^ϵ_η the difference quotient when the metric g is the Euclidean metric;

(ii) *∇F_ϵ is a Lipschitz continuous function on PC^p_ϵ.*

Proof. In order to prove (i) we calculate the Gâteaux derivative along a direction $v \in PC^{\frac{p}{p-1}}_\epsilon$:

$$\lim_{\delta\to 0}\frac{F_\epsilon(u+\delta v) - F_\epsilon(u)}{\delta}$$

$$= \lim_{\delta\to 0}\frac{1}{\delta}\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} h(x)\, e^{-|\eta|_g}\Big(\varphi_{\epsilon|\eta|_g}\big(D^\epsilon_{g\eta}u(x) + \delta D^\epsilon_{g\eta}v(x)\big) - \varphi_{\epsilon|\eta|_g}\big(D^\epsilon_{g\eta}u(x)\big)\Big)dx\,d\eta$$

$$= \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} h e^{-|\eta|_g}\phi_{\epsilon|\eta|_g}\big(D^\epsilon_{g\eta}u\big)\frac{|\eta|}{|\eta|_g}D^\epsilon_\eta v\,dx\,d\eta$$

formally integrating by parts the difference quotient

$$= -\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} D^{-\epsilon}_\eta\Big(h(x)e^{-|\eta|_g}\frac{|\eta|}{|\eta|_g}\phi_{\epsilon|\eta|_g}\big(D^\epsilon_{g\eta}u\big)\Big)(x)v(x)\,dx\,d\eta.$$

Finally, ∇F_ϵ is Lipschitz continuous because it is compositions of Lipschitz continuous functions. ∎

3.2. The one-dimensional case. Let us start with studying the simplest operator of the form (1.8) in $\mathbb{R}$:

$$\int_{\mathbb{R}} f(x)\varphi_{\epsilon|\eta|_g}\left(D^{\epsilon}_{g\eta}u(x)\right)dx, \tag{3.6}$$

where $\eta \in \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}$, is a continuous function such that

$$\lambda \le f(x) \le \Lambda \quad \forall x \in \mathbb{R},$$

with λ, Λ positive constants. A metric in $\mathbb{R}$ is simply defined by a continuous function b such that for every η,

$$|\eta|_{g(x)} = b(x)|\eta|.$$

Moreover, by simplicity in dimension 1 we will always assume that $\eta = 1$ so that the functional on an interval I reduces to

$$\hat{F}_{\epsilon,1,f,b}(u, I) = \int_I f(x)\varphi_{\epsilon b(x)}\Big(\frac{D^{\epsilon}u(x)}{b(x)}\Big)dx, \tag{3.7}$$

where $D^{\epsilon} = D^{\epsilon}_1$ is the difference quotient with respect to the Euclidean metric.

We will give sufficient conditions for the Γ-convergence of the functional $\hat{F}_{\epsilon,1,f,b}(\cdot, I)$ to the Mumford and Shah functional

$$MS_{f,b}(u, I) = \begin{cases} \alpha \int_I f(x)\left(\frac{|u'(x)|}{b(x)}\right)^p dx + \beta \int_{I\cap S(u)} \frac{f(x)}{b(x)} dH^0(x) \\ \qquad\qquad \text{if } \in SBV(I), \\ \\ \qquad +\infty \quad \text{otherwise,} \end{cases} \tag{3.8}$$

where α is defined in (1.9) and $\beta = \varphi(\pi)$.

We recall the following regularity result, which is proved, for example, in Theorem 2.6 in [10].

THEOREM 3.2. *The functional $MS_{f,b}(u, I)$ is lower semicontinuous in $L^1_{loc}(I)$.*

In order to prove the Γ-convergence result in $L^1_{loc}(\mathbb{R}, \mathbb{R}/2\pi\mathbb{Z})$, we need an approximation lemma for sequences converging in $L^1_{loc}(\mathbb{R}, \mathbb{R}/2\pi\mathbb{Z})$; see [10].

LEMMA 3.3. *Let $u_\epsilon \to u$ in $L^1_{loc}(\mathbb{R}, \mathbb{R}/2\pi\mathbb{Z})$. We call $T^{\epsilon}_y v(x)$ a function whose values on the interval $[y+\epsilon(k, k+1)]$, $k \in \mathbb{Z}$, are between $u_\epsilon(y+k\epsilon)$ and $u_\epsilon(y+(k+1)\epsilon)$. Then, for almost every $y \in (0, \epsilon)$ and all choices of functions $T^{\epsilon}_y v(x)$, the family $T^{\epsilon}_y v(x)$ converges to u in $L^1_{loc}(\mathbb{R})$.*

LEMMA 3.4. *Let us first assume that there are two positive constants $\tilde{\alpha}$ and $\tilde{\beta}$ such that*

$$\varphi(z) = \min\left\{\tilde{\alpha}z^p, \tilde{\beta}\right\}. \tag{3.9}$$

Then for every $u \in L^1_{loc}(\mathbb{R}, \mathbb{R}/2\pi\mathbb{Z})$, for every sequence $u_j \to u$ in $L^1_{loc}(\mathbb{R}, \mathbb{R}/2\pi\mathbb{Z})$ there exists a sequence $\epsilon_j \to 0$ such that

$$\liminf_{j\to+\infty} \hat{F}_{\epsilon_j,1,f,b}(u_j, \mathbb{R}) \ge MS_{f,b}(u, \mathbb{R}).$$

Proof. By simplicity of notation in the proof we will always denote $MS(u, I)$ instead of $MS_{f,b}(u, I)$, $\hat{F}_\epsilon(u, I)$ instead of $\hat{F}_{\epsilon,1,f,b}(u, I)$, and D^ϵ instead of D^ϵ_1.

We will call $s \in \mathbb{R}$ such that φ is constant in $[s, +\infty[$.

First we assume that I is a bounded open interval of $\mathbb{R}$. Let $u_j \to u$ in $L^1(I, \mathbb{R}/2\pi\mathbb{Z})$ and show that

$$\liminf_j \hat{F}_{\epsilon_j}(u_j, I) \geq MS(u, I).$$

Let $\delta > 0$ be fixed. Arguing as in Braides [10, p. 82], we can assume that there exists a subsequence, always denoted ϵ_j, and a sequence (y_j), with $y_j \in (0, \epsilon_j)$ satisfying the thesis of Lemma 3.3 such that

$$\hat{F}_{\epsilon_j}(u, I) + \delta \geq \epsilon_j \sum_{k \in J_j} f(k\epsilon_j + y_j)\varphi_{\epsilon_j b}\Big(\frac{D^{\epsilon_j} u(k\epsilon_j + y_j)}{b(k\epsilon_j + y_j)}\Big),$$

where we have denoted

$$J_j = \{k \in \mathbb{Z} :]\epsilon_j k + y_j, \epsilon_j(k+1) + y_j[\subset I\}.$$

This is a particular version of the mean value theorem for integrals, where we have only one inequality, since we are not free to choose y_j in an arbitrary way but only almost everywhere.

Since J_j is finite, we can write

$$J_j = \{k_1^j, \dots, k_{N_j}^j\}$$

and denote

$$J_j^1 = \left\{ k \in J_j : \frac{|(u_j((k+1)\epsilon_j + y_j) - u_j(k\epsilon_j + y_j))\, mod\, 2\pi|}{\epsilon_j b(k\epsilon_j + y_j)} \leq s\epsilon_j^{-1/p} \right\}, \quad J_j^2 = J_j \backslash J_j^1.$$

Then we define $v_j = T_y^{\epsilon_j} u_j$ as follows:

$$\begin{cases} \Big(\frac{t-y_j}{\epsilon_j} - k\Big)u_j(\epsilon_j(k+1) + y_j) + \Big((k+1) - \frac{t-y_j}{\epsilon_j}\Big)u_j(k\epsilon_j + y_j), & t \in\ y_j + \epsilon_j]k, k+1[,\ k \in J_j^1, \\ u_j(k\epsilon_j + y_j), & t \in\ y_j + \epsilon_j]k, k+1[,\ k \in J_j^2, \\ u_j(k_0^1\epsilon_j + y_j) & \text{if } t \leq y_j + k_1^j\epsilon_j, \\ u_j((k_{N_j}^k + 1)\epsilon_j + y_j) & \text{if } t \geq y_j + (k_{N_j}^j + 1)\epsilon_j. \end{cases}$$

The choice of y_j is made, according to Lemma 3.3, in such a way that $v_j \to u$ in $L^1(I)$.

With this notation the estimate of $\hat{F}_\epsilon$ becomes

$$\hat{F}_{\epsilon_j}(u, I) + \delta \geq \epsilon_j \sum_{k \in J_j} f(k\epsilon_j + y_j)\varphi_{\epsilon_j b}\Big(\frac{D^{\epsilon_j} u(k\epsilon_j + y_j)}{b(k\epsilon_j + y_j)}\Big)$$

$$= \tilde{\alpha} \sum_{k \in J_j^1} \epsilon_j f(k\epsilon_j + y_j)\Big|\frac{D^{\epsilon_j} u(k\epsilon_j + y_j)}{b(k\epsilon_j + y_j)}\Big|^p + \tilde{\beta} \sum_{k \in J_j^2} \frac{f(k\epsilon_j + y_j)}{b(k\epsilon_j + y_j)}$$

$$= \tilde{\alpha} \int_I f(x) \left|\frac{v_j'(x)}{b(x)}\right|^p dx + \tilde{\beta} \sum_{x \in S(v_j) \cap I} \frac{f(x)}{b(x)}.$$

The sequence v_j converges to u by its choice. On the other hand, the operator MS is lower semicontinuous so that

$$\lim_{j\to+\infty}\inf \hat{F}_{\epsilon_j}(u_j, I) \geq MS(u, I) - \delta.$$

The arbitrariness of $\delta > 0$ gives the thesis in the case where I is a bounded open interval. The result is still valid for $\mathbb{R}$ approximating from the interior by bounded and open interval I.

In order to deal with the general case, we recall the following theorem about supremum of family of positive measures, which can be found in [10].

PROPOSITION 3.5. *Let Ω be an open set and $A(\Omega)$ be the family of its open subsets. Let $\mu_1 : A(\Omega) \to [0, +\infty[$ be an open set function, superadditive on open sets with disjoint compact closures. Let μ be a positive measure, let ψ_i be positive Borel functions such that $\mu_1(A) \geq \int_A \psi_i d\mu$ for all $A \in A(\Omega)$, and let $\psi(x) = \sup \psi_i(x)$. Then $\mu_1(A) \geq \int_A \psi d\mu$ for all $A \in A(\Omega)$.* □

THEOREM 3.6. *Let ϕ and φ satisfy the assumptions stated in Theorem* 1.1. *Then for every $u \in L^1_{loc}(\mathbb{R}, \mathbb{R}/2\pi\mathbb{Z})$, for every sequence $u_j \to u$ in $L^1_{loc}(\mathbb{R}, \mathbb{R}/2\pi\mathbb{Z})$ there exists a sequence $\epsilon_j \to 0$ such that*

$$\lim_{j\to+\infty}\inf \hat{F}_{\epsilon_j,1,f,b}(u_j, \mathbb{R}) \geq MS_{f,b}(u, \mathbb{R}).$$

Proof. Let a_i and b_i be sequences of positive real numbers such that $\sup_i a_i = \alpha$, $\sup_i b_i = \beta$, and

$$\varphi_i(z) = \min\left\{a_i z^p b_i\right\} \leq \varphi(z) \quad \forall\, t \geq 0 \tag{3.10}$$

by Remark 3.1. Note that we do not require any monotonicity property on a_i and b_i so that their existence is ensured. From Lemma 3.4 we have

$$\lim_{j\to+\infty}\inf \hat{F}_{\epsilon_j,1,f,b}(u_j, \mathbb{R}) \geq a_i \int_I f(x) \left(\frac{|u'(x)|}{b(x)}\right)^p dx + b_i \int_{I\cap S(u)} \frac{f(x)}{b(x)} dH^0(x)$$

for every i. In order to apply Proposition 3.5 we set

$$\mu = \mathcal{L}_1 + \sum_{x\in S(u)} \delta_x,$$

where $\mathcal{L}_1$ is the Lebesgue measure and δ_x is the Dirac measure. We also set

$$\psi_i(x) = \begin{cases} a_i f(x) \left(\frac{|u'(x)|}{b(x)}\right)^p & \text{on } I\backslash S(u), \\ \\ b_i \frac{f(x)}{b(x)} & \text{on } S(u) \end{cases}$$

so that

$$\psi(x) = \sup \psi_i(x) = \begin{cases} \alpha f(x) \left(\frac{|u'(x)|}{b(x)}\right)^p & \text{on } I\backslash S(u), \\ \\ \beta \frac{f(x)}{b(x)} & \text{on } S(u). \end{cases}$$

By Proposition 3.5 we deduce

$$\lim_{j\to+\infty}\inf \hat{F}_{\epsilon_j,1,f,b}(u_j,\mathbb{R}) \geq \alpha\int_I f(x)\left(\frac{|u'(x)|}{b(x)}\right)^p dx + \beta\int_{I\cap S(u)}\frac{f(x)}{b(x)}dH^0(x).$$

This is the thesis. □

The opposite inequality is simpler. We start with a simple remark.

REMARK 3.2. *Let I be a real interval, not necessarily bounded, and let $u \in BV(\mathbb{R},\mathbb{R}/2\pi\mathbb{Z})$ such that $MS_{f,b}(u,I) < +\infty$. Then there exists a constant c_1 independent of ϵ such that for every*

$$A \subset \{x \in I : [x, x+\epsilon] \cap S(u) \neq \emptyset\}, \tag{3.11}$$

$$\int_A |D^\epsilon u(x)|^p dx \leq c_1 \int_{\tilde{A}_\epsilon} |u'(x)|^p dx, \tag{3.12}$$

where $\tilde{A}_\epsilon = \cup_{x\in A}[x, x+\epsilon]$.

Indeed,

$$\int_A |D^\epsilon u(x)|^p dx = \int_A \left|\int_0^1 u'(x+\epsilon s)ds\right|^p dx \leq c_1 \int_A\int_0^1 |u'(x+\epsilon s)|^p ds dx$$

(with the change of variable $y = x+\epsilon s$)

$$\leq c_1\int_0^1\int_{\tilde{A}_\epsilon}|u'(x)|^p dx ds = c_1\int_{\tilde{A}_\epsilon}|u'(x)|^p dx.$$

THEOREM 3.7. *Let ϕ and φ satisfy the assumptions stated in Theorem* 1.1. *Then for every $u \in L^1_{loc}(\mathbb{R},\mathbb{R}/2\pi\mathbb{Z})$,*

$$\limsup_{\epsilon\to 0^+} \hat{F}_{\epsilon,1,f,b}(u,\mathbb{R}) \leq MS_{f,b}(u,\mathbb{R}).$$

Proof. Let us fix $\delta > 0$. We can obviously assume that $MS_{f,b}(u,\mathbb{R}) < +\infty$, which implies that u has only a finite number of jumps. Then there exists $M > 0$ such that u has no jumps in $I\backslash[-M,M]$. Since $u' \in L^p$, by the previous remark we can also assume that M is chosen in such a way that for every ϵ,

$$\int_{I\backslash[-M,M]}(|D^\epsilon u|^p + |u'|^p)dx \leq \delta. \tag{3.13}$$

From the previous remark it also follows that there exists $\sigma > 0$ independent of ϵ such that, for every ϵ, for every A satisfying (3.11), $A \subset [-M,M]$ and with Lebesgue measure $|A| < \sigma$, the following estimate holds:

$$\int_A (|D^\epsilon u|^p + |u'|^p)dx \leq \delta. \tag{3.14}$$

In particular, if we call

$$I_{k\epsilon} = \{x \in [-M,M] : [x, x+\epsilon] \cap S(u) \neq \emptyset, |D_\epsilon u(x)| > k\},$$

then, always for the previous remark,

$$|I_{k\epsilon}| \le \frac{1}{k^p}\int_{I_{k\epsilon}} |D^\epsilon u|^p dx \le \frac{1}{k^p}\int_{I\setminus S(u)} |u'(x)|^p dx \to 0$$

as $k \to +\infty$, uniformly in ϵ. Then by (3.14) we can fix $k > 0$ such that for every ϵ,

$$\int_{I_{k\epsilon}} |D^\epsilon u|^p dx \le \delta. \tag{3.15}$$

Let us denote $\{x_1, \dots, x_s\}$ the set of jumps of u, and let us call

$$J = \{x \in [-M, M] : [x, x+\epsilon] \cap S(u) = \Omega, |D_\epsilon u(x)| \le k\}, \quad I_S = \bigcup_j [x_j - \epsilon, x_j].$$

By (3.15) and (3.13) and the fact that $\varphi_\epsilon(z) \le c_2 z^p$ for every $z \in \mathbb{R}$, with c_2 independent of ϵ, the discrete functional can be estimated as

$$\hat{F}_{\epsilon,1,f,b}(u, I) = 2c_2\delta + \sum_j \int_{x_j-\epsilon}^{x_j} f(x)\varphi_{\epsilon b}\Big(\frac{D^\epsilon u(x)}{b(x)}\Big)dx \tag{3.16}$$

$$+ \int_J f(x)\varphi_{\epsilon b}\Big(\frac{D^\epsilon u(x)}{b(x)}\Big)dx.$$

Each of the integrals in the first sum can be estimated using the definition of φ_ϵ and the fact that $\max\varphi = \beta$:

$$\int_{x_j-\epsilon}^{x_j} f(x)\varphi_{\epsilon b}\Big(\frac{D^\epsilon u(x)}{b(x)}\Big)dx \le \frac{\beta}{\epsilon}\int_{x_j-\epsilon}^{x_j} \frac{f(x)}{b(x)}dx \to \beta\frac{f(x_j)}{b(x_j)} \tag{3.17}$$

as ϵ tends to 0.

In the last integral of (3.16) we use the fact that $D^\epsilon u(x)$ takes values in the compact set $[-k, k]$ and punctually tends to u', while $\varphi_{\epsilon b(x)}(z) \to \alpha\frac{|z|^p}{b^p(x)}$ uniformly if (x, z) belong to a compact set. Hence

$$\varphi_{\epsilon b}\Big(\frac{D^\epsilon u(x)}{b(x)}\Big) \to \alpha\,\frac{|u'(x)|^p}{b^p(x)}$$

almost everywhere. Using again the fact that $D^\epsilon u(x)$ is bounded by k we can apply Lebesgue's dominate convergence theorem on the bounded set $[-M, M]\setminus S(u)$ and obtain

$$\int_J f(x)\varphi_{\epsilon b}\Big(\frac{D^\epsilon u(x)}{b(x)}\Big)dx \to \alpha\int_{[-M,M]\setminus S(u)} \frac{|u'(x)|^p}{b^p(x)}dx. \tag{3.18}$$

Putting together (3.16), (3.17), and (3.18) we obtain

$$\limsup_{\epsilon\to 0^+} \hat{F}_{\epsilon,1,f,b}(u, \mathbb{R}) \le 2c_2\delta + MS_{f,b}(u, \mathbb{R}),$$

and this implies the thesis, since δ is arbitrary. □

Finally, from Lemma 3.4, Theorem 3.6, and (3.5) we have the following corollary.

COROLLARY 3.8. *Let ϕ and φ satisfy the assumptions stated in Theorem* 1.1. *Then*

$$\Gamma - \lim_{\epsilon\to 0} \hat{F}_{\epsilon,1,f,b}(u,\mathbb{R}) = MS_{f,b}(u,\mathbb{R}) \quad in \quad L^1_{loc}(\mathbb{R},\mathbb{R}/2\pi\mathbb{Z}),$$

$$\lim_{\epsilon\to 0}\hat{F}_{\epsilon,1,f,b}(u,\mathbb{R}) = MS_{f,b}(u,\mathbb{R}) \quad for \ \ every \ \ u \in SBV(\mathbb{R},\mathbb{R}/2\pi\mathbb{Z}),$$

and

$$\hat{F}_{\epsilon,1,f,b}(u,\mathbb{R}) \le C\ MS_{f,b}(u,\mathbb{R}) \quad for \ \ every \ \ u \in L^1(\mathbb{R},\mathbb{R}/2\pi\mathbb{Z}),$$

with C a positive constant. □

3.3. The *n*-dimensional case. In this section we will deduce the general n-dimensional case from the one-dimensional result, using the slicing method already used in the nonperiodic, isotropic case by Braides [10].

This procedure is formally similar to a standard reduction in the integral so that we fix $\eta \in \mathbb{R}^n\backslash\{0\}$ and denote $\langle\eta\rangle^\perp = \{z \in \mathbb{R}^n : \langle\eta, z\rangle = 0\}$ the orthogonal space to η with respect to the Euclidean metrics. For every $y \in \langle\eta\rangle^\perp$ consider the function $u_{\eta y}$ defined by

$$u_{\eta y}(t) = u\Big(y + t\frac{\eta}{|\eta|}\Big), \quad t \in \mathbb{R}.$$

With these notations the operator F_ϵ defined in (3.2) becomes

$$F_\epsilon(u,\mathbb{R}^n) = \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} h(x)e^{-|\eta|_g}\varphi_{\epsilon|\eta|_g}\Big(D^\epsilon_{g\eta}u(x)\Big)dxd\eta \tag{3.19}$$

$$= \int_{\mathbb{R}^n}\int_{\langle\eta\rangle^\perp} \hat{F}_{\epsilon,\eta,f,b}(u_{\eta y},\mathbb{R})dyd\eta,$$

where

$$f(t) = \Big(he^{-|\eta|_g}\frac{|\eta|}{|\eta|_g}\Big)_{\eta y}(t), \qquad b(t) = \frac{|\eta|_g(y + t\frac{\eta}{|\eta|})}{|\eta|}, \tag{3.20}$$

and

$$\hat{F}_{\epsilon,\eta,f,b}(u_{\eta y}(t),\mathbb{R}) = \int_{\mathbb{R}} f(t)\varphi_{\epsilon\eta}\Big(\frac{D^\epsilon_{|\eta|}u_{\eta y}(t)}{b(t)}\Big)dt.$$

In this way the functional F_ϵ is represented in terms of one-dimensional sections.

Also the functional MS, defined in (1.10), can be represented in terms of its sections, and the function u belongs to BV if and only if its sections $u_{\eta y}$ belong to $BV(\mathbb{R})$.

THEOREM 3.9. (i) *Let $u \in SBV(\mathbb{R}^n,\mathbb{R}/2\pi\mathbb{Z})$. Then for all $\eta \in \mathbb{R}^n$ we have $u_{\eta y} \in SBV(\mathbb{R},\mathbb{R}/2\pi\mathbb{Z})$ for almost everywhere $y \in \langle\eta\rangle^\perp$ and, moreover,*

$$u'_{\eta y}(t) = \Big\langle \nabla u\Big(y + t\frac{\eta}{|\eta|}\Big), \frac{\eta}{|\eta|}\Big\rangle \quad for\ a.e.\ t \in \mathbb{R},$$

$$S(u_{\eta y}) = \left\{ t \in \mathbb{R} : y + t\frac{\eta}{|\eta|} \in S(u) \right\}, \quad u^+_{\eta y}(t) = u^+\Big(y + t\frac{\eta}{|\eta|}\Big), \quad u^-_{\eta y}(t) = u^-\Big(y + t\frac{\eta}{|\eta|}\Big),$$

where u^+ and u^- are defined at the end of Definition 2.1.

(ii) *Let $u \in L^1_{loc}(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z})$, and let $MS_{f,b}$ be the operator defined in* (3.8). *If*

$$\int_{\langle \eta \rangle^\perp} MS_{f,b}(u_{\eta y}, \mathbb{R}) dy < +\infty$$

for every $\eta \in B$, B a basis of vector space $\mathbb{R}^n$, then $u \in SBV(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z})$.

We refer to Ambrosio [1] for the proof.

Applying the previous result we get the expression of our Mumford and Shah functional.

THEOREM 3.10. *For every function $u \in L^1_{loc}(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z})$ we have that*

$$\int_{\mathbb{R}^n} \left(\int_{\langle \eta \rangle^\perp} MS_{f,b}(u_{\eta y}, \mathbb{R}) dy \right) d\eta = MS(u, \mathbb{R}^n).$$

Proof. We can assume that $u \in SBV(\mathbb{R}^n)$. In this case by Theorem 3.9 we have that

$$\int_{\langle \eta \rangle^\perp} MS_{f,b}(u_{\eta y}, \mathbb{R}) dy$$

$$= \alpha \int_{\langle \eta \rangle^\perp} \int_{\mathbb{R}} f(t) \Big(\frac{|u'_{\eta y}(t)|}{b(t)} \Big)^p dt dy + \beta \int_{\langle \eta \rangle^\perp} \int_{S(u_{\eta y})} \frac{f(t)}{b(t)} dH^0(t) dy$$

by Theorem 3.9 and by definition (3.20)

$$= \alpha \int_{\mathbb{R}^n} h(x) e^{-|\eta|_g} \Big| \Big\langle \nabla u(x), \frac{\eta}{|\eta|_g} \Big\rangle \Big|^p dx + \beta \int_{S(u)} h(x) e^{-|\eta|_g} \left| \Big\langle \nu, \frac{\eta}{|\eta|_g} \Big\rangle \right| dH^{n-1}(x),$$

where the equality follows from [10] for the second integral. Integrating in η the preceding equality we get

$$\int_{\mathbb{R}^n} \int_{\langle \eta \rangle^\perp} MS_{f,b}(u_{\eta y}, \mathbb{R}) dy d\eta$$

$$= \alpha \int_{\mathbb{R}^n} h(x) \left(\int_{\mathbb{R}^n} e^{-|\eta|_g} \Big| \Big\langle \nabla u(x), \frac{\eta}{|\eta|_g} \Big\rangle \Big|^p d\eta \right) dx$$

$$+ \beta \int_{S(u)} h(x) \left(\int_{\mathbb{R}^n} e^{-|\eta|_g} \left| \Big\langle \nu, \frac{\eta}{|\eta|_g} \Big\rangle \right| d\eta \right) dH^{n-1}(x)$$

(by Proposition 2.5)

$$= \alpha c_{np} \int_{\mathbb{R}^n} \frac{h(x)}{\sqrt{g(x)}} |\nabla_g u(x)|^p_g dx + \beta\, c_{n1} \int_{S(u)} \frac{h(x)}{\sqrt{g(x)}} |\nu_g|_g dH^{n-1}(x). \qquad \square$$

Proof of Theorem 1.1. We sketch the proof which follows from the convergence Corollary 3.8 and the representation of the limit functional provided in Theorem 3.10. Let $u, u_j \in L^1_{loc}(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z})$, $u_j \to u$ in $L^1_{loc}(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z})$, let $\epsilon_j \to 0$, and let us prove that

$$\lim_{j\to+\infty} \inf F_{\epsilon_j}(u_j) \geq MS(u, \mathbb{R}^n).$$

Indeed, by (3.19) and the Fatou lemma

$$\lim_{j\to+\infty} \inf F_{\epsilon_j}(u_j) \geq \int_{\mathbb{R}^n} \int_{\langle\eta\rangle^\perp} \lim_{j\to+\infty} \inf \hat{F}_{\epsilon_j,\eta,f,b}((u_j)_{\eta y}, \mathbb{R}) dy d\eta$$

(by Corollary 3.8)

$$\geq \int_{\mathbb{R}^n} \int_{\langle\eta\rangle^\perp} MS_{f,b}((u_j)_{\eta y}, \mathbb{R}) dy d\eta = MS(u, \mathbb{R}^n)$$

by Theorem 3.10. Finally, the dominated convergence asserted in Corollary 3.8 ensures that $MS(u) = \lim_\epsilon F_\epsilon(u)$ for every u, and this proves the second requirement in the definition of Γ-convergence. □

4. Existence of a minimum for the Mumford and Shah functional. We will give here an approximation result of the minimization problem for the Riemannian Mumford and Shah functional. It is based on the existence of the minimum for every F_ϵ, on the Γ-convergence property, and on a suitable compactness result.

4.1. An embedding theorem. In this section we will prove an embedding theorem which extends the classical compactness result in the space BV. Indeed, due to the particular expression of the functional F_ϵ, we will deal with family (u_ϵ) of functions such that the quantity

$$N(u_\epsilon) = \int_\Omega |u_\epsilon| dx + \int_{\mathbb{R}^n} e^{-|\eta|_g} \int_\Omega |D^\epsilon_\eta u_\epsilon(x)| dx\, d\eta \tag{4.1}$$

is bounded if Ω is bounded.

THEOREM 4.1. *Let (u_ϵ) be a family of functions in $L^1_{loc}(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z})$ such that for every bounded set Ω, $N(u_\epsilon)$ is bounded. Then there exists a sequence ϵ_j convergent to 0 and a function u in BV_{loc} such that u_{ϵ_j} converges to u in $L^1_{loc}(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z})$.*

Proof. Let us choose a nonnegative radially symmetric cut off function η of class $C^\infty_0(\mathbb{R}^n)$, supported in the unitary sphere, and with integral 1. For every $\epsilon > 0$ we set

$$u^\epsilon_\epsilon(x) = \int_{\mathbb{R}^n} \eta(\xi) u_\epsilon(x + \epsilon\xi) d\xi.$$

Then we have for Ω bounded

$$\int_\Omega |u^\epsilon_\epsilon(x)| dx \leq \int_\Omega \left(\int_{\mathbb{R}^n} \eta(\xi) u_\epsilon(x + \epsilon\xi) d\xi \right) dx < c_1,$$

since η is bounded in L^∞ and (u_ϵ) in L^1_{loc}. By definition the gradient of (u^ϵ_ϵ) is

$$\nabla u^\epsilon_\epsilon(x) = \frac{1}{\epsilon} \int_{\mathbb{R}^n} \nabla\eta(\xi) u_\epsilon(x + \epsilon\xi) d\xi$$

since η is radially symmetric

$$= \int_{\mathbb{R}^n} \nabla\eta(\xi)\Big(\frac{u_\epsilon(x+\epsilon\xi)-u_\epsilon(x)}{\epsilon}\Big)d\xi.$$

Then for any bounded Ω

$$\int_\Omega |\nabla u_\epsilon^\epsilon(x)|dx \le \int_{|\xi|\le 1}\int_\Omega |D_\xi^\epsilon u_\epsilon(x)|dxd\xi < c_2$$

by the assumption on N_ϵ. By the standard compactness theorem in BV_{loc} it follows that (u_ϵ^ϵ) has a subsequence $(u_{\epsilon_j}^{\epsilon_j})$ converging in L^1_{loc} to a BV_{loc} function u. On the other side,

$$(u_{\epsilon_j}-u_{\epsilon_j}^{\epsilon_j})(x) = \int_{\mathbb{R}^n}\eta(\xi)\Big(u_{\epsilon_j}(x+\epsilon_j\xi)-u_{\epsilon_j}(x)\Big)d\xi \le \epsilon_j\int_{|\xi|\le 1}\Big|D_\xi^{\epsilon_j}u_{\epsilon_j}(x)\Big|d\xi.$$

Integrating over Ω we get

$$\int_\Omega |u_{\epsilon_j}-u_{\epsilon_j}^{\epsilon_j}|(x)dx \le \epsilon_j\int_{|\xi|\le 1}\int_\Omega \Big|D_\xi^{\epsilon_j}u_{\epsilon_j}(x)\Big|dxd\xi \le c_3\epsilon_j.$$

It immediately follows that u_{ϵ_j} has the same limit as $u_{\epsilon_j}^{\epsilon_j}$ in L^1_{loc}. □

4.2. A compactness result. Let us now prove a compactness result for a family (u_ϵ) of functions such that $F_\epsilon(u_\epsilon)$ is bounded. Since the argument of the function φ_ϵ in the expression of F_ϵ is the difference quotient and the functions we are interested in have a different behavior when the argument is small or big, we will also denote the following: $D_{g\xi}^{\epsilon,+}u_\epsilon(x) = D_{g\xi}^\epsilon u_\epsilon(x)$ if $|D_{g\xi}^\epsilon u_\epsilon(x)| > \pi(\epsilon|\xi|)^{-\frac{1}{p}}$ and $D_{g\xi}^{\epsilon,+}u_\epsilon(x) = 0$ otherwise, and we will call

$$I_{\epsilon\xi}^+ = \{x\in\mathbb{R}^n | D_{g\xi}^{\epsilon,+}u_\epsilon(x)\neq 0\}. \tag{4.2}$$

This notation will be useful when studying the limit for $\epsilon\to 0$, since the term $D_{g\xi}^{\epsilon,-}u_\epsilon(x)$ will recover the gradient of u, while $D_{g\xi}^{\epsilon,+}u_\epsilon(x)$ will describe the jump set of the function.

THEOREM 4.2. *Let (u_ϵ) be a family of functions in $L^1_{loc}(\mathbb{R}^n,\mathbb{R}/2\pi\mathbb{Z})$ such that $F_\epsilon(u_\epsilon)$ is bounded; then $N_\epsilon(u_\epsilon)$ is bounded.*

Proof. Let us call c_1 a constant such that

$$\varphi_\epsilon(z)\ge c_1 z^p \tag{4.3}$$

for all z such that $\epsilon^{1/p}z\le\pi$. Note that c_1 is independent of ϵ. Let us now fix an open set $\Omega\subset\subset\mathbb{R}^n$ and estimate separately the integral on $I_{\epsilon\xi}^+$ and the complement set. Since u_ϵ takes values in $[-\pi,\pi]$, we have

$$\int_{\mathbb{R}^n} e^{-|\xi|_g}\int_{I_{\epsilon\xi}^+\cap\Omega}|D_{g\xi}^\epsilon u_\epsilon(x)|dxd\xi \le c_2\int_{\mathbb{R}^n}e^{-|\xi|_g}\int_{I_{\epsilon\xi}^+\cap\Omega}\frac{1}{\epsilon|\xi|_g}dxd\xi \tag{4.4}$$

(since φ takes constantly the value β in $[\pi,+\infty]$)

$$\le \frac{c_2}{\beta}\int_{\mathbb{R}^n}e^{-|\xi|_g}\int_{I_{\epsilon\xi}^+\cap\Omega}\varphi_{\epsilon|\xi|}\big(D_{g\xi}^\epsilon u_\epsilon\big)dxd\xi \le c_3F_\epsilon(u_\epsilon).$$

By condition (4.3) and the assumption (3.1) on $(g_{ij})_{ij}$,

$$\int_{\mathbb{R}^n}\int_{\mathbb{R}^n\setminus I^+_{\epsilon\xi}} e^{-|\xi|}\left|D^\epsilon_{g\xi}u_\epsilon(x)\right|^p dxd\xi \le c_4\int_{\mathbb{R}^n}\int_{\mathbb{R}^n\setminus I^+_{\epsilon\xi}} e^{-|\xi|_g}\varphi_{\epsilon|\xi|}\left(D^\epsilon_{g\xi}u_\epsilon(x)\right)dxd\xi$$

$$\le c_4 F_\epsilon(u_\epsilon).$$

Consequently,

$$\int_{\mathbb{R}^n} e^{-|\xi|_g}\int_\Omega |D^\epsilon_{g\xi}u_\epsilon(x)|dxd\xi = \int_{\mathbb{R}^n} e^{-|\xi|_g}\left(\int_{\Omega\setminus I^+_{\epsilon\xi}} + \int_{\Omega\cap I^+_{\epsilon\xi}}\right)|D^\epsilon_{g\xi}u(x)|dxd\xi$$

by (4.4) and Hölder inequality

$$\le c_3F_\epsilon(u_\epsilon) + \int_{\mathbb{R}^n} e^{-|\xi|_g}\left(\int_{\Omega\setminus I^+_{\epsilon\xi}} |D^\epsilon_{g\xi}u_\epsilon(x)|^p dx + c_5|\Omega|\right)d\xi$$

$$\le (c_3+1)\,F_\epsilon(u_\epsilon) + c_6|\Omega|$$

for suitable constants c_i. Here $|\cdot|$ indicate the Lebesgue measure in $\mathbb{R}^n$.

Then lemma is proved. □

4.3. Approximation of the minima for the Riemannian Mumford and Shah functional. Let us first modify the functional F_ϵ in such a way that its minimum is a BV function.

LEMMA 4.3. *Let $g \in L^q(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z})$ and for every $\epsilon > 0$ let us denote*

$$G_\epsilon(u) = \begin{cases} F_\epsilon(u) + \int_{\mathbb{R}^n}|u-g|^q dx & \text{if } u \in BV(\mathbb{R}^n,\mathbb{R}/2\pi\mathbb{Z}), |Du|(\mathbb{R}^n) \le \frac{1}{\epsilon}, \\ +\infty & \text{otherwise.}\end{cases} \tag{4.5}$$

Then the family $G_\epsilon(u)$ Γ-converges as $\epsilon \to 0$ in $L^1_{loc}(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z})$ to the functional

$$G_0(u) = MS(u,\mathbb{R}^n) + \int_{\mathbb{R}^n}|u-g|^q dx.$$

Proof. The lim inf-inequality follows from the Γ-convergence of (F_ϵ). The lim sup follows from the pointwise convergence of (F_ϵ) if $u \in SBV$ and by a truncation argument for all u. □

Proof of Theorem 1.2. Since the functional G_ϵ is lower semicontinuous in $L^1_{loc}(\mathbb{R}^n)$ and the set

$$\{u \in BV(\mathbb{R}^n,\mathbb{R}/2\pi\mathbb{Z}) : |Du|(\mathbb{R}^n) \le 1/\epsilon\}$$

is compact in $L^1_{loc}(\mathbb{R}^n)$, the existence of minimizers for G_ϵ follows from the direct method of the calculus of variations.

We then prove that all the minimizers belong to the same compact set K. Let (u_ϵ) be a family of minimizers. Since

$$G_\epsilon(u_\epsilon) \le G_\epsilon(0) \le |g|^q_{L^q(\mathbb{R}^n)},$$

we can apply Theorems 4.1 and 4.2 and deduce that the family (u_ϵ) is relatively compact in L^1_{loc} and has a limit in BV.

Finally, by the general property of Γ-convergence stated in Theorem 2.3, any limit point of (u_ϵ) is a minimizer for the problem (1.12) and $m_\epsilon \to m_0$ as $\epsilon \to 0$. □

5. The evolution problem. In this chapter we fix a function $u_0 \in BV$, approximate it by a piecewise constant function, and for every ϵ study the solution (u_ϵ) of problem (1.13) in section 1. Then we establish the properties of the limit of this family as $\epsilon \to 0$.

We now define the space

$$X = \{u \in SBV_{loc}(\mathbb{R}^n, \mathbb{R}/2\pi Z) : MS(u, \mathbb{R}^n) < +\infty\}.$$

Let $u_0 \in X$ be an initial datum for the parabolic problem. Since the functional F_ϵ is defined on piecewise constant functions, we consider an approximation of u_0 in the space $PC^p_\epsilon(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z})$ defined in section 2.

PROPOSITION 5.1. *If $u_0 \in X$, there exists a family $(u_{0\epsilon}) \in PC^p_\epsilon(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z})$ such that*

$$u_{0\epsilon} \to u_0 \quad in \quad L^p_{loc}(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z}),$$

$$\lim_{\epsilon \to 0} F_\epsilon(u_{0\epsilon}) = MS(u_0, \mathbb{R}^n),$$

and

$$\sup_{\epsilon > 0}\{F_\epsilon(u_{0\epsilon})\} < +\infty$$

(*see* [29, *p.* 167] *for the proof*).

Then we consider the evolution problem in (1.13). By the standard Cauchy–Lipschitz existence result (cf. [31]), we have the following theorem.

THEOREM 5.2. *For every $\epsilon > 0$ the initial value problem* (1.13) *has a unique solution $u_\epsilon \in C^1([0, +\infty[, PC^p_\epsilon)$ which depends continuously on the initial datum.*

Let us now study the limit of the family (u_ϵ).

LEMMA 5.3. *Let Ω be a compact set in $\mathbb{R}^n$, and let (u_ϵ) be the family of solutions of the initial value problem found in Theorem* 5.2. *There exists a sequence (ϵ_k) convergent to* 0 *such that (u_{ϵ_k}) is relatively compact in $C([0, +\infty[; L^p_{loc}(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z}))$ and has a limit $u \in C([0, +\infty[; L^p_{loc}(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z}))$ such that $u(t) \in BV(\Omega)$ for every $t \in [0, +\infty[$.*

Proof. We first note that the function $t \to F_\epsilon(u_\epsilon(t))$ is nonincreasing. Indeed,

$$\frac{d}{dt}F_\epsilon(u_\epsilon(t)) = \langle \nabla F_\epsilon(u_\epsilon(t)), u'_\epsilon(t)\rangle_{L^2(\mathbb{R}^n)} \tag{5.1}$$

$$= -||u'_\epsilon(t)||_{L^2(\mathbb{R}^n)} = -||\nabla F_\epsilon(u_\epsilon(t))||_{L^2(\mathbb{R}^n)}.$$

This implies that

$$F_\epsilon(u_\epsilon(t)) \le F_\epsilon(u_{0\epsilon}) \le \sup_{\epsilon > 0} F_\epsilon(u_{0\epsilon}) < +\infty$$

by the choice of the family $(u_{0\epsilon})$ in Proposition 5.1. By Theorems 4.1 and 4.2, this implies that the family $(u_\epsilon(t))$ is relatively compact in $L^1_{loc}(\mathbb{R}^n)$, and for every t the limit $u(t)$ belongs to BV.

We have to prove the continuity of this limit. Since the functions (u_ϵ) take values in $[-\pi, \pi]$, the compactness in L^1_{loc} implies compactness in L^p_{loc} for every p. Moreover,

$$||u_\epsilon(t_1) - u_\epsilon(t_2)||_{L^2(\mathbb{R}^n)} \le \int_{t_1}^{t_2} ||u'_\epsilon(t)||_{L^2(\mathbb{R}^n)} dt \le \left(\int_{t_1}^{t_2} ||u'_\epsilon(t)||^2_{L^2(\mathbb{R}^n)} dt\right)^{\frac{1}{2}} |t_1 - t_2|^{\frac{1}{2}}$$

$$\leq (F_\epsilon(u_{0\epsilon}(t)))^{\frac{1}{2}} |t_1 - t_2|^{\frac{1}{2}} \leq c\, |t_1 - t_2|^{\frac{1}{2}}$$

for any $\epsilon > 0$. Letting ϵ go to 0, we obtain the continuity of u. ∎

REMARK 5.1. *Let us note that if*

$$\frac{\phi(z)}{z} \to 2\alpha \quad as\ t \to 0, \tag{5.2}$$

then condition (1.9) *is satisfied with* $p = 2$*, and all the previous results hold true. Moreover, if* ϕ *is of class* C^2*, there exists a constant* c *such that*

$$|\phi_{\epsilon|\xi|}(z) - \alpha z| \leq c\sqrt{\epsilon}(\varphi_{\epsilon|\xi|}(z) + |\xi|^2) \text{ when } |z| \leq \pi(\epsilon|\xi|)^{-\frac{1}{2}}. \tag{5.3}$$

Let us prove the following theorem, where we will assume $p = 2$.

THEOREM 5.4. *Assume as before that* ϕ *is continuous, it is odd,* $\phi > 0$ *in* $[0, \pi[$*,* $\phi = 0$ *on* $[\pi, +\infty[$*, and assume that assumptions* (5.2) *and* (5.3) *are satisfied. If* u_ϵ *is the solution of problem* (1.13) *and* u *its limit, then*

$$\partial_t u_\epsilon \to 2\alpha\, c_{n2}\, \mathrm{div}\left(\frac{g^{ij}}{\sqrt{g}}\partial_j u\right) \textit{ weakly in } \ L^2_{loc}([0, +\infty[\times\mathbb{R}^n, \mathbb{R}).$$

Proof. Let us fix a bounded set Ω. By assumption we have

$$\int_0^T \int_{\mathbb{R}^n} \int_{\Omega \cap I^+_{\epsilon\eta}} \phi_{\epsilon|\eta|_g}\big(D^\epsilon_{g\eta} u_\epsilon(x)\big) dx d\eta dt = 0,$$

where $I^+_{\epsilon\eta}$ is defined in (4.2). If $U \subset\subset \mathbb{R}^n$ is bounded, by (5.3)

$$\int_0^T \int_U \int_{\Omega \setminus I^+_{\epsilon\eta}} \Big|\phi_{\epsilon|\eta|_g}\big(D^\epsilon_{g\eta} u_\epsilon(x)\big) - \alpha D^\epsilon_{g\eta} u_\epsilon(x)\Big| dx d\eta dt \leq \sqrt{\epsilon}(F_\epsilon(u_\epsilon) + c_1) \to 0$$

as $\epsilon \to 0$.

This means that

$$\phi_{\epsilon|\eta|_g}\big(D^\epsilon_{g\eta} u_\epsilon(x)\big) - \alpha D^\epsilon_{g\eta} u_\epsilon(x) \to 0 \text{ in } L^1_{loc}([0,T] \times \mathbb{R}^n \times \Omega) \quad \text{as} \quad \epsilon \to 0. \tag{5.4}$$

On the other side, by Lemma 3.6 in [29]

$$D^\epsilon_\eta u_\epsilon(x) \to \left\langle \nabla u, \frac{\eta}{|\eta|} \right\rangle \text{ weakly} * \text{ in } L^1_{loc}([0,T] \times \mathbb{R}^n \times \Omega)$$

so that

$$\phi_{\epsilon|\eta|_g}\big(D^\epsilon_{g\eta} u_\epsilon(x)\big) \to \left\langle \nabla u, \frac{\eta}{|\eta|} \right\rangle \text{ weakly} * \text{ in } L^1_{loc}([0,T] \times \mathbb{R}^n \times \Omega). \tag{5.5}$$

Now let $\Phi \in C^\infty_0(]0, +\infty[\times\mathbb{R}^n)$. Since u_ϵ is a solution of the evolution equation, we have

$$\int_0^\infty \int_{\mathbb{R}^n} u_\epsilon \frac{\partial \Phi}{\partial t} dx dt$$

$$= \int_0^{+\infty} \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} h\, e^{-|\eta|_g} \phi_{\epsilon|\eta|_g}\big(D^\epsilon_{g\eta} u_\epsilon(x)\big) D^\epsilon_{g\eta} \Phi(x,t) d\eta \right) dx dt$$

by (5.5) and the uniform convergence of $D^{\epsilon}_{g\eta}\Phi$ to $\langle \nabla\Phi, \frac{\eta}{|\eta|_g}\rangle$ as $\epsilon \to 0$

$$\to \alpha \int_0^{+\infty} \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} h\, e^{-|\eta|_g} \left\langle \nabla u, \frac{\eta}{|\eta|_g} \right\rangle \left\langle \nabla\Phi, \frac{\eta}{|\eta|_g} \right\rangle d\eta dx dt$$

by Proposition 2.5

$$= \alpha\, c_{n2} \int_0^{+\infty} \int_{\mathbb{R}^n} \frac{g^{hk}(x)}{\sqrt{g}} \partial_h u \partial_k \Phi dx dt.$$

On the other side, $\partial_t u_\epsilon$ is bounded in $L^2_{loc}([0,+\infty[\times\mathbb{R}^n, \mathbb{R})$, and the thesis is proved. □

COROLLARY 5.5. *Under the assumptions of Theorem* 5.4 *the function* u *belongs to* $C([0,+\infty[; L^2_{loc}(\mathbb{R}^n, \mathbb{R}/2\pi\mathbb{Z}))$ *and satisfies the following:* $u(0) = u_0$, $MS(u(t), \mathbb{R}^n) \leq MS(u_0, \mathbb{R}^n)$, *for every* $t \geq 0$. *Moreover, the function* $u = u(x,t)$ *is a distributional solution in* $]0,+\infty[\times\mathbb{R}^n$ *of the equation*

$$\frac{\partial u}{\partial t} = 2\alpha\, c_{n2} D \left(\frac{g^{ij}}{\sqrt{g}} \nabla u \right),$$

where D *is the distributional* x*-derivative, out of the jump set of* u.

Proof. It is a consequence of the results we have proved on the function u in the previous theorems. □

6. A numerical example. We consider here a simple numerical example showing how the phase equation (1.5) is able to segment an object by reaching phase locking in semantically homogeneous areas of an image and by decoupling phases between object and background. We will consider the figure completion of the well-known square of Kanizsa (Figure 6.1). In this example we consider an image $(x_1, x_2) \to I(x_1, x_2)$ as a real positive function defined in a rectangular domain $\Omega \subset \mathbb{R}^2$. Following [37], we suppose that the image induces a local change of the connectivity e in proximity of its discontinuities in such a way that hypercolumns appear decoupled across the boundaries of a figure. We choose a simple edge indicator as the connectivity function, namely

$$s(x_1, x_2) = \frac{1}{1 + (|\nabla G_\sigma(x_1, x_2) \star I(x_1, x_2)|/c)^2}, \tag{6.1}$$

where

$$G_\sigma(x_1, x_2) = \frac{\exp(-(|(x_1, x_2)|/\sigma)^2)}{\sigma\sqrt{\pi}}, \tag{6.2}$$

and $\star$ denotes the convolution. The denominator is the gradient magnitude of a smoothed version of the initial image. Thus, the value of s is closer to 1 in flat areas ($|\nabla I| \to 0$) and closer to 0 in areas with large changes in image intensity, i.e., the local edge features. The minimal size of the details that are detected is related to the size of the kernel, which acts like a scale parameter. By viewing s as a potential function, we note that its minima denote the position of edges, as depicted in Figure 6.1.

The edge indicator s also induces a metric $g\delta_{ij}$, where $g = \frac{1}{s^2}$ and δ_{ij} is the Kronecker function. Since this metric is conformal, we get

$$|\eta|_g = g|\eta|, \quad D^{\epsilon}_{g\eta} u = \frac{1}{\sqrt{g}} D^{\epsilon}_{\eta} u,$$

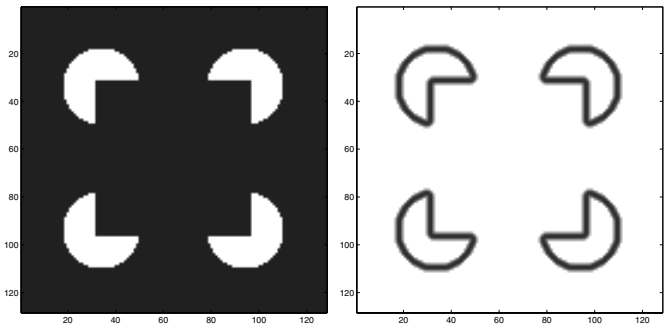


FIG. 6.1. *The Kanizsa square (left) with the connectivity map g (right).*

and, in order to study a curvature equation, we will choose

$$h = g.$$

The phase equation (1.7) becomes

$$\partial_t u(t) = -\int_{\mathbb{R}^n} D_\eta^{-\epsilon}\left(e^{-\sqrt{g}|\eta|}\phi_{\epsilon|\eta|_g}\left(D^\epsilon_{g\eta}u\right)\sqrt{g}\right)d\eta$$

(using the definition of difference quotient)

$$= -\int_{\mathbb{R}^n}\left(\frac{e^{-\sqrt{g}|\eta|}}{(\epsilon|\eta|)^{3/2}}\phi\Big(\frac{u(x)-u(x-\epsilon\eta)}{\sqrt{\epsilon|\eta|g}}\Big) - \frac{e^{\sqrt{g}|\eta|}}{(\epsilon|\eta|)^{3/2}}\phi\Big(\frac{u(x+\epsilon\eta)-u(x)}{\sqrt{\epsilon|\eta|g}}\Big)\right)d\eta$$

since ϕ is odd

$$= 2\int_{\mathbb{R}^n}\frac{e^{-\sqrt{g}|\eta|}}{(\epsilon|\eta|)^{3/2}}\phi\Big(\frac{u(x+\epsilon\eta)-u(x)}{\sqrt{\epsilon|\eta|g}}\Big)d\eta.$$

We note that the exponential kernel $e^{-\sqrt{g}|\eta|}$ can be substituted by a compactly supported kernel $\chi = \chi(\sqrt{g}|\eta|)$. The new equation and the corresponding functional F_ϵ satisfy the same convergence results as before. We will assume that χ is the indicatrix function of the square $[-1,1]^2$ so that in the numerical simulations the integral will be approximated with the sum on the vectors

$$\eta = (i,j), \quad i,j \in \{0,1,-1\}.$$

According to the introduction, the function ϕ will be the sin function, extended with zero, outside of the interval $[-\pi,\pi]$. To perform numerical simulations the phase equation has been approximated by forward differences in time:

$$u^{n+1}_{l,m} = u^n_{l,m} + 2\Delta t \sum_{(i,j)\in\{0,-1,1\}} \frac{1}{\epsilon^{3/2}(i^2+j^2)^{3/4}}\sin\Big(\frac{u^n(l+i,m+j)-u^n(l,m)}{\sqrt{\epsilon(i^2+j^2)^{1/2}g(l+i/2,m+j/2)}}\Big),$$

where $\epsilon = 0.03$ is the space increment and $\Delta t = 0.01$ is the time discretization. As in [37], the initial condition is given by a function $u_0 = \mathcal{D}$ that is proportional to

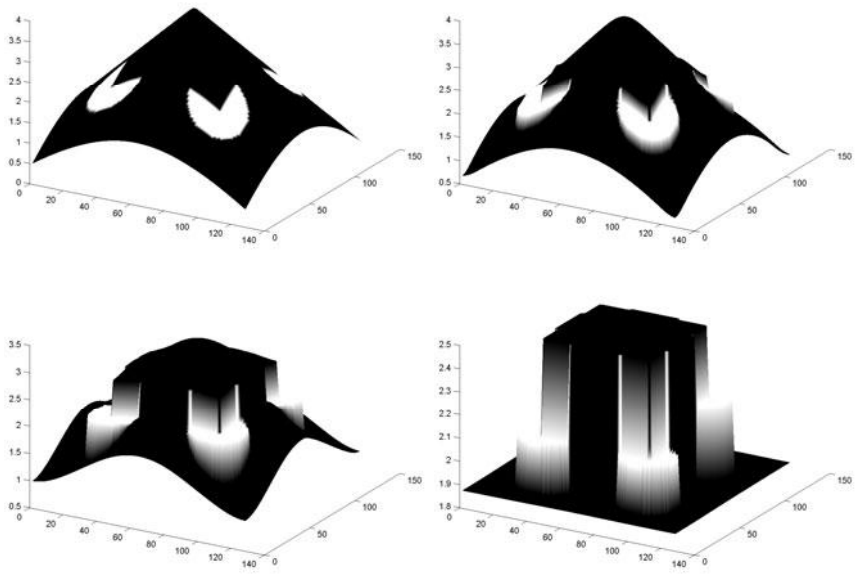


FIG. 6.2. *Evolution of the phase equation towards the phase locking solution segmenting the Kanizsa square.*

the distance from a point internal to the object. We impose Neumann boundary conditions.

During the flow, the surface evolves towards the piecewise constant solution by continuation and closing of the boundary fragments and the filling in of the homogeneous regions (Figure 6.2). In regions of the image where edge information exists, the level sets of the surface get attracted to the edges and accumulate. Consequently, the spatial gradient increases, and the surface begins to develop a discontinuity. In the regions of the image corresponding to subjective contours (i.e., contours that are perceived without any existing discontinuity in the image) discontinuities of u are propagated from existing edge fragments (Figure 6.2).

Acknowledgment. The authors thank Prof. L. Ambrosio and Prof. B. Franchi for some useful conversations on the subject of this work.

REFERENCES

[1] L. AMBROSIO, *A compactness theorem for a new class of functions of bounded variation*, Bull. Un. Mat. Ital. B (7), 3 (1989), pp. 857–851.
[2] L. AMBROSIO, *Existence theory for a new class of variational problems*, Arch. Rational Mech. Anal., 111 (1990), pp. 291–322.
[3] L. AMBROSIO AND D. PALLARA, *Partial regularity of free discontinuity sets.* I, Ann. Sc. Norm. Sup. Pisa Cl. Sci. (4), 24 (1997), pp. 1–38.
[4] L. AMBROSIO, N. FUSCO, AND D. PALLARA, *Partial regularity of free discontinuity sets.* II, Ann. Sc. Norm. Sup. Pisa Cl. Sci. (4), 24 (1997), pp. 39–62.
[5] L. AMBROSIO AND V. M. TORTORELLI, *On the approximation of free discontinuity problems*, Boll. Uni. Mat. Ital. B (7), 6 (1992), pp. 105–123.
[6] A. BALDI, *Weighted BV functions*, Houston J. Math., 27 (2001), pp. 683–705.

[7] G. Bellettini and M. Paolini, *Anisotropic motion by mean curvature in the context of Finsler geometry*, Hokkaido Math. J., 25 (1996), pp. 537–566.
[8] A. Bonnet, *On the regularity of edges in image segmentation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 13 (1996), pp. 485–528.
[9] A. Bonnet and G. David, *Cracktip is a global Mumford-Shah minimizer*, Astérisque, 274 (2001).
[10] A. Braides, *Approximation of Free-Discontinuity Problems*, Lecture Notes in Math. 1694, Springer-Verlag, Berlin, 1998.
[11] A. Braides and M. S. Gelli, *From Discrete to Continuum: A Variational Approach*, Lecture Notes, SISSA, Trieste, 2000.
[12] A. Braides and M. S. Gelli, *Continuum limits of discrete systems without convexity hypotheses*, Math. Mech. Solids, 7 (2002), pp. 41–66.
[13] A. Braides and G. Dal Maso, *Non-local approximation of the Mumford-Shah functional*, Calc. Var. Partial Differential Equations, 5 (1997), pp. 293–322.
[14] A. Braides, G. Dal Maso, and A. Garroni, *Variational formulation of softening phenomena in fracture mechanics: The one-dimensional case*, Arch. Ration. Mech. Anal., 146 (1999), pp. 23–58.
[15] A. Chambolle, *Finite-differences discretizations of the Mumford-Shah functional*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 261–288.
[16] A. Chambolle and G. Dal Maso, *Discrete approximation of the Mumford-Shah functional in dimension two*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 651–672.
[17] G. Cortesani, *Sequences of non-local functionals which approximate free-discontinuity problems*, Arch. Rational Mech. Anal., 144 (1998), pp. 357–402.
[18] G. David, *C^1-arcs for minimizers of the Mumford–Shah functional*, SIAM J. Appl. Math., 56 (1996), pp. 783–888.
[19] G. Dal Maso, *An Introduction to Γ-Convergence*, Birkhäuser Boston, Boston, 1993.
[20] E. De Giorgi, M. Carriero, and A. Leaci, *Existence theorem for a minimum problem with free discontinuity set*, Arch. Rational Mech. Anal., 108 (1989), pp. 195–218.
[21] F. Demengel, *Some remarks on variational problems on* $\mathrm{BV}(\Omega, S^1)$ *and* $W^{1,1}(\Omega, S^1)$, Comm. Partial Differential Equations, 18 (1993), pp. 1055–1068.
[22] A. Engel, P. Konig, C. Gray, and W. Singer, *Temporal coding by coherent oscillations as a potential solution to the binding problem: Physiological evidence*, in Nonlinear Dynamics and Neural Networks, H. Schuster, ed., Springer-Verlag, Berlin, 1992.
[23] L. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC, Boca Raton, FL, 1992.
[24] H. Federer, *Geometric Measure Theory*, Springer-Verlag, Berlin, 1969.
[25] D. J. Field, A. Hayes, and R. F. Hess, *Contour integration by the human visual system: Evidence for a local association field*, Vis. Res., 2 (1993), pp. 173–193.
[26] I. Fonseca and N. Fusco, *Regularity results for anisotropic image segmentation models*, Ann. Sc. Norm. Sup. Pisa Cl. Sci. (4), 24 (1997), pp. 463–499.
[27] N. Fusco, G. Mingione, and C. Trombetti, *Regularity of minimizers for a class of anisotropic free discontinuity problems*, J. Convex Anal., 8 (2001), pp. 349–367.
[28] M. Gobbino, *Finite difference approximation of the Mumford-Shah functional*, Comm. Pure Appl. Math., 51 (1998), pp. 197–227.
[29] M. Gobbino, *Gradient flow for the one dimensional Mumford-Shah functional*, Ann. Sc. Norm. Sup. Pisa Cl. Sci. (4), 27 (1998), pp. 145–193.
[30] M. Gobbino and M. G. Mora, *Finite-difference approximation of free-discontinuity problems*, Proc. Roy. Soc. Edinburgh, Sect. A 131 (2001), pp. 567–595.
[31] P. Hartmann, *Ordinary Differential Equations*, Birkhäuser Boston, Boston, 1982.
[32] J. Jost, *Riemannian Geometry and Geometric Analysis*, Springer-Verlag, Berlin, 1995.
[33] G. Kanizsa, *Organization in Vision*, Praeger, New York, 1979.
[34] N. Kopell and G. B. Ermentrout, *Symmetry and phaselocking in chains of weakly coupled oscillators*, Comm. Pure Appl. Math., 39 (1986), pp. 623–660.
[35] N. Kopell and G. B. Ermentrout, *Phase transitions and other phenomena in chains of coupled oscillators*, SIAM J. Appl. Math., 50 (1990), pp. 1014–1052.
[36] D. Mumford and J. Shah, *Optimal approximation by piecewise smooth functions and associated variational problems*, Comm. Pure Appl. Math., 17 (1989), pp. 577–685.
[37] A. Sarti, R. Malladi, and J. A. Sethian, *Subjective surfaces: A method for completion of missing boundaries*, Proc. Nat. Acad. Sci. USA, 97 (2000), pp. 6258–6263.
[38] H. G. Schuster and P. Wagner, *A model for neuronal oscillators in the visual cortex. I: Mean field theory and derivation of the phase equations*, Biol. Cybern., 64 (1990), pp. 77–82.

[39] H. G. Schuster and P. Wagner, *A model for neuronal oscillators in the visual cortex.* II: *Phase description of the feature dependent synchronization*, Biol. Cybern., 64 (1990), pp. 83–85.

[40] C. Trombetti, *Existence of minimizers for a class of anysotropic free discontinuity problems*, Ann. Mat. Pura Appl. (4), 177 (1999), pp. 277–292.

[41] H. R. Wilson and J. D. Cowan, *Excitatory and inhibitory interactions in localized populations of model neurons*, Biophys. J., 12 (1972), pp. 1–24.

SIAM J. MATH. ANAL.
Vol. 35, No. 6, pp. 1420–1450

DESTABILIZATION OF FRONTS IN A CLASS OF BISTABLE SYSTEMS*

ARJEN DOELMAN†, DAVID IRON‡, AND YASUMASA NISHIURA§

Abstract. In this article, we consider a class of bistable reaction-diffusion equations in two components on the real line. We assume that the system is singularly perturbed, i.e., that the ratio of the diffusion coefficients is (asymptotically) small. This class admits front solutions that are asymptotically close to the (stable) front solution of the "trivial" scalar bistable limit system $u_t = u_{xx}+u(1-u^2)$. However, in the system these fronts can become unstable by varying parameters. This destabilization is caused by either the essential spectrum associated to the linearized stability problem or by an eigenvalue that exists near the essential spectrum. We use the Evans function to study the various bifurcation mechanisms and establish an explicit connection between the character of the destabilization and the possible appearance of saddle-node bifurcations of heteroclinic orbits in the existence problem.

Key words. pattern formation, bistable systems, geometric singular perturbation theory, stability analysis, Evans functions

AMS subject classifications. 35B25, 35B32, 35B35, 35K57, 35P20, 34A26, 34C37

DOI. 10.1137/S0036141002419242

1. Introduction.

The class of bistable reaction-diffusion equations we consider in this paper is given by

$$\begin{cases} U_t &= \varepsilon^2 U_{xx} &+ (1+V-U^2)U, \\ \tau V_t &= V_{xx} &+ F(U^2,V;\varepsilon), \end{cases} \tag{1.1}$$

where $F(U^2,V;\varepsilon)$ is a smooth function of U^2, V, and ε such that $F(1,0;\varepsilon)\equiv 0$ and $\lim_{\varepsilon\to 0} F(U^2,V;\varepsilon)$ exists; $\tau > 0$ is a parameter. Thus, the system is such that the background state $(U,V)\equiv(\pm 1,0)$ is always a solution. We furthermore assume that the ratio of the two diffusion coefficients, ε^2, is asymptotically small; thus, the problem has a singularly perturbed nature. We consider the system on the (unbounded) line, i.e., $(U,V) = (U(x,t),V(x,t))$ with $(x,t)\in\mathbb{R}\times\mathbb{R}^+$. Note that (1.1) is (by construction) symmetric under

$$U \to -U. \tag{1.2}$$

To motivate the structure of (1.1) we introduce the fast variable

$$\xi = \frac{x}{\varepsilon} \tag{1.3}$$

*Received by the editors December 10, 2002; accepted for publication (in revised form) August 22, 2003; published electronically February 18, 2004.
http://www.siam.org/journals/sima/35-6/41924.html

†Centrum voor Wiskunde en Informatica, P. O. Box 94079, 1090 GB Amsterdam, The Netherlands and Korteweg-deVries Instituut, Universiteit van Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands (doelman@science.uva.nl). This author was supported by the "Research Training Network (RTN): Front Singularities" (RTN contract HPRN-CT-2002-00274).

‡Korteweg-deVries Instituut, Universiteit van Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands (diron@math.uci.edu). This author was supported by the NSERC by way of a postdoctoral fellowship and the "Research Training Network (RTN): Front Singularities" (RTN contract HPRN-CT-2002-00274).

§Laboratory of Nonlinear Studies and Computation, Research Institure for Electronic Science, Hokkaido University, Kita-ku, Sapporo, 060, Japan (nishiura@aurora.es.hokudai.ac.jp).

so that (1.1) can be written in its equivalent "fast" form,

$$\left\{ \begin{array}{rcl} U_t & = & U_{\xi\xi} + (1+V-U^2)U, \\ \varepsilon^2\tau V_t & = & V_{\xi\xi} + \varepsilon^2 F(U^2,V;\varepsilon). \end{array} \right. \tag{1.4}$$

Since $U(x,t)$ and $V(x,t)$ are a priori supposed to be bounded on the entire domain $\mathbb{R}\times\mathbb{R}^+$, we find in the natural (fast reduced) limit, i.e., $\varepsilon\to 0$ in (1.4), that $V\equiv V_0$ and that U is a solution of the well-studied, scalar (standard) bistable or Nagumo equation,

$$U_t = U_{\xi\xi} + (1+V_0-U^2)U. \tag{1.5}$$

In this paper we interpret the original system, (1.1) or (1.4), as a scalar bistable Nagumo equation (1.5) in which the coefficient of the linear term is allowed to evolve by reaction and diffusion on a long, or slow, spatial scale. Note that the (slow) dynamics of the V-component are allowed to be completely general, except that it is assumed that the full system conserves the symmetry (1.2) and the background states $U\equiv\pm 1$, at $V\equiv 0$, of the scalar limit (see also Remark 1.1). A priori, one expects that the V-component of front-like solutions will remain small ($\mathcal{O}(\varepsilon)$) due to the "boundary conditions" $V=0$ at $\pm\infty$ so that the effect of the slowly varying $V(x,t)$-component cannot have a significant influence on the (well-understood) dynamics of the scalar Nagumo equation. An important motivation of the research in this paper is to find out whether or not this intuition is correct.

We will focus completely on the existence and stability issues associated to the persistence of the asymptotically stable stationary front solutions of the bistable equation (1.5) with $V_0=0$. In fact, this paper can also be seen as a first step towards analyzing the dynamics (and possibly defects) of striped patterns in a class of relatively simple bistable reaction-diffusion equations, i.e., (1.1) for $(U,V)=(U(x,y,t),V(x,y,t))$ with $(x,y)\in\mathbb{R}^2$. The methods and techniques developed in this paper are supposed to carry over to the analysis of the existence and stability of spatially periodic solutions of (1.1) and their two-dimensional counterparts (the planar fronts and the stripe patterns). See also section 5.

The problem of the persistence of the stable front solution of the scalar bistable equation (1.5) is quite subtle, as can be expected in light of recent results on the stability of pulses in singularly perturbed reaction-diffusion equations of the Gray–Scott and Gierer–Meinhardt type [4, 5]. Such systems can also be written in the form (1.4); however, the scalar limit systems are monostable, i.e., in essence of the form $U_t = U_{\xi\xi} - U + U^2$. The pulses correspond in this (fast reduced) limit to the stationary homoclinic solution of $u_{\xi\xi} - u + u^2 = 0$. Thus, one would expect that the pulses of the full system cannot be stable, since the stability problem associated to the homoclinic solution has an $\mathcal{O}(1)$ unstable eigenvalue. Nevertheless, stable pulses of this type do exist in the Gray–Scott and the Gierer–Meinhardt equation [4, 5]. On the other hand, the stability of the pulses in these monostable equations is strongly related to the freedom one has in these systems to scale the magnitude of the pulses; i.e., the amplitude of the stable pulses is asymptotically large in ε in these monostable systems. Such scalings are not possible for the fronts in the bistable case, since the background states $(\pm 1,0)$ are fixed (and $\mathcal{O}(1)$).

In the analysis of the front solutions, we will find that it is natural to decompose $F(U^2,V;\varepsilon)$ into a component that has a factor of $(1+V-U^2)$ and a rest term $G(V;\varepsilon)$

that does not depend on U^2. Hence, we write (1.1) as

$$\left\{ \begin{array}{rcl} U_t &=& \varepsilon^2 U_{xx} + (1+V-U^2)U, \\ \tau V_t &=& V_{xx} + (1+V-U^2)H(U^2,V;\varepsilon) + G(V;\varepsilon), \end{array} \right. \tag{1.6}$$

with $G(0,\varepsilon) \equiv 0$. Note that this decomposition induces no restriction on $F(U^2,V;\varepsilon)$ since we have assumed that F is smooth. In fact,

$$G(V) = F(1+V,V) \text{ and } (1+V-U^2)H(U^2,V) = F(U^2,V) - F(1+V,V).$$

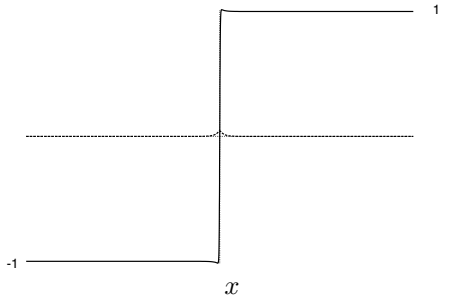
We will find that the quantities $\frac{\partial G}{\partial V}(0;\varepsilon)$ and $H(1,0;\varepsilon)$ have a crucial impact on the structure and the dynamics of the front-like solutions. Therefore, we define

$$G_1(\varepsilon) = \frac{\partial G}{\partial V}(0;\varepsilon) \text{ and } H_0 = H(1,0;\varepsilon); \tag{1.7}$$

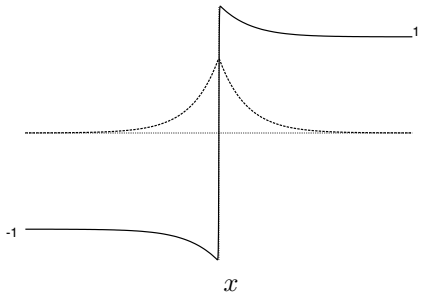
G_1 is the main bifurcation parameter used in this paper. Throughout this paper we assume that $H(U^2,V)$ is nondegenerate, i.e., that $H(1+V,V)$ is not identically 0, and that $\tau = \mathcal{O}(1)$ (see Remark 4.13).

In section 2 we will show that as long as $G_1 < 0$ and $\mathcal{O}(1)$, the front solutions of (1.5) with $V_0 = 0$ persist in a regular fashion, in the sense that the system (1.1) has a front solution with U-components that are asymptotically and uniformly close to a front in (1.5) with $V_0 = 0$ and with V-components that are asymptotically and uniformly small (Theorem 2.1). However, if G_1 becomes $O(\varepsilon^2)$, these fronts become truly singular, in the sense that V becomes $\mathcal{O}(1)$, while the U-component is close to a front of (1.5) with $V_0 \neq 0$ on the fast spatial scale (and it converges to $U = \pm 1$ on the slow spatial scale). Moreover, the front solutions are no longer uniquely determined; there can be several types of heteroclinic front solutions if $G_1 = \mathcal{O}(\varepsilon^2)$ that may or may not merge in saddle-node bifurcations of heteroclinic orbits when G_1 is varied (Theorems 2.3 and 2.5). It should be noted here that for simplicity we consider $G(V) = -\varepsilon^2 \gamma V$ in (1.1) in the singular limit $G_1 = \mathcal{O}(\varepsilon^2)$ throughout this paper—see Remark 2.4. We refer to Figure 1.1 for a numerical representation of a regular front (Figure 1.1(a)) and a singular front (Figure 1.1(b)). The magnitude of G_1 is also extremely relevant in the stability analysis. It can be shown that the (regular) front solutions are asymptotically stable as long as $G_1 < 0$ and $\mathcal{O}(1)$ and $H_0 + G_1 - 2\tau < 0$ and $\mathcal{O}(1)$—see Theorem 4.3. It seems, at leading order, that the destabilization of the front is caused by the essential spectrum, σ_{ess}, associated to the stability of the front (σ_{ess} reaches the imaginary axis exactly at $G_1 = 0$ or at $H_0 + G_1 - 2\tau = 0$—see Lemma 3.1). However, the analysis also shows that there can be eigenvalues near the "tips" of σ_{ess} and that it is possible that the destabilization is caused by such an eigenvalue, i.e., by an element of the discrete spectrum and not by σ_{ess}. These "new" eigenvalues do not have counterparts in the (scalar) fast reduced limit problem; they have a singular slow-fast nature and may appear through edge bifurcations from the essential spectrum.

In section 4 we study in detail the nature of the destabilization as $G_1 < 0$ increases towards 0. In this section it becomes clear that there is an intimate relation between the geometrical character of the singularly perturbed existence problem and the character of the destabilization of the front. This is a natural and frequently encountered relation—see, for instance, [14] and the references therein. We establish that a front solution destabilizes at a critical value of $G_1 = -\varepsilon^2 \gamma_{\text{double}} < 0$ by an eigenvalue if and only if it merges with another front solution in a saddle-node bifurcation of heteroclinic orbits. Moreover, we are able to determine the explicit value



(a) Regular front, $G_1 = -1.0$.



(b) Singular front, $G_1 = -2\varepsilon^2$.

FIG. 1.1. *Two stable front solutions of* (1.1)/(1.6) *plotted on the slow spatial scale* x *(by a numerical simulation). Here* $H(U^2, V) = H_0 U^2$, $G(V) = G_1 V$, $\varepsilon = 0.1$, *and* $H_0 = 1$. *The solid curves represent the* U*-coordinates, and the dotted curves represent the* V*-coordinates.*

of this bifurcation value to be $\gamma_{\text{double}} > 0$. If the front does not "encounter" such a saddle-node as G_1 increases to 0, the front will be destabilized by σ_{ess} at $G_1 = 0$—see Theorems 4.6 and 4.10.

Another way to motivate the analysis of this paper is as follows. In this paper we show that the technique of decomposing the Evans function associated to the stability of a "localized structure" (a (traveling) pulse or front) into the product of an analytic "fast" and a meromorphic "slow" transmission function [4, 5] can be extended to a class of bistable equations. We show that the slow transmission function ($t_2(\lambda, \varepsilon)$) is a natural tool for analyzing the existence or appearance of eigenvalues near or from the essential spectrum and that such eigenvalues play a crucial role in the stability of the front. Note that in this sense the theme of this paper is similar to that of [15], where Evans function techniques are developed to study eigenvalues near σ_{ess} in a class of nearly integrable systems.

The paper is organized as follows. The existence problem is studied in section 2. In section 3 the basic properties of the linearized stability problem are studied and (the decomposition of) the Evans function is introduced. Section 4 is the main section of the paper; in it we develop an approach by which the (possible) location and existence of "slow-fast eigenvalues" near the essential spectrum can be studied. This section is split into three parts: a subsection on the regular problem, a subsection in

which we study an explicit example ($G(V) = -\varepsilon^2\gamma$, $H(U^2, V) = H_0U^2$) in full detail, and a subsection in which we study the "fate" of the regular front as G_1 approaches 0 in the general case. In section 5 we present simulations which clearly exhibit the impact of the distinction between a destabilization by the discrete or by the essential spectrum. Moreover, we discuss some related issues and topics of future research.

Remark 1.1. Large parts of the theory developed in this paper can be generalized to systems of the type (1.1)/(1.4) in which the fast reduced limit system is of the type $U_t = U_{\xi\xi} + B(U^2; V_0)U$ for some function B, i.e., to bistable systems of a more general nature. We focused on the standard case, i.e., $B = 1 + V_0 - U^2$, since the analysis is more transparent. If one drops the condition on the symmetry (1.2), the fronts will, in general, travel with a certain (nonzero) speed. Although the symmetry is used throughout this paper, there is no reason to expect that such asymmetric systems cannot be studied along the lines of the methods presented here.

Notation and definitions. Let $\rho(\varepsilon)$ be a function of $\varepsilon \geq 0$ that is smooth and positive for $\varepsilon > 0$ such that $\lim_{\varepsilon\downarrow 0} \rho(\varepsilon) = 0$ ($\rho(\varepsilon)$ is called an order function [8]). Let $R(z;\varepsilon) \in \mathbb{R}^m$ or $\in \mathbb{C}^m$ ($m \geq 1$) be a certain expression that depends on ε (among other variables or parameters $z \in \mathbb{R}^p, \mathbb{C}^p$ for some $p \geq 0$) such that the limit $\varepsilon \downarrow 0$ exists, i.e., $\lim_{\varepsilon\downarrow 0} R(z;\varepsilon) \stackrel{\text{def}}{=} R_0(z)$. Throughout this paper, the following notation will be used to describe the rate of convergence of $R(z;\varepsilon)$ to $R_0(z)$:

$$R(z;\varepsilon) = R_0(z) + \mathcal{O}(\rho(\varepsilon)).$$

By definition, this is equivalent to the statement that there exists a constant $C > 0$, which is independent of ε, and an $\varepsilon_0 > 0$ such that $\|R(z;\varepsilon) - R_0(z)\| < C\rho(\varepsilon)$ for $0 < \varepsilon < \varepsilon_0$ (here $\|.\|$ is the standard Euclidean norm). Note that both C and ε_0 may be z-dependent. As is usual in (singular) perturbation theory, the precise structure of $\rho(\varepsilon)$ is crucial at many steps in the forthcoming analysis. If this is not the case, $R(z;\varepsilon)$ is often said to be "asymptotically close" to $R_0(z)$, or, equivalently, $\|R(z;\varepsilon) - R_0(z)\| \ll 1$, which implies only that there exists an (unspecified) $\rho(\varepsilon)$ such that $R(z;\varepsilon) = R_0(z) + \mathcal{O}(\rho(\varepsilon))$, i.e., that $\lim_{\varepsilon\downarrow 0} R(z;\varepsilon) = R_0(z)$.

In this paper, the expression $R(z;\varepsilon)$ is said to be $\mathcal{O}(1)$ with respect to ε if there are constants $C^{\pm} > 0$ and $\varepsilon_0 > 0$ such that $C^- < \|R(z;\varepsilon)\| < C^+$ for $0 < \varepsilon < \varepsilon_0$. We refer the reader to [8] for more details on order functions (including the definitions of "$\gg$" and $\mathcal{O}(\frac{1}{\rho(\varepsilon)})$).

2. The existence problem. We analyze the existence of stationary one-dimensional patterns through geometric singular perturbation theory [9, 12] using the methods developed in [6, 5]. Therefore, we write the ODE associated to (1.6) as a dynamical system in $\mathbb{R}^4$,

$$\left\{\begin{array}{rcl} \dot{u} &=& p, \\ \dot{p} &=& -(1 + v - u^2)u, \\ \dot{v} &=& \varepsilon q, \\ \dot{q} &=& \varepsilon\left[-(1 + v - u^2)H(u^2, v;\varepsilon) - G(v;\varepsilon)\right], \end{array}\right. \tag{2.1}$$

where $\dot{}$ denotes the derivative with respect to the spatial variable ξ (1.3) (i.e., ξ "plays the role of time"). Note that this system inherits two symmetries of (1.6),

$$\xi \to -\xi, \quad p \to -p, \quad q \to -q \ \text{ and } \ u \to -u, \quad p \to -p. \tag{2.2}$$

We consider the "superslow" case in which $G_1(\varepsilon) = \mathcal{O}(\varepsilon^2)$ separately in sections 2.2 and 2.3. Note that in the fast reduced limit, i.e., $\varepsilon \to 0$ in (2.1), the monotonically increasing heteroclinic front solution is given by (u_0, p_0, v_0, q_0), where

(2.3)
$$(u_0(\xi; v_0), p_0(\xi; v_0)) = \left(\sqrt{1+v_0} \tanh\left(\sqrt{\frac{1+v_0}{2}} \xi \right), \frac{1+v_0}{\sqrt{2}} \mathrm{sech}^2 \left(\sqrt{\frac{1+v_0}{2}} \xi \right) \right),$$

and v_0 and q_0 are constants.

2.1. The regular case. The main result of this section is the following theorem.

THEOREM 2.1. *Let $G_1(\varepsilon)$ (1.7) be $\mathcal{O}(1)$ and negative. Then, for $\varepsilon > 0$ small enough, system (2.1) has a symmetric pair of heteroclinic orbits: $\Gamma_h^+(\xi;\varepsilon) = (u_h(\xi;\varepsilon), p_h(\xi;\varepsilon), v_h(\xi;\varepsilon), q_h(\xi;\varepsilon))$ and $\Gamma_h^-(\xi;\varepsilon) = (-u_h(\xi;\varepsilon), -p_h(\xi;\varepsilon), v_h(\xi;\varepsilon), q_h(\xi;\varepsilon))$, with $\lim_{\xi\to\pm\infty} \Gamma_h^+(\xi;\varepsilon) = (\pm 1, 0, 0, 0)$ and $\lim_{\xi\to\pm\infty} \Gamma_h^-(\xi;\varepsilon) = (\mp 1, 0, 0, 0)$; $u_h(\xi;\varepsilon)$ and $q_h(\xi;\varepsilon)$ are odd and monotonic as functions of ξ, and $v_h(\xi;\varepsilon)$ and $p_h(\xi;\varepsilon)$ are even. Moreover, $|u_h(\xi;\varepsilon) - u_0(\xi;0)| = \mathcal{O}(\varepsilon)$ (2.3) and $|v_h(\xi;\varepsilon)|, |q_h(\xi;\varepsilon)| = \mathcal{O}(\varepsilon)$, both uniformly on $\mathbb{R}$; $v_h(0;\varepsilon)$ is the extremal value of $v_h(\xi;\varepsilon)$, with*

$$v_h(0;\varepsilon) = \frac{\varepsilon}{2\sqrt{-G_1(0)}} \int_{-\infty}^{\infty} \left(1 - u_0^2(\xi;0)\right) H(u_0^2(\xi;0), 0) d\xi + \mathcal{O}(\varepsilon^2). \tag{2.4}$$

The orbits $\Gamma^{\pm}(\xi;\varepsilon)$ correspond to the (stationary) front patterns $(\pm U_h(\xi;\varepsilon), V_h(\xi;\varepsilon))$ of (1.6) with $U_h(\xi;\varepsilon) = u_h(\xi;\varepsilon)$ odd as a function of ξ, $V_h(\xi;\varepsilon) = v_h(\xi;\varepsilon) = \mathcal{O}(\varepsilon)$ even, $\lim_{\xi\to\pm\infty} U_h(\xi;\varepsilon) = \pm 1$, and $\lim_{\xi\to\pm\infty} V_h(\xi;\varepsilon) = 0$.

Proof. As the system is singularly perturbed, we also consider (2.1) with the slow scaling $x = \varepsilon\xi$; equation (2.1) is given by

$$\left\{ \begin{array}{rcl} \varepsilon u' &=& p, \\ \varepsilon p' &=& -(1+v-u^2)u, \\ v' &=& q, \\ q' &=& \left[-(1+v-u^2) H(u^2, v; \varepsilon) - G(v;\varepsilon) \right], \end{array} \right. \tag{2.5}$$

where $'$ refers to differentiation with respect to x. System (2.5) is referred to as the slow system. We begin by finding the locally invariant manifolds of (2.5) in the limit $\varepsilon \to 0$. In this limit, the first two equations of (2.5) will reduce to

$$p = 0\,, \quad -(1+v-u^2)u = 0\,. \tag{2.6}$$

The manifold given by $(u, p, v, q) = (0, 0, v, q)$ is not normally hyperbolic and will not be considered. However, the manifolds, denoted $\mathcal{M}_0^{\pm}$, determined by $(u, p, v, q) = (\pm\sqrt{1+v}, 0, v, q)$ are normally hyperbolic and thus, by [9, 12], equation (2.5) possesses locally invariant manifolds $\mathcal{M}_\varepsilon^{\pm}$, which are $\mathcal{O}(\varepsilon)$ close to $\mathcal{M}_0^{\pm}$. We now determine the leading order correction to these manifolds. Let the manifold $\mathcal{M}_\varepsilon^{\pm}$ be given by

$$\mathcal{M}_\varepsilon^{\pm} = \{ u = \pm\sqrt{1+v} + \varepsilon U^{\pm}(v, q; \varepsilon), p = \varepsilon P^{\pm}(v, q; \varepsilon), v, q \}\,. \tag{2.7}$$

To obtain successive approximations of $\mathcal{M}_\varepsilon^{\pm}$, we can expand $U^{\pm} = u_1^{\pm} + \varepsilon u_2^{\pm} + \cdots$, and $P^{\pm} = p_1^{\pm} + \varepsilon p_2^{\pm} + \cdots$. Using the first two lines of (2.5) we find

$$\begin{aligned} p_1^{\pm} &= \frac{q}{2\sqrt{1+v}}, \quad p_2^{\pm} = \frac{\partial u_1^{\pm}}{\partial v} q - \frac{\partial u_1^{\pm}}{\partial q} G(v;\varepsilon), \\ u_1^{\pm} &= 0, \quad u_2^{\pm} = \mp \frac{q^2}{4(1+v)^{5/2}} \mp \frac{G(v;\varepsilon)}{(1+v)^{3/2}}. \end{aligned} \tag{2.8}$$

Hence, the (slow) flow on the slow manifold is given by

$$v'' = -G(v;\varepsilon) + \mathcal{O}(\varepsilon^2)\,. \tag{2.9}$$

To leading order, this flow is integrable. The point $(v,q) = (0,0)$, which corresponds to $(\pm 1,0,0,0)$, is a critical point on $\mathcal{M}_\varepsilon^\pm$. Since $G_1 < 0$, $(0,0)$ is a saddle on $\mathcal{M}_\varepsilon^\pm$ with unstable direction $(1,\sqrt{-G_1})$ and stable direction $(-1,\sqrt{-G_1})$.

A heteroclinic orbit $\Gamma_h^\pm$ from $(\mp 1,0,0,0)$ to $(\pm 1,0,0,0)$ is both an element of $W^u(\mathcal{M}_\varepsilon^\mp)$ and of $W^s(\mathcal{M}_\varepsilon^\pm)$. Here we will consider only Γ_h^+. The existence of Γ_h^- follows from the symmetry (2.2). The orbit Γ_h^+ remains exponentially close to $W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ before it "takes off" and makes a "jump" through the fast field, i.e., the region in phase space in which $(u_\xi, p_\xi) = \mathcal{O}(1)$. After that, it "touches down" on $\mathcal{M}_\varepsilon^+$ and remains exponentially close to it (and to $W^u(1,0,0,0)|_{\mathcal{M}_\varepsilon^+}$)—see Figure 2.1. The change in q by the passage through the fast field is $\mathcal{O}(\varepsilon)$ (2.1); therefore Γ_h^+ must take off from $\mathcal{M}_\varepsilon^-$ and touch down on $\mathcal{M}_\varepsilon^+$ with a q-coordinate that is $\mathcal{O}(\varepsilon)$. Since Γ_h^+ is asymptotic to the saddle points $(0,0) \in \mathcal{M}_\varepsilon^\pm$, it follows that the v-coordinate of Γ_h^+ must also be $\mathcal{O}(\varepsilon)$. Note that we have used here implicitly that $G_1 = \mathcal{O}(1)$.

We will determine whether such a trajectory, as Γ_h^+, is possible using a Melnikov method. Both $W^u(\mathcal{M}_\varepsilon^-)$ and $W^s(\mathcal{M}_\varepsilon^+)$ are $\mathcal{O}(\varepsilon)$ close to the family of heteroclinic orbits in the fast reduced limit of (2.1) given in (2.3). The leading order distance between $W^u(\mathcal{M}_\varepsilon^-)$ and $W^s(\mathcal{M}_\varepsilon^+)$ can be determined by a Melnikov function for slowly varying systems [19]. Both $W^u(\mathcal{M}_\varepsilon^-)$ and $W^s(\mathcal{M}_\varepsilon^+)$ intersect the hyperplane $\{u = 0\}$ transversally. Note that $W^{s,u}(\mathcal{M}_\varepsilon^\pm) \cap \{u = 0\}$ is two-dimensional; thus, since $\{u = 0\}$ is three-dimensional, one expects a one-dimensional intersection $W^u(\mathcal{M}_\varepsilon^-) \cap W^s(\mathcal{M}_\varepsilon^+) \cap \{u = 0\}$. The separation between $W^u(\mathcal{M}_\varepsilon^-)$ and $W^s(\mathcal{M}_\varepsilon^+)$ is, at leading order, measured by the integral,

$$\Delta = \int_{-\infty}^{\infty} \begin{pmatrix} p(\xi) \\ u(\xi) + u^3(\xi) - u(\xi)v_0 \end{pmatrix} \wedge \begin{pmatrix} 0 \\ -u(\xi)\frac{\partial q}{\partial \delta}(\xi) \end{pmatrix} d\xi\,. \tag{2.10}$$

Here the wedge product refers to the scalar cross product, and $\frac{\partial q}{\partial \delta}$ solves the differential equation, $\frac{d}{d\xi}(\frac{\partial q}{\partial \delta}) = q_0\xi$, $\frac{\partial q}{\partial \delta}(0) = 0$. Substituting (2.3) into (2.10) results in the following expression for the leading order splitting distance:

$$\Delta = -\int_{-\infty}^{\infty} \frac{q\xi}{\sqrt{2}} \tanh\left(\frac{\xi}{\sqrt{2}}\right) \operatorname{sech}^2\left(\frac{\xi}{\sqrt{2}}\right) d\xi \quad = -q_0\sqrt{2}\,.$$

Thus, $W^u(\mathcal{M}_\varepsilon^-) \cap W^s(\mathcal{M}_\varepsilon^+) \cap \{u = 0\}$ must be $\mathcal{O}(\varepsilon)$ close to $q = 0$. By the symmetries (2.2), we conclude that $W^u(\mathcal{M}_\varepsilon^-) \cap W^s(\mathcal{M}_\varepsilon^+) \cap \{u = 0\}$ must be identically $q = 0$. Hence, again by (2.2), any solution that connects $\mathcal{M}_\varepsilon^-$ to $\mathcal{M}_\varepsilon^+$ must have a u component that is odd with respect to ξ and a v component that is even with respect to ξ.

We are now ready to determine the take off (touch down) curves $T_o^- \subset \mathcal{M}_\varepsilon^-$ ($T_d^+ \subset \mathcal{M}_\varepsilon^+$) [6, 5]. These curves represent the points at which the one-dimensional family of orbits in $W^u(\mathcal{M}_\varepsilon^-) \cap W^s(\mathcal{M}_\varepsilon^+)$ leave (land on) $\mathcal{M}_\varepsilon^\pm$. Let the elements of this family be denoted $\gamma(\xi;p)$, where the parameter $p > 0$ corresponds to the p-component of $\gamma(\xi;p)$ as it crosses through $\{u = q = 0\}$. Note that the γ-family forms the Fenichel fibering of $W^u(\mathcal{M}_\varepsilon^-) \cap W^s(\mathcal{M}_\varepsilon^+)$ [9] and that each $\gamma(\xi;p)$ is asymptotically close to an unperturbed orbit given in (2.3). To each $\gamma(\xi;p)$ we associate two orbits, $\gamma_{\mathcal{M}_\varepsilon^-}(\xi;p) \subset \mathcal{M}_\varepsilon^-$ and $\gamma_{\mathcal{M}_\varepsilon^+}(\xi;p) \subset \mathcal{M}_\varepsilon^+$, by the fact that $||\gamma(\xi;p) - \gamma_{\mathcal{M}_\varepsilon^\pm}(\xi;p)||$ is

exponentially small if $\pm\xi > \mathcal{O}(\varepsilon^{-1})$. We define T_o^- and T_d^+ as the collections of base points of the Fenichel fibers on $\mathcal{M}_\varepsilon^-$ and on $\mathcal{M}_\varepsilon^-$,

$$T_o^- = \bigcup_{p>0} \gamma_{\mathcal{M}_\varepsilon^-}(0;p)\,, \quad T_d^+ = \bigcup_{p>0} \gamma_{\mathcal{M}_\varepsilon^+}(0;p)\,. \tag{2.11}$$

We can compute the leading order structure of T_o^- and T_d^+ by considering the effect of the journey through the fast field on the slow variables v and q. Since $v_\xi = \varepsilon q$ and $q = \mathcal{O}(\varepsilon)$ it follows that the change in v through the fast field is of higher order, i.e., $\mathcal{O}(\varepsilon^2)$. By construction, q will be an odd function of ξ; thus, the value of q for a given v on T_o^- must be $-\frac{1}{2}\Delta q(v)$, where $\Delta q(v)$ is the change in q due to one full pass through the fast field (during which v remains (at leading order) constant, $v = v_0$). Similarly, the value of q on T_d^- must be $\frac{1}{2}\Delta q(v)$. Since we already know that both v and q must be $\mathcal{O}(\varepsilon)$ in this regular case, we compute $\Delta q(0)$ (by (2.1), (2.3)):

$$\Delta q(0) = \int_{-\infty}^{\infty} \dot{q}|_{v=0}\, d\xi = -\varepsilon \int_{-\infty}^{\infty} \left[1 - \tanh^2\left(\frac{\xi}{\sqrt{2}}\right)\right] H\left(\tanh^2\left(\frac{\xi}{\sqrt{2}}\right), 0\right) d\xi + \mathcal{O}(\varepsilon^2).$$

To establish the existence of the heteroclinic orbit $\Gamma_h^+(\xi)$, we consider the intersection $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ on $\mathcal{M}_\varepsilon^-$, $\mathcal{O}(\varepsilon)$ close to $(-1,0,0,0)$. Thus, T_o^- and $W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ are given by $\{q = -\frac{1}{2}\Delta q(0) + \mathcal{O}(\varepsilon^2)\}$ and $\{q = \sqrt{-G_1}v + \mathcal{O}(\varepsilon^2)\}$. Figure 2.1 shows the superposition of T_o^- with $W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ and of T_d^+ with $W^s(1,0,0,0)|_{\mathcal{M}_\varepsilon^+}$. The v-coordinate of $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ is given in (2.4). Thus, we have established the existence of an orbit $\Gamma_h^+ \in W^u(\mathcal{M}_\varepsilon^-) \cap W^s(\mathcal{M}_\varepsilon^+)$ that is asymptotic to $(-1,0,0,0) \in \mathcal{M}_\varepsilon^-$. Since Γ_h^+ passes through $\{u = 0, q = 0\}$ during its jump through the fast field, it follows by the symmetries (2.2) that Γ_h^+ is indeed the orbit described in the statement of the theorem. As already mentioned, the existence of Γ_h^- also follows immediately from (2.2). ∎

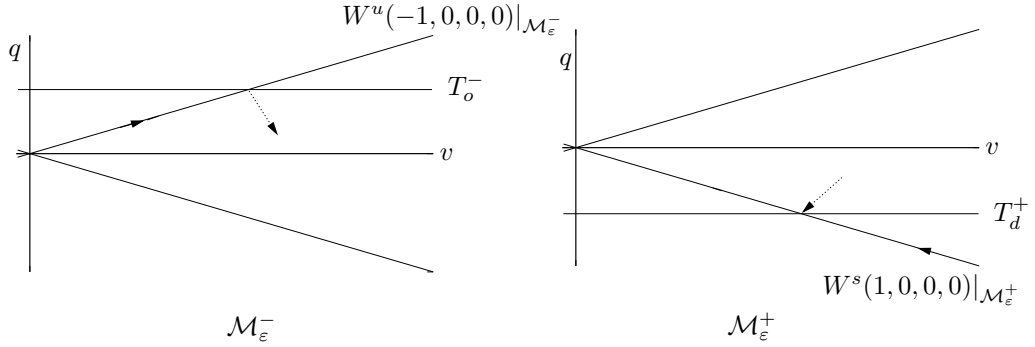


FIG. 2.1. *Superposition of the take off and touch down curves $T_{o,d}^\pm$ with $W^{s,u}(\pm 1,0,0,0)|_{\mathcal{M}_\varepsilon^\pm}$. The intersections $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ and $T_d^+ \cap W^s(1,0,0,0)|_{\mathcal{M}_\varepsilon^+}$ determine the heteroclinic front solution $\Gamma_h^+(\xi;\varepsilon)$. The dotted arrows indicate the orbit $\Gamma_h^+(\xi)$ "taking off" and "touching down."*

Remark 2.2. We note that if $G_1 = \mathcal{O}(\varepsilon^\sigma)$ for some $\sigma \in [0,2)$, then the intersection of T_o^- and $W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ will result in a value of v_0 of $\mathcal{O}(\varepsilon^{1-\frac{1}{2}\sigma}) \ll 1$ (2.4); thus $\Gamma_h^+(\xi)$ will still be a regular perturbation of the orbit in the scalar limit. Moreover, this argument also shows that singular orbits may exist for $G_1 = \mathcal{O}(\varepsilon^2)$.

2.2. The superslow limit: An example. In this section we consider the "significant degeneration" $G_1(\varepsilon) = \mathcal{O}(\varepsilon^2)$. For simplicity, we consider only the case in which the flow on the slow manifolds $\mathcal{M}_\varepsilon^\pm$ is linear, i.e., $G(v;\varepsilon) = \varepsilon^2 G_1(\varepsilon)v \stackrel{\text{def}}{=} -\varepsilon^2\gamma v$, where γ does not depend on ε. Moreover, we first consider an explicit expression for $H(u^2, v; \varepsilon)$, $H(u^2, v; \varepsilon) = H_0 u^2$. The case of a general $H(U^2, V)$ will be considered in the next subsection. We refer the reader to Remark 2.4 for a brief discussion of the case of a general function $G(V)$. System (2.1) reduces to

$$\left\{\begin{array}{rcl} \dot{u} &=& p, \\ \dot{p} &=& -(1+v-u^2)u, \\ \dot{v} &=& \varepsilon q, \\ \dot{q} &=& \varepsilon\left[-(1+v-u^2)H_0u^2 + \varepsilon^2\gamma v\right]. \end{array}\right. \tag{2.12}$$

This system has various types of (singular) heteroclinic orbits.

THEOREM 2.3. *Assume that $G(V) = -\varepsilon^2\gamma V$, that $H(U^2, V) = H_0U^2$, and that ε is small enough.*

(i) *$H_0 > 0$. If $\gamma > \gamma_{\text{double}}$, where $\gamma_{\text{double}} = \frac{3}{2}H_0^2 + \mathcal{O}(\varepsilon)$, equation* (2.12) *has two pairs of heteroclinic orbits, $\Gamma_h^{+,j}(\xi;\varepsilon) = (u_h^j(\xi), p_h^j(\xi), v_h^j(\xi), q_h^j(\xi))$, $j = 1, 2$, and their symmetrical counterparts $\Gamma_h^{-,j}(\xi;\varepsilon) = (-u_h^j(\xi), -p_h^j(\xi), v_h^j(\xi), q_h^j(\xi))$, with $\lim_{\xi\to\pm\infty}\Gamma_h^{+,j}(\xi;\varepsilon) = (\pm1, 0, 0, 0)$. In the fast field $u_h^j(\xi)$ (resp., $v_h^j(\xi)$) is asymptotically and uniformly close to $u_0(\xi; v_j)$* (2.3) *(resp., v_j); the constants v_j are the zeros of $\sqrt{\gamma}v = \frac{2}{3}\sqrt{2}H_0(v+1)^{3/2}$ so that $0 < v_1 < 2 < v_2$ (at leading order). In the slow field, $\Gamma_h^{+,j}(\xi;\varepsilon)$ is exponentially close to $W^{u,s}(\pm1, 0, 0, 0)|_{\mathcal{M}_\varepsilon^\pm} \subset \mathcal{M}_\varepsilon^\pm$. The orbits $\Gamma_h^{\pm,1}(\xi;\varepsilon)$ and $\Gamma_h^{\pm,2}(\xi;\varepsilon)$ merge in a saddle-node bifurcation of heteroclinic orbits as $\gamma \downarrow \gamma_{\text{double}}$. There are no heteroclinic orbits for $\gamma < \gamma_{\text{double}}$.*

(ii) *$H_0 < 0$. The relation $\sqrt{\gamma}v = \frac{2}{3}\sqrt{2}H_0(v+1)^{3/2}$ has a unique zero for all $\gamma > 0$, and there is one pair of heteroclinic orbits $\Gamma_h^\pm(\xi;\varepsilon)$ for all $\gamma > 0$. These orbits have the same structure as described in* (i).

The orbits $\Gamma_h^{\pm(,j)}(\xi;\varepsilon)$ correspond to the front solutions $(U_h^{\pm(,j)}(\xi;\varepsilon), V_h^{\pm(,j)}(\xi;\varepsilon))$ of (1.6) *with $U_h^{\pm(,j)}(\xi;\varepsilon) = \pm u_h^j(\xi;\varepsilon)$ odd and $V_h^{\pm(,j)}(\xi;\varepsilon) = v_h^j(\xi;\varepsilon)$ even as functions of ξ.*

Proof. The essence of the analysis of the superslow system is similar to that of the regular case. The important difference is that, although the change in q by a "jump" through the fast field is still $\mathcal{O}(\varepsilon)$, the v-coordinate of the heteroclinic orbit may now be $\mathcal{O}(1)$, due to the superslow character of the flow on $\mathcal{M}_\varepsilon^\pm$. It is this difference that will cause the bifurcation and the formation of the second orbit in case (i). The flow on the slow manifold is now $\mathcal{O}(\varepsilon^2)$, i.e., superslow, and is at leading order governed by

$$v'' = \varepsilon^2\gamma v\,. \tag{2.13}$$

Since the right-hand side of this equation is $\mathcal{O}(\varepsilon^2)$, one might expect that one needs to incorporate the higher order corrections to the approximation of $\mathcal{M}_\varepsilon^\pm$ (2.8) to determine the leading order flow on $\mathcal{M}_\varepsilon^\pm$. However, the $\mathcal{O}(\varepsilon^2)$ correction contains a term with a q^2 factor and a term with $G(v)$ (2.8). Since we consider $q = \mathcal{O}(\varepsilon)$ on $\mathcal{M}_\varepsilon^\pm$ and since $G(v) = \mathcal{O}(\varepsilon^2)$, the resulting correction will not be of leading order.

Again the equilibria on $\mathcal{M}_\varepsilon^\pm$ are saddles, with stable and unstable directions, $(\pm1, \varepsilon\sqrt{\gamma})$. As in Theorem 2.1 we consider only the orbit that jumps from $\mathcal{M}_\varepsilon^-$ to $\mathcal{M}_\varepsilon^+$ (the others follows from the symmetry (2.2). We repeat the Melnikov calculations and again conclude that $W^u(\mathcal{M}_\varepsilon^-) \cap W^s(\mathcal{M}_\varepsilon^+) \cap \{u = 0\}$ must be identically $q = 0$. Hence,

again by (2.2), any solution that connects $\mathcal{M}_\varepsilon^-$ to $\mathcal{M}_\varepsilon^+$ must have a u component that is odd with respect to ξ and a v component that is even with respect to ξ.

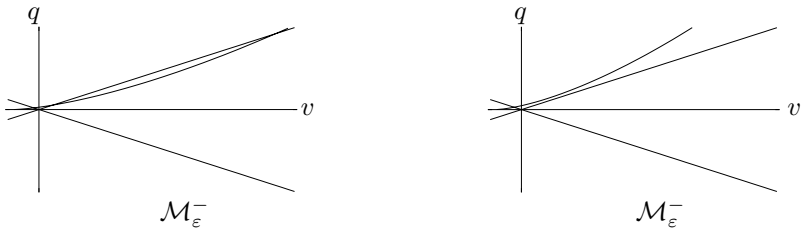
We define the take off, T_o^-, and touch down, T_d^+, curves as in (2.11). We find the leading order behavior of T_d^+ and T_o^- by calculating the change in q as we traverse the fast field. As in the regular case, v remains a constant up to $\mathcal{O}(\varepsilon^2)$, and the value of q on the take off (touch down) curve must be $-\frac{1}{2}\Delta q(v_0)$ ($\frac{1}{2}\Delta q(v_0)$), where v_0 is the (leading order) constant value of the v-coordinate of the orbit that is heteroclinic to $\mathcal{M}_\varepsilon^+$ in the fast field. The calculation of the change in q is similar to that of the regular case except that v_0 now effects the leading order term (2.3),

$$\begin{aligned}\Delta q(v_0) &= -\varepsilon H_0(1+v_0)^2\int_{-\infty}^{\infty}\left[1-\tanh^2\left(\sqrt{\tfrac{v_0+1}{2}}\xi\right)\right]\tanh^2\left(\sqrt{\tfrac{v_0+1}{2}}\xi\right)\,d\xi+\mathcal{O}(\varepsilon^2)\\ &= -\varepsilon\tfrac{2\sqrt{2}}{3}H_0(v_0+1)^{3/2}+\mathcal{O}(\varepsilon^2)\,.\end{aligned}$$

The heteroclinic orbits are again determined by $T_o^-\cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$, where $T_o^- = \{q=-\frac{1}{2}\Delta q(v_0)+\mathcal{O}(\varepsilon^2)\}$ and $W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}=\{q=\varepsilon\sqrt{\gamma}v+\mathcal{O}(\varepsilon^2)\}$,

$$\frac{1}{3}\sqrt{2}H_0(v_0+1)^{3/2}=\sqrt{\gamma}v_0\,; \tag{2.14}$$

see Figure 2.2. Thus, in the superslow case, a heteroclinic orbit may leave $\mathcal{M}_\varepsilon^-$ with a v-coordinate of $\mathcal{O}(1)$. Now if $H_0>0$ and $\gamma>\gamma_{\text{double}}=\frac{3}{2}H_0^2+\mathcal{O}(\varepsilon)$, (2.14) has two possible solutions, $v_0=v_j$, $j=1,2$, with $0<v_1<2<v_2$ (at leading order). These intersections correspond to the heteroclinic orbits $\Gamma_h^{+,j}(\xi)$. For $\gamma<\gamma_{\text{double}}$, there are no solutions to (2.14), and thus no heteroclinic connections exist: the orbits $\Gamma_h^{+,1}(\xi)$ and $\Gamma_h^{+,2}(\xi)$ have coalesced at $\gamma=\gamma_{\text{double}}$. In the case that $H_0<0$, (2.14) has a unique solution for all values of $\gamma>0$; there is only one pair of heteroclinic orbits. ∎



(a) (b)

FIG. 2.2. *Superposition of T_o^- with $W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ in the superslow case with $H_0>0$ for $\gamma>\gamma_{\text{double}}$* (a) *and $\gamma<\gamma_{\text{double}}$* (b).

Remark 2.4. If $G(V)$ is not linear in the singular limit (i.e., $G_1=\mathcal{O}(\varepsilon^2)$), then the analysis becomes more involved, but there are no essentially new phenomena. In this case, the magnitude (with respect to ε) of the second derivative of $G(v)$ at $v=0$ will start to play a role comparable to G_1. Moreover, the flow on $\mathcal{M}_\varepsilon^\pm$ is nonlinear so that $W^{u,s}(\pm1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ is no longer a straight line (at leading order); therefore, many "new" intersections of $T_o^-\cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$, and thus "new" heteroclinic orbits, may appear.

2.3. The superslow limit: The general case. We now consider the general superslow problem; i.e., $H(U^2,V)$ is a general (smooth) function of U^2 and V in this

section. As in section 2.2 and motivated in Remark 2.4, we choose to consider only the case of $G(V;\varepsilon)$ linear; i.e., $G(v;\varepsilon) = -\varepsilon^2\gamma v$ in (2.1). The treatment of the general superslow case and (2.12) is in essence identical to that of the previous section. However, the statement of the main results cannot be formulated as explicitly as in Theorem 2.3, as long as there is no explicit expression given for $H(U^2, V)$. Nevertheless, the character of the existence result is similar to that of Theorem 2.3; there can be various kinds of heteroclinic orbits that might coalesce in saddle-node bifurcations.

As in the proofs of Theorems 2.1 and 2.3, the existence of the heteroclinic orbits is established by the intersection of T_o^- and $W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$, i.e., by the solution v_0 of

$$\sqrt{\gamma}v_0 = \frac{1}{2}\int_{-\infty}^{\infty} \left[1 + v_0 - u_0^2(\xi; v_0)\right] H(u_0^2(\xi; v_0), v_0)d\xi, \tag{2.15}$$

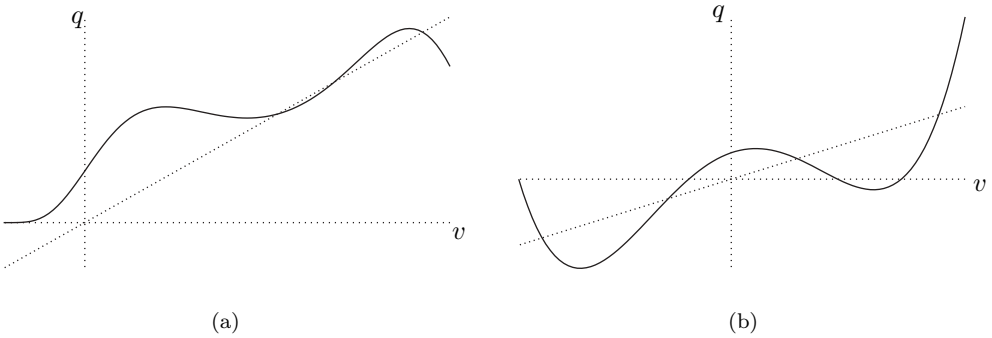
at leading order. Note that the right-hand side equals $-\frac{1}{2}\Delta q(v_0)$, i.e., half the accumulated change in q during a circuit through the fast field, and that we have used (2.3).

THEOREM 2.5. *Assume that $G(V) = -\varepsilon^2\gamma V$ and that ε is small enough. System (2.1) has $n \geq 0$ pairs of heteroclinic orbits, $\Gamma_h^{\pm,j}(\xi;\varepsilon) = (\pm u_h^{\pm,j}(\xi), \pm p_h^{\pm,j}(\xi), v_h^{\pm,j}(\xi), q_h^{\pm,j}(\xi))$, where $j = 1, \ldots, n$, with $\lim_{\xi\to\pm\infty}\Gamma_h^{+,j}(\xi;\varepsilon) = (\pm 1, 0, 0, 0)$. The number $n = n(\gamma)$ is given by the number of solutions v_j of (2.15). In the fast field $u_h^j(\xi)$ (resp., $v_h^j(\xi)$) is asymptotically and uniformly close to $u_0(\xi; v_j)$ (2.3) (resp., v_j), where the constant v_j is the jth zero of (2.15). In the slow field, $\Gamma_h^{+,j}(\xi;\varepsilon)$ is exponentially close to $W^{u,s}(\pm 1, 0, 0, 0)|_{\mathcal{M}_\varepsilon^\pm} \subset \mathcal{M}_\varepsilon^\pm$.*

Two orbits $\Gamma_h^{\pm,j}(\xi;\varepsilon)$ and $\Gamma_h^{\pm,j+1}(\xi;\varepsilon)$ coalesce in a saddle-node bifurcation of heteroclinic orbits at a certain value $\gamma = \gamma_{\text{double}}^j$ if the zeros $v_j \leq v_{j+1}$ of (2.15) merge, i.e., if the intersection $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ is nontransversal.

The orbits $\Gamma_h^{\pm,j}(\xi;\varepsilon)$ correspond to the front solutions $(U_h^{\pm,j}(\xi;\varepsilon), V_h^{\pm,j}(\xi;\varepsilon))$ of (1.6) with $U_h^{\pm,j}(\xi;\varepsilon) = \pm u_h^j(\xi;\varepsilon)$ odd and $V_h^{\pm,j}(\xi;\varepsilon) = v_h^j(\xi;\varepsilon)$ even as functions of ξ.

The proof of this result is in essence identical to that of Theorem 2.3. In Figure 2.3 two examples of the possible richness of the intersection $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ are given.



(a) (b)

FIG. 2.3. *Two examples of the possible character of the intersection $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ for a given $H(U^2, V)$;* (a) *there are* 3 *different singular heteroclinic orbits and* (b) 4 *heteroclinic orbits.*

3. The stability of fronts.

3.1. The essential spectrum. The essential spectrum associated to the stability of the front patterns $(U,V) = (U_h(\xi;\varepsilon), V_h(\xi;\varepsilon))$ is fully determined by the spectrum of the linear stability problem for the (trivial) background states (at $\pm\infty$) $(U,V) \equiv (\pm 1, 0)$ [11]. Therefore, we introduce $k \in \mathbb{R}$ and $\alpha, \beta, \lambda \in \mathbb{C}$ by

$$U(x,t) = \pm 1 + \alpha e^{ik\xi+\lambda t}, \quad V(x,t) = \beta e^{ik\xi+\lambda t}$$

and substitute this expression into (1.6) (using (1.3)). This yields the matrix equation

$$\begin{pmatrix} -k^2-2 & \pm 1 \\ \mp 2\varepsilon^2 H_0 & -k^2+\varepsilon^2(H_0+G_1) \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \lambda \begin{pmatrix} \alpha \\ \varepsilon^2\tau\beta \end{pmatrix},$$

where G_1 and H_0 have been introduced in (1.7). Thus, $\lambda = \lambda(k^2)$ is a solution of the characteristic equation

$$Q(\lambda,k) = (\lambda + k^2 + 2)(\varepsilon^2\tau\lambda + k^2 - \varepsilon^2(H_0+G_1)) + 2\varepsilon^2 H_0 = 0. \tag{3.1}$$

Note that this equation holds for both background states $(\pm 1, 0)$, due to the symmetry (1.2). We may conclude the following lemma.

LEMMA 3.1. *The essential spectrum σ_{ess} associated to* (3.3) *is given by the solutions $\lambda = \lambda(k^2)$ of* (3.1) *with $k \in \mathbb{R}$; σ_{ess} is stable, i.e., $\sigma_{\text{ess}} \in \{\text{Re}(\lambda) < 0\}$, if $G_1 < 0$ and $H_0 + G_1 - 2\tau < 0$.*

Proof. The two conditions in this lemma are obtained directly from

$$\begin{array}{rcl} \lambda_1+\lambda_2 &=& \frac{1}{\varepsilon^2\tau}\left[\varepsilon^2(H_0+G_1-2\tau) - k^2(1+\varepsilon^2\tau)\right] \;<\; 0 \quad \forall\, k, \\ \lambda_1\lambda_2 &=& \frac{1}{\varepsilon^2\tau}\left[k^4 + k^2(2-\varepsilon^2(H_0+G_1)) - 2\varepsilon^2 G_1\right] \;>\; 0 \quad \forall\, k. \end{array} \tag{3.2}$$

Both relations attain their extremal value at $k = 0$. ∎

However, we need to have more information on the essential spectrum than just this stability result. In section 4 we will see that the appearance of edge bifurcations is closely related to the structure of σ_{ess}. We focus on the stable case $G_1 < 0$ and $H_0+G_1-2\tau < 0$. It is straightforward to check that (3.1) has two solutions $\lambda_{1,2}(k) \in \mathbb{R}$ for all $k \in \mathbb{R}$ if $H_0 < 0$. As H_0 passes through zero two k-intervals, $(-k^+, -k^-)$ and (k^-, k^+), $(0 < k^- < k^+)$ appear in which $\lambda_{1,2}(k)$ are complex valued. These intervals merge (i.e., $k^- \downarrow 0$) as H_0 approaches $(\sqrt{2\tau} - \sqrt{-G_1})^2$. For $(\sqrt{2\tau} - \sqrt{-G_1})^2 < H_0 < 2\tau - G_1$ (which is a nonempty region), $\lambda_{1,2}(k) \in \mathbb{C}$ if $-k^+ < k < k^+$. See Figure 3.1.

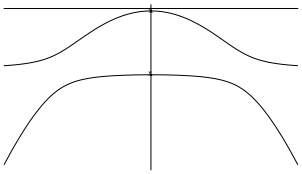
3.2. The linearized stability problem. With a slight abuse of notation we (re-)introduce $u(\xi)$ and $v(\xi)$ by

$$U(\xi,t) = U_h(\xi;\varepsilon) + u(\xi)e^{\lambda t}, \quad V(\xi,t) = V_h(\xi;\varepsilon) + v(\xi)e^{\lambda t},$$

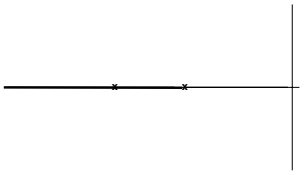
substitute this into (1.6), and linearize

$$\begin{array}{rcl} u_{\xi\xi} &+& (1 + V_h - 3U_h^2 - \lambda)u = -U_h v \\ v_{\xi\xi} &=& \varepsilon^2 \left\{ 2\left[H(U_h^2,V_h) - (1+V_h-U_h^2)\frac{\partial H}{\partial U^2}(U_h^2,V_h)\right] U_h u \right. \\ && \left. - \left[H(U_h^2,V_h) - (1+V_h-U_h^2)\frac{\partial H}{\partial V}(U_h^2,V_h) + \frac{\partial G}{\partial V}(V_h) - \tau\lambda\right] v \right\}. \end{array} \tag{3.3}$$

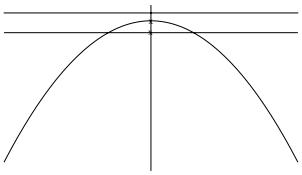
Note that the front pattern $(U_h(\xi), V_h(\xi))$ corresponds to any of the regular or singular heteroclinic orbits $\Gamma_h^{\pm,j}(\xi)$ of Theorems 2.1, 2.3, and 2.5. In the stability analysis of forthcoming sections we will consider only the front patterns of +-type, i.e., those



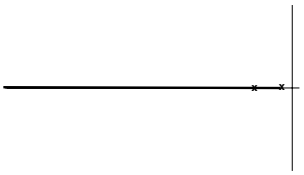
(a) $\mathrm{Re}(\lambda)$ vs. k with $H_0 < 0$.



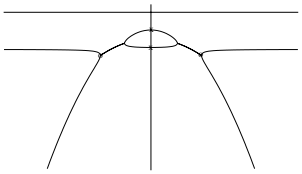
(b) $\mathrm{Re}(\lambda)$ vs. $\mathrm{Im}(\lambda)$ with $H_0 < 0$.



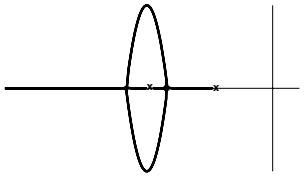
(c) $H_0 = 0$.



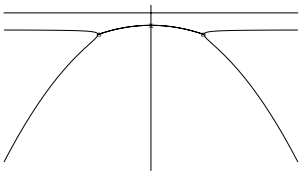
(d) $H_0 = 0$.



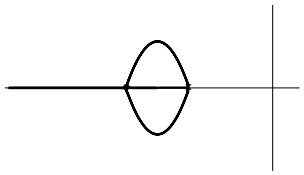
(e) $H_0 \in (0, (\sqrt{2\tau} - \sqrt{-G_1})^2)$.



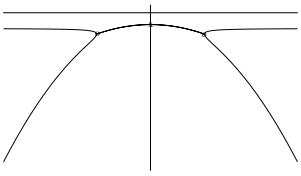
(f) $H_0 \in (0, (\sqrt{2\tau} - \sqrt{-G_1})^2)$.



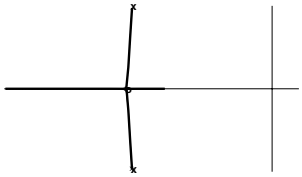
(g) $H_0 = (\sqrt{2\tau} - \sqrt{-G_1})^2$.



(h) $H_0 = (\sqrt{2\tau} - \sqrt{-G_1})^2$.



(i) $H_0 \in ((\sqrt{2\tau} - \sqrt{-G_1})^2, 2\tau - G_1)$.



(j) $H_0 \in ((\sqrt{2\tau} - \sqrt{-G_1})^2, 2\tau - G_1)$.

FIG. 3.1. *The five possible different structures of the stable essential spectrum. On the left we plot* $\mathrm{Re}(\lambda)$ *vs.* k *and on the right* $\mathrm{Re}(\lambda)$ *vs.* $\mathrm{Im}(\lambda)$.

fronts for which $\lim_{\xi\to\pm\infty} U_h(\xi;\varepsilon) = \pm 1$. Thus, we do not explicitly consider their symmetric counterparts. Due to the symmetry (1.2) this is, of course, also not necessary. The coupled system of second order equations (3.3) is equivalent to a linear system in $\mathbb{C}^4$,

$$\phi_\xi = A(\xi;\lambda,\varepsilon)\phi \ \text{ with } \ \phi(\xi) = (u(\xi), p(\xi), v(\xi), q(\xi)), \tag{3.4}$$

where $A(\xi;\lambda,\varepsilon)$ is a 4×4 matrix with $\mathrm{Tr}(A(\xi;\lambda,\varepsilon)\equiv 0$, and $u_\xi = p$, $v_\xi = \varepsilon q$. It follows that

$$\lim_{\xi\to\pm\infty} A(\xi;\lambda,\varepsilon) \stackrel{\text{def}}{=} A^{\pm}_{\infty}(\lambda,\varepsilon) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 2+\lambda & 0 & \mp 1 & 0 \\ 0 & 0 & 0 & \varepsilon \\ \pm 2\varepsilon H_0 & 0 & -\varepsilon(H_0+G_1-\tau\lambda) & 0 \end{pmatrix}; \tag{3.5}$$

see (1.7). The matrices $A^{\pm}_{\infty}$ have the same set of eigenvalues $\Lambda_i(\lambda,\varepsilon)$, $i=1,2,3,4$,

$$\Lambda^2_{1,4}(\lambda,\varepsilon) = \lambda + 2 + \mathcal{O}(\varepsilon^2), \quad \Lambda^2_{2,3}(\lambda,\varepsilon) = \varepsilon^2 \frac{\tau\lambda^2 - \lambda(G_1+H_0-2\tau) - 2G_1}{\lambda+2} + \mathcal{O}(\varepsilon^4). \tag{3.6}$$

Note that both expansions break down as λ approaches -2 (see Remark 3.2). We define a branch cut such that for $z\in\mathbb{C}$ $\arg(\sqrt{z}) \in (-\frac{1}{2}\pi, \frac{1}{2}\pi]$ so that the Λ_i's can be ordered

$$\mathrm{Re}(\Lambda_4(\lambda,\varepsilon)) < \mathrm{Re}(\Lambda_3(\lambda,\varepsilon)) < 0 < \mathrm{Re}(\Lambda_2(\lambda,\varepsilon)) < \mathrm{Re}(\Lambda_1(\lambda,\varepsilon)). \tag{3.7}$$

This ordering, of course, breaks down if $\lambda\in\sigma_{\mathrm{ess}}$, the essential spectrum associated to (3.3)/(3.4), since σ_{ess} coincides with values of λ for which either $\mathrm{Re}(\Lambda_{1,4}(\lambda,\varepsilon)) = 0$ or $\mathrm{Re}(\Lambda_{2,3}(\lambda,\varepsilon)) = 0$ [11]; see also section 3.1. The eigenvectors $E^{\pm}_i(\varepsilon,\lambda)$ of the matrices $A^{\pm}_{\infty}(\varepsilon,\lambda)$ associated to $\Lambda_i(\lambda,\varepsilon)$ are given by

$$E^{\pm}_{1,4}(\varepsilon,\lambda) = \begin{pmatrix} 1 \\ \Lambda_{1,4}(\lambda,\varepsilon) \\ \mathcal{O}(\varepsilon^2) \\ \pm\frac{2H_0}{\Lambda_{1,4}(\lambda,\varepsilon)}\varepsilon + \mathcal{O}(\varepsilon^3) \end{pmatrix}, \quad E^{\pm}_{2,3}(\varepsilon,\lambda) = \begin{pmatrix} \pm\frac{1}{\lambda+2} + \mathcal{O}(\varepsilon^2) \\ \mathcal{O}(\varepsilon^2) \\ 1 \\ \frac{1}{\varepsilon}\Lambda_{2,3}(\lambda,\varepsilon) \end{pmatrix} \tag{3.8}$$

(for $\lambda+2 \gg \varepsilon$—see Remark 3.2).

Remark 3.2. The expansions (3.6) and (3.8) are valid only for $\lambda+2 \gg \varepsilon$. It is straightforward to check that $\Lambda^2_{1,4}(\lambda,\varepsilon) = \mathcal{O}(\varepsilon) = \Lambda^2_{2,3}(\lambda,\varepsilon)$ if $\lambda+2=\mathcal{O}(\varepsilon)$ and that, in general, when $\lambda+2=\mathcal{O}(\varepsilon^\sigma)$ for some $\sigma\in[0,1]$, $\Lambda^2_{1,4}(\lambda,\varepsilon) = \mathcal{O}(\varepsilon^\sigma)$ and $\Lambda^2_{2,3}(\lambda,\varepsilon) = \mathcal{O}(\varepsilon^{2-\sigma})$. Thus, $\Lambda_{1,4}$ cannot be assumed to be large/fast compared to $\Lambda_{2,3}$ if $\lambda+2=\mathcal{O}(\varepsilon)$. Since $\lambda=-2+\mathcal{O}(\varepsilon)$ is way into the stable region, we do not consider this degeneration further and assume throughout this paper that $|\Lambda_{2,3}| \ll |\Lambda_{1,4}|$.

3.3. The Evans function. The use of the Evans function in the analysis of linear systems associated to the stability of traveling waves is by now well established. Here, we give a brief exposition of the characteristics of the Evans function in reaction-diffusion systems. We refer the reader to [1, 18, 10, 4, 5] for the full analytic details of the statements in this section.

We define the complement of the essential spectrum by

$$\mathcal{C}_e = \mathbb{C}\backslash\sigma_{\mathrm{ess}}. \tag{3.9}$$

For $\lambda\in\mathcal{C}_e$ the ordering (3.7) holds so that we have the following lemma.

LEMMA 3.3. *For all $\lambda \in \mathcal{C}_e$ there exist two two-dimensional families of solutions $\Phi_-(\xi;\lambda,\varepsilon)$ and $\Phi_+(\xi;\lambda,\varepsilon)$ to* (3.4) *such that $\lim_{\xi\to\pm\infty}\phi_\pm(\xi;\lambda,\varepsilon) = (0,0,0,0)^t$ for all $\phi_\pm(\xi;\lambda,\varepsilon) \in \Phi_\pm(\xi;\lambda,\varepsilon)$; $\Phi_\pm(\xi;\lambda,\varepsilon)$ depend analytically on λ.*

An eigenfunction of (3.4) must be in the intersection of $\Phi_-(\xi;\lambda,\varepsilon)$ and $\Phi_+(\xi;\lambda,\varepsilon)$. We define the Evans function $\mathcal{D}(\lambda,\varepsilon)$ by

$$\mathcal{D}(\lambda,\varepsilon) = \det[\phi_1(\xi;\lambda,\varepsilon),\phi_2(\xi;\lambda,\varepsilon),\phi_3(\xi;\lambda,\varepsilon),\phi_4(\xi;\lambda,\varepsilon)], \tag{3.10}$$

where $\{\phi_1,\phi_2\}$ (resp., $\{\phi_3,\phi_4\}$) span the space $\Phi_-(\xi;\lambda,\varepsilon)$ (resp., $\Phi_+(\xi;\lambda,\varepsilon)$). Since $\mathrm{Tr}(A) \equiv 0$, it follows by Abel's theorem that $\mathcal{D}(\lambda,\varepsilon)$ is independent of ξ. Moreover, $\mathcal{D}(\lambda,\varepsilon) = 0$, by construction, at an eigenvalue, since an eigenfunction must be in $\Phi_+(\xi;\lambda,\varepsilon)\cap\Phi_+(\xi;\lambda,\varepsilon)$. The Evans function is analytic in $\lambda \in \mathcal{C}_e$, and its zeros correspond one-to-one with eigenvalues of (3.4), counting multiplicities [1, 18]. Of course, this definition does not determine $\mathcal{D}(\lambda)$ uniquely. However, this can be achieved by choosing $\phi_1(\xi)$ and $\phi_2(\xi)$ as follows.

LEMMA 3.4. *For all $\lambda \in \mathcal{C}_e$ there is a unique solution $\phi_1(\xi;\lambda,\varepsilon) \in \Phi_-(\xi;\lambda,\varepsilon)$ of* (3.4) *such that*

$$\lim_{\xi\to-\infty} \phi_1(\xi;\lambda,\varepsilon)e^{-\Lambda_1(\lambda,\varepsilon)\xi} = E_1^-(\lambda,\varepsilon);$$

see (3.6), (3.8). *There exists an analytic transmission function $t_1(\lambda,\varepsilon)$ such that*

$$\lim_{\xi\to\infty} \phi_1(\xi;\lambda,\varepsilon)e^{-\Lambda_1(\lambda,\varepsilon)\xi} = t_1(\lambda,\varepsilon)E_1^+(\lambda,\varepsilon).$$

For $\lambda \in \mathcal{C}_e$ such that $t_1(\lambda,\varepsilon) \neq 0$ there is a unique solution $\phi_2(\xi;\lambda,\varepsilon) \in \Phi_-(\xi;\lambda,\varepsilon)$ of (3.4)*, which is independent of $\phi_1(\xi;\lambda,\varepsilon)$, that satisfies*

$$\lim_{\xi\to-\infty} \phi_2(\xi;\lambda,\varepsilon)e^{-\Lambda_2(\lambda,\varepsilon)\xi} = E_2^-(\lambda,\varepsilon) \text{ and } \lim_{\xi\to\infty} \phi_2(\xi;\lambda,\varepsilon)e^{-\Lambda_1(\lambda,\varepsilon)\xi} = (0,0,0,0)^t.$$

There exists a second meromorphic transmission function $t_2(\lambda,\varepsilon)$ that is determined by

$$\lim_{\xi\to\infty} \phi_2(\xi;\lambda,\varepsilon)e^{-\Lambda_2(\lambda,\varepsilon)\xi} = t_2(\lambda,\varepsilon)E_2^+(\lambda,\varepsilon).$$

The solutions $\phi_{3,4}(\xi;\lambda,\varepsilon) \in \Phi_+(\xi;\lambda,\varepsilon)$ of (3.4) can be defined likewise. Since $\sum_{i=1}^4 \Lambda_i(\lambda,\varepsilon) \equiv 0$ (3.6),

$$\mathcal{D}(\lambda,\varepsilon) = \det[\phi_1(\xi)e^{-\Lambda_1\xi},\phi_2(\xi)e^{-\Lambda_2\xi},\phi_3(\xi)e^{-\Lambda_3\xi},\phi_4(\xi)e^{-\Lambda_4\xi}]$$

so that $\mathcal{D}(\lambda,\varepsilon)$ can be decomposed into a product of $t_1(\lambda,\varepsilon)$ and $t_2(\lambda,\varepsilon)$ by taking the limit $\xi \to +\infty$.

LEMMA 3.5. *Let $\lambda \in \mathcal{C}_e$; then*

$$\mathcal{D}(\lambda,\varepsilon) = t_1(\lambda,\varepsilon)t_2(\lambda,\varepsilon)\det\left[E_1^+(\lambda,\varepsilon),E_2^+(\lambda,\varepsilon),E_3^+(\lambda,\varepsilon),E_4^+(\lambda,\varepsilon)\right]. \tag{3.11}$$

We conclude that the eigenvalues of (3.4) correspond to zeros of the transmission functions $t_1(\lambda,\varepsilon)$ and $t_2(\lambda,\varepsilon)$. However, we will see that a zero of $t_1(\lambda,\varepsilon)$ does not necessarily correspond to a zero of $\mathcal{D}(\lambda,\varepsilon)$, since $t_2(\lambda,\varepsilon)$ can have poles (see section 4.1 and [4, 5]).

3.4. The fast eigenvalues. The next section will be devoted to the analysis of (the zeros of) $t_2(\lambda,\varepsilon)$; here we consider the zeros of the fast transmission function $t_1(\lambda,\varepsilon)$. In order to do so, we first consider the stability problem associated to the front solution $U_f(\xi;V_0)$, with $U_f(\xi;V_0)\to\pm\sqrt{1+V_0}$ as $\xi\to\pm\infty$, of the scalar fast reduced limit problem (1.5),

$$w_{\xi\xi}+(1+V_0-3u_0^2(\xi;V_0)-\lambda)w=0, \tag{3.12}$$

since $U_f(\xi;V_0)=u_0(\xi;V_0)$ (2.3). This system can be written as a linear system in $\mathbb{C}^2$,

$$\psi_\xi=B(\xi;\lambda)\psi \ \text{ with } \ \psi(\xi)=(u(\xi),p(\xi)), \tag{3.13}$$

where $B(\xi;\lambda)$ is a 2×2 matrix of which the coefficients are by construction $\mathcal{O}(\varepsilon)$ close (uniformly in ξ) to those of the 2×2 block in the upper left corner of the 4×4 matrix $A^{\pm}(\xi;\lambda,\varepsilon)$ defined in (3.4), if we set $V_0=V_h(0)$. The Evans function associated to this problem can be written as $\mathcal{D}_f(\lambda)=\det[\psi_1(\xi,\lambda),\psi_4(\xi,\lambda)]$, in which $\psi_1(\xi)$ and $\psi_4(\xi)$ are solutions of (3.4) determined by $\lim_{\xi\to-\infty}\psi_1(\xi)e^{-\sqrt{\lambda+2}\xi}=(1,\sqrt{\lambda+2})^t$ and $\lim_{\xi\to\infty}\psi_4(\xi)e^{\sqrt{\lambda+2}\xi}=(1,-\sqrt{\lambda+2})^t$ (where $\pm\sqrt{\lambda+2}$ and $(1,\pm\sqrt{\lambda+2})^t$ are the eigenvalues and eigenvectors of the matrix $B_\infty(\lambda)=\lim_{\xi\to\pm\infty}B(\xi;\lambda)$ (compare to (3.6), (3.8))). As for the full system, we can define an analytic fast reduced transmission function $t_f(\lambda)$ by $\lim_{\xi\to\infty}\psi_1(\xi)e^{-\sqrt{\lambda+2}\xi}=t_f(\lambda)(1,\sqrt{\lambda+2})^t$ so that

$$\mathcal{D}_f(\lambda)=\lim_{\xi\to\infty}\det[\psi_1(\xi),\psi_4(\xi)]=\det[t_f(\lambda)(1,\sqrt{\lambda+2})^t,(1,-\sqrt{\lambda+2})^t]=-2t_f(\lambda)\sqrt{\lambda+2}.$$

The transmission function $t_1(\lambda)$ is, by construction, asymptotically close to its fast reduced limit $t_f(\lambda)$.

LEMMA 3.6. *For all $\lambda_i^f\in\mathcal{C}_e$ such that $t_f(\lambda_i^f)=0$, there is a uniquely determined $\lambda_i(\varepsilon)$ with $\lim_{\varepsilon\to0}\lambda_i(\varepsilon)=\lambda_i^f$ such that $t_1(\lambda_i(\varepsilon),\varepsilon)=0$; $t_1(\lambda,\varepsilon)\neq0$ for $\lambda\neq\lambda_i(\varepsilon)$.*

The (quite technical) proof of this lemma is analogous to the proofs of similar statements in [1, 10, 4, 5] and is based on the "elephant trunk" procedure [1, 10]—see especially the proof of Corollary 3.9 (and thus of Theorem 3.7) in [4] for a complete and detailed analysis.

Thus, by Lemma 3.6, we can find (the leading order behavior of) the zeros of $t_1(\lambda,\varepsilon)$ by computing the spectrum of (3.12). By (2.3) and by introducing $\eta=\sqrt{\frac{1}{2}(1+V_0)}$ we can write (3.12) as

$$w_{\eta\eta}+\left(\frac{6}{\cosh^2\eta}-P^2\right)w=0 \ \text{ with } \ P^2=\frac{2\lambda}{1+V_0}+4,$$

which is a well-studied problem of Schrödinger/Sturm–Liouville type (see, for instance, [21, 5]). It has discrete spectrum at $P=1$ and $P=4$ and essential spectrum for $P\in i\mathbb{R}$. We conclude that the eigenvalues of (3.12), and thus the leading order approximations of the zeros of $t_1(\lambda)$, are given by

$$\lambda_1^f=0,\quad \lambda_2^f=-\frac{3}{2}(1+V_0)<0. \tag{3.14}$$

The essential spectrum of (3.12) is given by

$$\sigma_{\text{ess}}^f=\{\lambda\le-2(1+V_0)\}. \tag{3.15}$$

We conclude this subsection by stating two simple, but useful results.

LEMMA 3.7. *Let $(u(\xi;\varepsilon), v(\xi;\varepsilon))$ be a pair of eigenfunction solutions of (3.3) associated to a simple eigenvalue $\lambda(\varepsilon)$; then either $u(\xi)$ is even as a function of ξ and $v(\xi)$ odd, or $u(\xi)$ is odd and $v(\xi)$ even.*

Proof. We write (3.3) in the following way:

$$v_{\xi\xi} = \varepsilon^2[F_o(\xi)u + F_e(\xi)v]\,. \tag{3.16}$$

By construction, U_h is an odd function of ξ and V_h is an even function of ξ. It thus follows that the above functions, F_o and F_e, must be odd and even functions of ξ, respectively. Let (u, v) be an eigenfunction associated to the eigenvalue λ. We decompose (u, v) into odd and even components, $u = u_o + u_e$, $v = v_o + v_e$, where u_o, v_o are odd and u_e, v_e are even. By the parity of the functions U_h, V_h, F_o, and F_e it is clear that (u_o, v_e) and (u_e, v_o) form two independent solutions of the eigenvalue problem associated to the eigenvalue λ. Since we have assumed that λ is simple, we have a contradiction. ∎

LEMMA 3.8. *Assume that the eigenfunction solution $v(\xi)$ of (3.3) with eigenvalue $\lambda(\varepsilon)$ is odd; then $\lambda(\varepsilon) \equiv 0$ so that $(u(\xi), v(\xi)) = (U_{h,\xi}(\xi;\varepsilon), V_{h,\xi}(\xi;\varepsilon))$.*

We will see in section 4 that there can be several eigenvalues for which $u(\xi)$ is odd and $v(\xi)$ even.

Proof. It is clear that there is an eigenvalue $\lambda = 0$ associated to the derivative of the front $(u(\xi), v(\xi)) = (U_{h,\xi}(\xi;\varepsilon), V_{h,\xi}(\xi;\varepsilon))$. We assume there is another eigenfunction with v odd. Since $v_{\xi\xi}$ is $\mathcal{O}(\varepsilon^2)$ and v is odd, it follows that $|v| \ll 1$ on the fast spatial scale. Hence, the equation for the u-component is to leading order homogeneous and given by (3.12) (with w replaced by u). Lemma 3.7 implies that u is even. Since the only even eigenfunction of (3.12) is $U_{h,\xi}$ with eigenvalue 0, it follows that the leading order behavior of u is given by $U_{h,\xi}$ and that λ is asymptotically close to 0. We thus write

$$u = U_{h,\xi} + \delta(\varepsilon)u_1\,, \quad v = V_{h,\xi} + \delta(\varepsilon)v_1\,, \quad \lambda = \delta(\varepsilon)\hat{\lambda}(\epsilon)\,, \tag{3.17}$$

where $\delta(\varepsilon) \to 0$ as $\varepsilon \to 0$ and $\hat{\lambda}(0) \neq 0$ (i.e., $\hat{\lambda}(\varepsilon) = \mathcal{O}(1)$, $\delta(\varepsilon)$ represents the leading order magnitude of λ). We substitute (3.17) into (3.3) to get the following equation for u_1:

$$u_{1,\xi\xi} + (1 - 3U_h^2)u_1 = \hat{\lambda}U_{h,\xi} - U_h v_1\,.$$

Note that this equation implies that u_1 is $\mathcal{O}(1)$, i.e., that the scaling chosen for $u - U_{h,\xi}$ in (3.17) is indeed the correct one. This equation has the solvability condition, $\int_{-\infty}^{\infty}(\hat{\lambda}U_{h,\xi} - U_h v_1)U_{h,\xi}\,dy = 0$. Now, since also $v_{1,\xi\xi}$ is $\mathcal{O}(\varepsilon^2)$ (3.3) and odd, it again follows that $|v_1| \ll 1$ on the fast spatial scale. Thus, we observe by the fast exponential decay of $U_h(\xi)$ that $\int_{-\infty}^{\infty} U_h U_{h,\xi} v_1\,dy \to 0$ as $\varepsilon \to 0$. Hence, we conclude from the solvability condition that $\lim_{\varepsilon\to 0}\hat{\lambda}(\varepsilon) = 0$, which contradicts the assumption that $\hat{\lambda}(\varepsilon) = \mathcal{O}(1)$. So the only possible eigenfunctions with v odd must correspond to a 0 eigenvalue, and hence $(u, v) = (U_{h,\xi}, V_{h,\xi})$. ∎

4. Slow-fast eigenvalues and edge bifurcations. The "slow-fast eigenvalues" are the eigenvalues that exist due to the interaction of the fast U-equation and the slow V-equation in (1.6); thus, these eigenvalues do not have a counterpart in the fast reduced scalar limit problem (1.5). The slow-fast eigenvalues correspond to the zeros of the $t_2(\lambda;\varepsilon)$, since this transmission function is based on a balance between slow and fast effects. See also Remark 4.5.

In order to study the combined effect of slow and fast terms, we need to define the region in which the fast ξ-jump takes place more accurately:

$$I_f = \left\{ \xi \in \left(-\frac{1}{\sqrt{\varepsilon}}, \frac{1}{\sqrt{\varepsilon}} \right) \right\} \text{ or } \{x \in (-\sqrt{\varepsilon}, \sqrt{\varepsilon})\}; \tag{4.1}$$

see (1.3). Note that the exact choice of the boundaries of I_f is not relevant; any choice will be suitable as long as it is in the transition zone between x and ξ (i.e., on the boundary of I_f we must have $|x| \ll 1$ and $|\xi| \gg 1$).

4.1. The regular case. Again, we first consider the case $G_1 = \mathcal{O}(1)$ (1.7). In the slow coordinate x (1.3), i.e., outside the region I_f, the equation for u reads

$$(1 - 3U_h^2 - \lambda + \mathcal{O}(\varepsilon))u = -U_h v + \mathcal{O}(\varepsilon^2 u_{xx}) \tag{4.2}$$

(see (3.3)), since $V_h(\xi) = \mathcal{O}(\varepsilon)$ on $\mathbb{R}$ (Theorem 2.1). Thus, u can be expressed in terms of v outside the fast ξ-region I_f (4.1). Using that $U_h^2(\xi;\varepsilon) = 1 + \mathcal{O}(\varepsilon)$ outside I_f (Theorem 2.1), we find for the v-equation of (3.3) on the slow x-scale

$$\begin{aligned} v_{xx} &= \left[2H(1,0)U_h + \mathcal{O}(\varepsilon)\right] u - \left[H(1,0) + \tfrac{\partial G}{\partial V}(0) - \tau\lambda + \mathcal{O}(\varepsilon)\right] v \\ &= \left[\tfrac{2H_0}{\lambda+2} - H_0 - G_1 + \tau\lambda + \mathcal{O}(\varepsilon)\right] v + \mathcal{O}(\varepsilon^2 v_{xx}); \end{aligned}$$

see (1.7). Hence, outside I_f,

$$v_{xx} = \left[\frac{-H_0\lambda + \lambda(\lambda+2)\tau - G_1(\lambda+2)}{\lambda+2} + \mathcal{O}(\varepsilon) \right] v, \tag{4.3}$$

uniformly in ξ. The v-equation is thus at leading order of constant coefficients type. By (3.1) and (3.6) we have on the ξ-scale

$$v_{\xi\xi} = \left[\frac{Q(\lambda;0)}{\lambda+2} + \mathcal{O}(\varepsilon^3) \right], \quad v = \left[\Lambda_{2,3}^2(\lambda,\varepsilon) + \mathcal{O}(\varepsilon^3)\right] v. \tag{4.4}$$

In order to determine an expression for $t_2(\lambda,\varepsilon)$, we need to control the solution $\phi_2(\xi;\lambda,\varepsilon)$ (Lemma 3.4) of (3.4). This is done in the following lemma.

LEMMA 4.1. *For all $\lambda \in \mathcal{C}_e$ such that $t_1(\lambda,\varepsilon) \neq 0$ there exist $\mathcal{O}(1)$ constants $C_-, C_+ > 0$ and a third meromorphic transmission function $t_3(\lambda,\varepsilon)$ such that*

$$\phi_2(\xi;\lambda,\varepsilon) = \begin{cases} \left[E_2^-(\lambda) + \mathcal{O}(\varepsilon)\right] e^{\Lambda_2(\lambda)\xi} + \mathcal{O}(e^{C_-\xi}) & \text{for} \quad \xi < -\frac{1}{\sqrt{\varepsilon}}, \\ t_2(\lambda)E_2^+(\lambda)e^{\Lambda_2(\lambda)\xi} + t_3(\lambda)E_3^+(\lambda)e^{\Lambda_3(\lambda)\xi} + \mathcal{O}(e^{-C_+\xi}) & \text{for} \quad \xi > \frac{1}{\sqrt{\varepsilon}}. \end{cases} \tag{4.5}$$

Moreover, there exists an $\mathcal{O}(1)$ constant C_f such that $||\phi_2(\xi)|| \leq C_f$ for $\xi \in I_f$. The v-coordinate of $\phi_2(\xi)$ satisfies $v(\xi) = 1 + \mathcal{O}(\sqrt{\varepsilon})$ on I_f so that

$$t_2(\lambda,\varepsilon) + t_3(\lambda,\varepsilon) = 1 + \mathcal{O}(\sqrt{\varepsilon}). \tag{4.6}$$

Proof. It follows from the above analysis that the (leading order) behavior of $\phi_2(\xi)$ outside I_f is determined by equations with constant coefficients. In other words, outside I_f, the matrix $A(\xi;\lambda,\varepsilon)$ of (3.4) can be approximated the constant coefficients matrix $A_\infty^\pm(\lambda,\varepsilon)$ of (3.5). Thus, the approximation (4.5) for $\xi < -1/\sqrt{\varepsilon}$ follows from the definition of $\phi_2(\xi)$ (i.e., the (boundary) conditions on $\phi_2(\xi)$ as $\xi \to -\infty$; see

Lemma 3.4). This same lemma establishes the leading order term in (4.5) for $\xi \to \infty$. The transmission function $t_3(\lambda, \varepsilon)$ measures the component of $\phi_2(\xi)$ that decays on the slow spatial scale x. Inside I_f, $v_{\xi\xi} = \mathcal{O}(\varepsilon^2)$ (3.3) and $\Lambda^2_{2,3}(\lambda, \varepsilon) = \mathcal{O}(\varepsilon^2)$ (3.6) so that (4.6) follows. As in section 3.3 we refrain from giving the full analytic details of this result, since these are essentially the same as in [10, 4, 5]. ∎

The transmission function $t_2(\lambda, \varepsilon)$ can be determined by the methods originally developed in [3]. We deduce from Lemma 4.1 and (4.4) that the total change in v_ξ over I_f is given by

$$\Delta_{\text{slow}} v_\xi = 2\varepsilon(t_2(\lambda) - 1)\sqrt{\frac{\tilde{Q}(\lambda; 0)}{\lambda + 2}} + \mathcal{O}(\varepsilon\sqrt{\varepsilon}), \tag{4.7}$$

where $\tilde{Q}(\lambda; 0) = \mathcal{O}(1)$ is defined by $Q(\lambda; 0) = \varepsilon\tilde{Q}(\lambda; 0)$ (3.1). This change in v_ξ must be an effect of the evolution on the fast ξ-scale, that is, given by

$$\Delta_{\text{fast}} v_\xi = \int_{-\frac{1}{\sqrt{\varepsilon}}}^{\frac{1}{\sqrt{\varepsilon}}} v_{\xi\xi}|_{\{u=u_{\text{in}}, v=1\}} d\xi + \mathcal{O}(\varepsilon^2\sqrt{\varepsilon}), \tag{4.8}$$

where $u_{\text{in}}(\xi; \lambda)$ is a bounded solution of the inhomogeneous problem

$$u_{\xi\xi} + (1 - 3U_h^2(\xi; 0) - \lambda)u = -U_h(\xi; 0) \tag{4.9}$$

(recall that $v(\xi) = 1 + \mathcal{O}(\sqrt{\varepsilon})$ in I_f). The transmission function $t_2(\lambda; \varepsilon)$ is determined by combining (4.7) and (4.8). Since, a priori $\Delta_{\text{slow}} v_\xi = \mathcal{O}(\varepsilon)$ and $\Delta_{\text{fast}} v_\xi = \mathcal{O}(\varepsilon\sqrt{\varepsilon})$ we are led to the conclusion that $t_2(\lambda) - 1$ must be $\mathcal{O}(\sqrt{\varepsilon})$ for λ not close to a zero of $\Delta_{\text{slow}} v_\xi$ or a singularity of $\Delta_{\text{fast}} v_\xi$.

LEMMA 4.2. *Consider $\lambda \in \mathcal{C}_e \cap \{\text{Re}(\lambda) > -2 + \delta\}$ for some $\delta > 0$ independent of ε. Let $\lambda_2^f = -\frac{3}{2}$ be the second eigenvalue of the limit system* (3.12) *with $V_0 = 0$* (3.14)*, and let $\lambda^+(0)$ and $\lambda^-(0)$ be the solutions of $Q(\lambda, 0) = 0$* (3.1)*. Then $t_2(\lambda) = 1 + \mathcal{O}(\sqrt{\varepsilon})$ if $|\lambda - \lambda_2^f|, |\lambda - \lambda^+(0)|, |\lambda - \lambda^-(0)| = \mathcal{O}(1)$; $t_2(\lambda) = 1 + \mathcal{O}(\varepsilon^{\frac{1}{2}-\sigma})$ if $|\lambda - \lambda_2^f| = \mathcal{O}(\varepsilon^\sigma)$, $|\lambda - \lambda^+(0)| = \mathcal{O}(\varepsilon^{2\sigma})$, or $|\lambda - \lambda^-(0)| = \mathcal{O}(\varepsilon^{2\sigma})$ for some $\sigma \in (0, \frac{1}{2})$.*

Thus, this lemma establishes that $t_2(\lambda, \varepsilon)$ can only be zero in $\{\text{Re}(\lambda) > -2\}$ if $\lambda \in \mathcal{C}_e$ is $\mathcal{O}(\sqrt{\varepsilon})$ close to λ_2^f or $\mathcal{O}(\varepsilon)$ close to $\lambda^+(0)$ or $\lambda^-(0)$, so we have to study only λ near these three points to determine the slow-fast eigenvalues of (3.4). Note that the fast reduced (scalar) limit problem has an eigenvalue $\lambda_2^f = -\frac{3}{2}$ ((3.14), since $V_0 = V_h(0) \to 0$ as $\varepsilon \to 0$ (Theorem 2.1)). We will prove below that $t_2(\lambda)$ has a (simple) zero close to λ_2^f, i.e., that the fast reduced eigenvalue λ_2^f persists.

However, before going further into the details of the (possible) existence of eigenvalues near λ_2^f, $\lambda^+(0)$ or $\lambda^-(0)$, we formulate a result that is an immediate consequence of Lemma 4.2 and that establishes the stability of the wave for values of G_1 and H_0 such that the essential spectrum, and hence $\lambda^+(0)$ and $\lambda^-(0)$, is in the negative half-plane and not too close to the imaginary axis (see Lemma 3.1).

THEOREM 4.3. *Let $\varepsilon > 0$ be small enough, and let $G_1 < 0$ and $H_0 + G_1 - 2\tau < 0$ be such that $|G_1|, |H_0 + G_1 - 2\tau| \gg \varepsilon$. The spectrum of the eigenvalue problem* (3.3) *associated to the stability of the solution $(U_h(\xi; \varepsilon), V_h(\xi; \varepsilon))$ consists of a (simple) eigenvalue at $\lambda = 0$ and a part that is embedded in the region $\{\text{Re}(\lambda) < -\varepsilon\}$. Therefore, $(U_h(\xi; \varepsilon), V_h(\xi; \varepsilon))$ is (spectrally) stable.*

Note that the operator defined by (3.3) is clearly sectorial in this case (see section 3.1) so that the nonlinear (orbital) stability of $(U_h(\xi; \varepsilon), V_h(\xi; \varepsilon))$ follows by standard arguments (see, for instance, [11]).

Proof of Lemma 4.2. We first note that indeed $\Delta_{\text{fast}} v_\xi = \mathcal{O}(\varepsilon\sqrt{\varepsilon})$ and $\Delta_{\text{slow}} v_\xi = \mathcal{O}(\varepsilon)$, and thus $t_2(\lambda) = 1 + \mathcal{O}(\sqrt{\varepsilon})$, for $\lambda \in \mathcal{C}_e$ that are not asymptotically close to the possible degenerations of (4.8) and (4.7).

The solution $u_{\text{in}}(\xi;\lambda)$ of the inhomogeneous problem (4.9) may become unbounded as λ approaches an eigenvalue, $\lambda_1^f = 0$ or $\lambda_2^f = -\frac{3}{2}$, or the essential spectrum σ_{ess}^f (3.15) of the linear problem associated to the fast reduced limit (3.12) with $V_0 = 0$ (i.e., the homogeneous part of (4.9)). To avoid irrelevant technicalities near σ_{ess}^f we assume that $\lambda \in \mathcal{C}_e \cap \{\text{Re}(\lambda) > -2 + \delta\}$. The eigenfunction associated to λ_1^f, i.e., $U_{h,\xi}(\xi;0)$, is odd, which implies that the inhomogeneous (and even) term $U_h(\xi;0)$ satisfies the solvability condition associated to (4.9) at $\lambda = 0$. Hence, $u_{\text{in}}(\xi;\lambda)$ remains bounded as $\lambda \to 0$ so that $t_2(\lambda) = 1 + \mathcal{O}(\sqrt{\varepsilon})$ also near $\lambda = 0$ [4, 5]. The eigenfunction associated to the second eigenvalue of the homogeneous part of equation (4.9), λ_2^f, is even; thus, the solution $u_{\text{in}}(\xi;\lambda)$ of (4.9) grows as $1/(\lambda_2^f - \lambda)$ as $\lambda \to \lambda_2^f$ [21, 4, 5]. Since u_{in} appears in $\Delta_{\text{fast}} v_\xi$ (4.8), we conclude that $t_2(\lambda,\varepsilon) - 1 = \mathcal{O}(\varepsilon^{\frac{1}{2}-\sigma})$ if $|\lambda - \lambda_2^f| = \mathcal{O}(\varepsilon^\sigma)$ for some $\sigma \in (0, \frac{1}{2})$.

The behavior of $t_2(\lambda)$ near the degenerations of (4.7), i.e., the zeros $\lambda^\pm(0)$ of $Q(\lambda;0)$, follows from observing that $\Delta_{\text{slow}} v_\xi = (t_2 - 1) \times \mathcal{O}(\varepsilon^{1+\sigma})$ if λ is $\mathcal{O}(\varepsilon^{2\sigma})$ close to $\lambda^+(0)$ or to $\lambda^-(0)$ for some $\sigma \in (0, \frac{1}{2})$. ∎

Although it is not essential for the forthcoming analysis, we note that we may also conclude from the proof of this lemma that $t_2(\lambda)$ indeed has a (simple) pole that approaches λ_2^f as $\varepsilon \to 0$. This pole is generated by the solution $u_{\text{in}}(\xi;\lambda)$ of the inhomogeneous problem (4.9) that appears in the expression for $\Delta_{\text{fast}} v_\xi$ (4.8). This solution necessarily develops a singularity near λ_2^f, an eigenvalue of the homogeneous part of (4.9). Since the Evans function $\mathcal{D}(\lambda)$ is analytic as function of λ, it follows that $t_2(\lambda)$ can only have its pole exactly at the zero $\lambda_2(\varepsilon)$ of $t_1(\lambda)$ (that is also asymptotic to λ_2^f—see Lemma 3.6). Note that this is fully consistent with Lemma 3.4, in which the existence of $t_2(\lambda)$ can only be proved for λ such that $t_1(\lambda) \neq 0$. Nevertheless, it can be shown, by a (standard) winding number argument [1, 4, 5], that the eigenvalue λ_2^f persists as an eigenvalue of the full system (3.3) if it is not embedded in the essential spectrum.

LEMMA 4.4. *Let G_1 and H_0 be such that σ_{ess} does not intersect an $\mathcal{O}(\varepsilon^\sigma)$ neighborhood of λ_2^f for some $\sigma < \frac{1}{2}$. Then there is an eigenvalue $\lambda_2(\varepsilon)$ of* (3.3) *with* $\lim_{\varepsilon\to 0} \lambda_2(\varepsilon) = \lambda_2^f = -\frac{3}{2}$.

Proof. By the assumptions in the lemma, there exists a contour K in the complex λ-plane, which does not intersect σ_{ess}, that encircles an $\mathcal{O}(\varepsilon^\sigma)$ neighborhood of λ_2^f and that is $\mathcal{O}(\varepsilon^\sigma)$ close to λ_2^f. It follows from Lemma 4.2 that $t_2(\lambda) = 1 + \mathcal{O}(\varepsilon^{\frac{1}{2}-\sigma})$ for $\lambda \in K$; thus, the winding number of $t_2(\lambda)$ over K is 0. However, $t_2(\lambda)$ must have a (simple) pole in the interior of K, as observed above. We conclude that $t_2(\lambda)$ must also have a (simple, real) zero in the interior of K. ∎

The possible existence of slow-fast eigenvalues near $\lambda^+(0)$ or $\lambda^-(0)$ is much more subtle. Since such eigenvalues become relevant only to the stability of the solution $(U_h(\xi;\varepsilon), V_h(\xi;\varepsilon))$ as G_1 (or $H_0 + G_1 - 2\tau$) approaches 0 (Theorem 4.3) we will consider this issue in forthcoming sections.

Remark 4.5. The eigenvalues $\lambda_1(\varepsilon) = 0$ and $\lambda_2(\varepsilon) \to -\frac{3}{2}$ as $\varepsilon \to 0$ can be interpreted as "fast" eigenvalues, since they correspond to eigenvalues of the fast reduced limit problem. However, strictly speaking, both eigenvalues also have the slow-fast structure described in the beginning of this section.

First, we of course know that $\lambda_1(\varepsilon) = 0$ is an eigenvalue—see also Lemma 3.8. Thus, it is a zero of $\mathcal{D}(\lambda,\varepsilon)$. Since $t_2(\lambda) = 1 + \mathcal{O}(\sqrt{\varepsilon})$ for λ near 0 (see the proof

of Lemma 4.2), we conclude that $t_1(0;\varepsilon) \equiv 0$ (note that this (in a sense) obvious result does not follow directly from Lemma 3.6). Thus, the solution $\phi_1(\xi;0,\varepsilon)$ of (3.4) that by construction has a purely fast structure for $\xi \ll -1$ does not blow up as $e^{\Lambda_1(0,\varepsilon)\xi}$ as $\xi \to \infty$ (Lemma 3.4). Nevertheless, the eigenfunction associated to $\lambda = 0$, $(U_{h,\xi}(\xi), V_{h,\xi}(\xi))$, has a clear slow-fast structure that it inherits from $(U_h(\xi), V_h(\xi))$ (Theorem 2.1). Hence, $\phi_1(\xi;0,\varepsilon)$ is not the eigenfunction associated to $\lambda = 0$. Neither is $\phi_2(\xi;0,\varepsilon)$, since $t_2(0) \neq 0$. It follows that the eigenfunction associated to $\lambda = 0$ must be a linear combination of $\phi_1(\xi;0,\varepsilon)$ and $\phi_2(\xi;0,\varepsilon)$, and thus that $\phi_1(\xi;0,\varepsilon)$ does not decay as $\xi \to \infty$, but instead grows exponentially (and slowly), as $e^{\Lambda_2(0,\varepsilon)\xi}$ (like $\phi_2(\xi;0,\varepsilon)$). The linear combination is such that the two growth terms $e^{\Lambda_2(0,\varepsilon)\xi}$ (for $\xi \to \infty$) cancel.

Second, $\lambda_2(\varepsilon)$ is not a zero of $t_1(\lambda)$, although it is asymptotically close to such a zero, but it is a zero of $t_2(\lambda)$. Thus, $\phi_2(\xi;\lambda_2(\varepsilon),\varepsilon)$ is the eigenfunction of (3.4) at $\lambda = \lambda_2(\varepsilon)$ (and $\phi_1(\xi;\lambda_2(\varepsilon),\varepsilon)$ blows up fast, as $e^{\Lambda_1(\lambda_2(\varepsilon),\varepsilon)\xi}$).

4.2. The superslow case: An example. In the previous section we have seen that the front might destabilize as G_1 approaches 0 (if we assume that $H_0 + G_1 - 2\tau < 0$). In this case, Theorem 2.1 can no longer be used to establish the existence of the front $(U_h(\xi), V_h(\xi))$. Thus, the question about the stability of the front is closely related to the characteristics of the existence problem (as is usual in the analysis of (traveling) waves; see also [14]). In this section we consider the bifurcation as G_1 approaches 0. Therefore, we assume that $H_0 - 2\tau < 0$ and $\mathcal{O}(1)$ with respect to ε. As in section 2 we consider in the superslow case the simplified system in which the general function $G(V)$ is replaced by a linear expression: $G(V) = G_1 V = -\varepsilon^2 \gamma V$ (see Remark 2.4). Note that Theorem 4.3 a priori predicts a possible destabilization as G_1 becomes $\mathcal{O}(\varepsilon)$, i.e., already before $G_1 = -\gamma\varepsilon^2$, but it will be shown in the next section that the estimate in Theorem 4.3 is not sharp, in the sense that a bifurcation occurs only as G_1 decreases to $\mathcal{O}(\varepsilon^2)$.

One of the main differences between the analysis in this section and that of the regular case is the fact that $V_h(\xi)$ is no longer $\mathcal{O}(\varepsilon)$; i.e., $V_h(\xi)$ does not contribute only to the higher order terms in the stability analysis of the front solutions. Nevertheless, we follow the approach of the previous section and express the solution u of (3.3) in terms of v, outside I_f (see (4.2)):

(4.10)
$$u = -\frac{U_h}{1 + V_h - 3U_h^2 - \lambda} v + \mathcal{O}(\varepsilon^2 u_{xx}) = \left[\frac{U_h}{2(1+V_h) + \lambda} + \mathcal{O}(\varepsilon^4) \right] v + \mathcal{O}(\varepsilon^2 v_{xx})$$

since $1 + V_h(\xi;\varepsilon) - U_h^2(\xi;\varepsilon) = \mathcal{O}(\varepsilon^4)$ (see (2.8); recall that q^2 and G are $\mathcal{O}(\varepsilon^2)$ in the superslow case). This yields that

(4.11)
$$\begin{aligned} v_{xx} &= \left\{ 2\left[H(U_h^2, V_h) + \mathcal{O}(\varepsilon^4) \right] U_h u - \left[H(U_h^2, V_h) + \mathcal{O}(\varepsilon^4) - \varepsilon^2\gamma - \tau\lambda \right] v \right\} \\ &= \left\{ \tfrac{2H(U_h^2,V_h)U_h^2}{2(1+V_h)+\lambda} - H(U_h^2, V_h) + \tau\lambda + \varepsilon^2\gamma + \mathcal{O}(\varepsilon^4) \right\} v + \mathcal{O}(\varepsilon^2 v_{xx}) \\ &= \left\{ \lambda \left[\tau - \tfrac{H(1+V_h,V_h)}{2(1+V_h)+\lambda} \right] + \varepsilon^2\gamma + \mathcal{O}(\varepsilon^4) \right\} v + \mathcal{O}(\varepsilon^2 v_{xx}). \end{aligned}$$

It follows from section 3.1 that one of the "tips" of σ_{ess}, $\lambda^+(0)$, is $\mathcal{O}(\varepsilon^2)$ if $G_1 = \mathcal{O}(\varepsilon^2)$ (and $H_0 - 2\tau < 0$), while the other one, $\lambda^-(0)$, is $\mathcal{O}(1)$ and negative (3.2). Thus, the destabilization of the front will be caused by either σ_{ess} at $G_1 = 0 = \gamma$ or possibly by

a slow-fast eigenvalue λ that is close to $\lambda^+(0)$ (Lemma 4.2). Therefore, we introduce $\tilde{\lambda}$ by

$$\lambda = \varepsilon^2 \tilde{\lambda}, \tag{4.12}$$

which implies that (4.11) can also be written as a superslow system,

$$v_{xx} = \varepsilon^2 \left\{ \tilde{\lambda} \left[\tau - \frac{H(1+V_h, V_h)}{2(1+V_h)} \right] + \gamma + \mathcal{O}(\varepsilon^2) \right\} v. \tag{4.13}$$

As in section 2.2, we first consider the explicit example in which $H(U^2, V) = H_0 U^2$. Thus, the existence of (several kinds of) front solutions is established by Theorem 2.3. In this case, the equation for v is, on the ξ-scale, given by

$$v_{\xi\xi} = \varepsilon^4 \left[\tilde{\lambda} \left(\tau - \frac{1}{2} H_0 \right) + \gamma + \mathcal{O}(\varepsilon^2) \right] v = \left[\Lambda_{2,3}^2(\lambda, \varepsilon) + \mathcal{O}(\varepsilon^6) \right] v; \tag{4.14}$$

see (3.6). Note that this equation is of constant coefficients type, and, at leading order, the same as in the equation for $v_{\xi\xi}$ in the regular case (4.4). Hence, we can copy the arguments leading to Lemma 4.1 and conclude that the fundamental solution $\phi_2(\xi; \varepsilon^2\tilde{\lambda}, \varepsilon)$ of (3.4) can again be expressed as in (4.5) outside the region I_f. Moreover, as in Lemma 4.1, we may conclude that $t_2(\tilde{\lambda}, \varepsilon) + t_3(\tilde{\lambda}, \varepsilon) = 1 + \mathcal{O}(\sqrt{\varepsilon})$ (4.6).

We may now proceed as in the preceding section (and as in [4, 5]) and determine $t_2(\tilde{\lambda})$ by measuring the change in the $q = v_\xi$-coordinate of $\phi_2(\xi)$ over the fast field. It follows from (4.14) that

$$\Delta_{\text{slow}} v_\xi = 2\varepsilon^2 (t_2(\tilde{\lambda}) - 1) \sqrt{\tilde{\lambda} \left(\tau - \frac{1}{2} H_0 \right) + \gamma} + \mathcal{O}(\varepsilon^2 \sqrt{\varepsilon}). \tag{4.15}$$

Note that we have to assume that $\tilde{\lambda}(\tau - \frac{1}{2}H_0) + \gamma > 0$, i.e., $\Lambda_{2,3}^2(\lambda, \varepsilon) > 0$, which is a natural assumption, since

$$\tilde{\lambda}_{\text{tip}} = \tilde{\lambda}^+(0) = -\frac{2\gamma}{2\tau - H_0} < 0 \tag{4.16}$$

determines the "tip" of σ_{ess} (recall that $H_0 - 2\tau < 0$); i.e., $t_2(\tilde{\lambda})$ is not defined if $\tilde{\lambda} \le \tilde{\lambda}_{\text{tip}}$. By definition, $\Delta_{\text{fast}} v_\xi$ is given by (4.8). Since, at leading order $V_h(\xi) = V_h(0) = v_0$ and $U_h(\xi) = u_0(\xi; v_0)$ (uniformly) in I_f(Theorem 2.3), and since $u_0(\xi; v_0)$ decays exponentially fast on the (fast) ξ-scale, it follows that

$$\Delta_{\text{fast}} v_\xi = \varepsilon^2 H_0 \int_{-\infty}^{\infty} \left\{ 2 \left[2u_0^2(\xi; v_0) - 1 - v_0 \right] u_0(\xi; v_0) u_{\text{in}}(\xi; v_0) - u_0^2(\xi; v_0) \right\} d\xi + \mathcal{O}(\varepsilon^2 \sqrt{\varepsilon}), \tag{4.17}$$

where $u_{\text{in}}(\xi; v_0) = u_{\text{in}}(\xi; \lambda = 0; v_0)$ is the (uniquely determined) bounded solution of the inhomogeneous problem

$$u_{\xi\xi} + (1 + v_0 - 3u_0^2(\xi; v_0))u = -u_0(\xi; v_0).$$

Since we already know one solution of the homogeneous problem, $u(\xi) = u_{0,\xi}(\xi; v_0)$, we can determine $u_{\text{in}}(\xi; v_0)$ explicitly:

$$u_{\text{in}}(\xi; v_0) = \frac{1}{2(1+v_0)} \left(u_0(\xi; v_0) + \xi u_{0,\xi}(\xi; v_0) \right). \tag{4.18}$$

Thus, by (2.3), $\Delta_{\text{fast}} v_\xi$ can be computed explicitly (at leading order):

$$\Delta_{\text{fast}} v_\xi = -\varepsilon^2 H_0 \sqrt{2}\sqrt{1+v_0} + \mathcal{O}(\varepsilon^2\sqrt{\varepsilon}).$$

Combining this with (4.15) yields an explicit expression for $t_2(\tilde{\lambda})$,

$$t_2(\tilde{\lambda},\varepsilon) = 1 - H_0 \sqrt{\frac{2(1+v_0)}{\tilde{\lambda}(\tau - \frac{1}{2}H_0) + \gamma}} + \mathcal{O}(\sqrt{\varepsilon}), \tag{4.19}$$

for $\tilde{\lambda} > \tilde{\lambda}_{\text{tip}}$ (4.16). It follows that $t_2(\tilde{\lambda}) \geq 1 + \mathcal{O}(\sqrt{\varepsilon})$ for $H_0 \leq 0$ and $t_2(\tilde{\lambda}) < 1 + \mathcal{O}(\sqrt{\varepsilon})$ for $H_0 > 0$. Hence, $t_2(\tilde{\lambda})$ cannot have zeros if $H_0 \leq 0$. In other words, there cannot be an eigenvalue near the tip of the essential spectrum in case (ii) of Theorem 2.3. On the other hand, $t_2(\tilde{\lambda})$ can be 0 for $H_0 > 0$; i.e., in case (i) of Theorem 2.3 there indeed is a "new" slow-fast eigenvalue of (3.3); it is given by

$$\lambda_{\text{edge}} = \varepsilon^2 \tilde{\lambda}_{\text{edge}} = \frac{-2\gamma + H_0^2(1+v_0)}{2\tau - H_0}\varepsilon^2 + \mathcal{O}(\varepsilon^2\sqrt{\varepsilon}) > \varepsilon^2 \tilde{\lambda}_{\text{tip}} = \lambda_{\text{tip}}; \tag{4.20}$$

see (4.16). Note that the eigenvalue λ_{edge} merges with λ_{tip} and thus with σ_{ess} as $H_0 \downarrow 0$. This is, of course, a leading order result; the accuracy of our analysis allows us only to conclude that $|\lambda_{\text{tip}} - \lambda_{\text{edge}}| \leq \mathcal{O}(\varepsilon^2\sqrt{\varepsilon})$ as $H_0 \downarrow 0$ and that λ_{edge} does not exist for $H_0 < 0$. Nevertheless, we conclude that λ_{edge} appears from the essential spectrum as H_0 increases through 0. In other words, λ_{edge} is created, or annihilated, by an edge bifurcation. Note that the new eigenvalue appears exactly as σ_{ess} becomes complex valued (see Figure 3.1).

The existence or nonexistence of λ_{edge} is crucial to the character of the destabilization (see also the numerical simulations in section 5). For $H_0 < 0$, the front solution $(U_h(\xi), V_h(\xi))$ destabilizes as γ, or equivalently G_1, crosses through 0. The destabilization is due to the essential spectrum, which implies that also the "background states" $(U(x,t), V(x,t)) \equiv (\pm 1, 0)$ destabilize at $\gamma = 0$. However, in the case $H_0 > 0$ the eigenvalue is λ_{edge} is $\varepsilon^2 H_0^2(1+v_0)/(2\tau - H_0)$ ahead of σ_{ess} (4.20), in the sense that it reaches the axis $\text{Re}(\lambda) = 0$ before σ_{ess} as $\gamma > 0$ decreases to 0. Thus, if $H_0 > 0$ the front solution $(U_h(\xi), V_h(\xi))$ destabilizes by an element of the discrete spectrum of (3.3) at $\gamma = \gamma_{\text{double}}$, defined as the solution of $\gamma = \frac{1}{2}H_0^2(1 + v_0(\gamma)) + \mathcal{O}(\sqrt{\varepsilon})$ (> 0). As a consequence, the background states $(\pm 1, 0)$ remain stable as $(U_h(\xi), V_h(\xi))$ destabilizes for $H_0 > 0$, contrary to the case $H_0 < 0$. The bifurcation at γ_{double} is associated to the saddle-node bifurcation of heteroclinic orbits described in Theorem 2.3.

THEOREM 4.6. *Assume that $G(V) = -\varepsilon^2\gamma V$, that $H(U^2, V) = H_0 U^2$, that $H_0 - 2\tau < 0$ and $\mathcal{O}(1)$, and that $\varepsilon > 0$ is small enough.*

(i) *Let $(U_h^{+,1}(\xi), V_h^{+,1}(\xi))$ and $(U_h^{+,2}(\xi), V_h^{+,2}(\xi))$ be the two types of heteroclinic front solutions that exist for $H_0 > 0$ and $\gamma \geq \gamma_{\text{double}} = \frac{3}{2}H_0^2 + \mathcal{O}(\sqrt{\varepsilon})$ with, at leading order, $0 < V_h^{+,1}(0) = v_1 \leq 2 \leq v_2 = V_h^{+,2}(0)$ (Theorem 2.3). The front solution $(U_h^{+,1}(\xi), V_h^{+,1}(\xi))$ is (nonlinearly) stable for $\gamma > \gamma_{\text{double}}$, and the front $(U_h^{+,2}(\xi), V_h^{+,2}(\xi))$ is unstable; $(U_h^{+,1}(\xi), V_h^{+,1}(\xi))$ destabilizes by an element of the discrete spectrum, λ_{edge}, at $\gamma = \gamma_{\text{double}}$ and merges with $(U_h^{+,2}(\xi), V_h^{+,2}(\xi))$ in a saddle-node bifurcation of heteroclinic orbits.*

(ii) *Let $(U_h^{+}(\xi), V_h^{+}(\xi))$ be a heteroclinic front solution that exists for $H_0 < 0$ and (all) $\gamma > 0$ (Theorem 2.3); $(U_h^{+}(\xi), V_h^{+}(\xi))$ is (nonlinearly) stable for all $\gamma > 0$; it is destabilized at $\gamma = 0$ by the essential spectrum σ_{ess}.*

Remark 4.7. As in the regular case, spectral stability implies nonlinear orbital stability in this superslow case, since the linear operator associated to the stability problem remains sectorial as long as $\varepsilon > 0$.

Proof of Theorem 4.6. We first note that the condition $H_0 - 2\tau < 0$ and $\mathcal{O}(1)$ determines that $\sigma_{\rm ess}$ can only cross, or come close to, the $\mathrm{Re}(\lambda) = 0$-axis at $\lambda = 0$ (Lemma 3.1 with $G_1 = \mathcal{O}(\varepsilon^2)$).

(i) The eigenvalue "in front of" $\sigma_{\rm ess}$, $\lambda^{1,2}_{\rm edge}(v_{1,2})$, is given by (4.20), where $v_0 > 0$ is a solution of $9\gamma v^2 = 2H_0^2(1+v)^3$, and $v_0 = v_1 \le 2$ (at leading order) for $(U_h^{+,1}(\xi), V_h^{+,1}(\xi))$, while $v_0 = v_2 \ge 2$ (at leading order) for $(U_h^{+,2}(\xi), V_h^{+,2}(\xi))$ (Theorem 2.3). Thus, by (4.20), $\lambda^1_{\rm edge}(v_1) < 0$ and $\lambda^2_{\rm edge}(v_2) > 0$ if $\gamma < \gamma_{\rm double} = \frac{3}{2}H_0^2 + \mathcal{O}(\sqrt{\varepsilon})$. As a consequence, $\lambda^1_{\rm edge}(v_1) \uparrow 0$ and $\lambda^2_{\rm edge}(v_2) \downarrow 0$ as $\gamma \downarrow \gamma_{\rm double}$, at which the saddle-node bifurcation takes place.

(ii) We have already shown that there can be no eigenvalues in front of the tip of $\sigma_{\rm ess}$. Therefore, the statement of the theorem follows. ∎

Remark 4.8. Since $t_2(\lambda) = 0$, the slow-fast eigenfunction associated to the bifurcation at $\gamma = \gamma_{\rm double}$ is given by $\phi_2(\xi)$. It follows from Lemmas 3.7 and 3.8 that the u-component of ϕ_2 is odd, and the v-component even, as functions of ξ.

4.3. Bifurcations in the general superslow problem. We now consider the stability of a front solution in the general superslow limit. Thus, we assume that we have established the existence of a front $(U_h(\xi), V_h(\xi))$ for a certain given function $H(U^2, V)$ (Theorem 2.5). To analyze its stability, we again try to determine $t_2(\lambda)$ by measuring $\Delta_{\rm fast} v_\xi$ and $\Delta_{\rm slow} v_\xi$.

In order to determine $\Delta_{\rm slow} v_\xi$ we follow the derivation of (4.13) in the previous section. Hence, we again conclude that nontrivial eigenvalues near 0 are possible only for $\lambda = \mathcal{O}(\varepsilon^2)$; thus, we again introduce $\tilde\lambda$ (4.12) (see also the proof of Theorem 4.10 for more details on the necessity of this scaling). Note that both G_1 and λ are now $\mathcal{O}(\varepsilon^2)$; thus, we can immediately obtain a leading order expression for $\Delta_{\rm fast} v_\xi$ in terms of $H(U^2, V)$,

$$\begin{aligned}\Delta_{\rm fast} v_\xi \quad = \quad & \varepsilon^2 \int_{-\infty}^{\infty} \Big\{ 2\left[H(u_0^2, v_0) - (1 + v_0 - u_0^2)\tfrac{\partial H}{\partial U^2}(u_0^2, v_0)\right] u_0 u_{\rm in} \\ & - \left[H(u_0^2, v_0) - (1 + v_0 - u_0^2)\tfrac{\partial H}{\partial V}(u_0^2, v_0)\right] \Big\}\, d\xi + \mathcal{O}(\varepsilon^2\sqrt{\varepsilon})\end{aligned} \tag{4.21}$$

(see (3.3)), where $u_{\rm in}(\xi)$ is given in (4.18)—recall that $v = 1 + \mathcal{O}(\sqrt{\varepsilon})$ in I_f. As in the previous section, we have approximated $U_h(\xi)$ by $u_0(\xi; v_0)$ (2.3), $V_h(\xi)$ by v_0, and I_f by $\mathbb{R}$ (Theorem 2.5). Note that the integral converges and that $\Delta_{\rm fast} v_\xi$ is (at leading order) independent of γ and $\tilde\lambda$.

It is in principle possible to determine $\Delta_{\rm slow} v_\xi$ in terms of $t_2(\lambda)$ from (4.13); however, this equation is in general not of constant coefficients type (unlike for the example problem in section 4.2). If we introduce the superslow coordinate X by $X = \varepsilon x = \varepsilon^2 \xi$, we can write (4.13) as

$$v_{XX} = \left\{ \tilde\lambda \left[\tau - \frac{H(1 + V_h(X), V_h(X))}{2(1 + V_h(X))} \right] + \gamma + \mathcal{O}(\varepsilon^2) \right\} v; \tag{4.22}$$

i.e., the functions $V_h(X)$ introduce explicit X-dependent terms in the equation (in section 2.3, $V_h(X)$ behaves as $e^{\mp\sqrt{\gamma}X}$ on $\mathcal{M}^{\pm}_\varepsilon$). Nevertheless, we can in principle determine the v-components of the solution $\phi_2(\xi)$ of (3.4) outside the fast region I_f. However, the analysis is much less transparent. For instance, the decomposition (4.5) as in Lemma 4.1 now holds only for $X \gg 1$; therefore the relation between $t_3(\tilde\lambda)$ and

$t_2(\tilde{\lambda})$ that is obtained from the value of v in I_f will in general be more complicated than in (4.6). Moreover, $\tilde{\lambda}[\tau - \frac{H(1+V_h(X),V_h(X))}{2(1+V_h(X))}] + \gamma$ might change sign as a function of X so that the solution $v(X)$ of (4.22) can have oscillatory parts.

Thus, we conclude that it is not a straightforward extension of the approach in the previous section to determine $t_2(\tilde{\lambda})$ for general values of $\tilde{\lambda}$. It should also be noted that a similar problem occurs in the regular case in the study of possible eigenvalues near $\lambda^{\pm}(0)$ (Lemma 4.2). If one introduces $\tilde{\lambda}^{\pm}$ by $\lambda = \lambda^{\pm}(0) + \varepsilon\tilde{\lambda}^{\pm}$ and derives the leading order equation for v_{xx} (4.3) in this case, then one finds an equation like (4.22), i.e., an equation with spatially dependent coefficients (these x-dependent terms originate from the $\mathcal{O}(\varepsilon)$ corrections corresponding to $V_h(x) = \mathcal{O}(\varepsilon)$ in (4.2) and (4.3)). Hence, at this point it is not yet possible to determine in full detail whether or not eigenvalues exist near the tips of σ_{ess} for general nonlinearities $H(U^2, V)$ and general λ. Moreover, it is also not possible to explicitly describe how and when eigenvalues appear from, or disappear into, σ_{ess}. On the other hand, it is clear from (4.21) and (4.22) that the number of zeros of $t_2(\tilde{\lambda})$ depends (for instance) on H_0. It thus follows that eigenvalues will be created/annihilated near the tip of σ_{ess} in the general case (as in the example system considered in the previous section). The analysis of eigenvalues near the tip of σ_{ess} is therefore a continuing subject of research in progress; see also section 5.

Nevertheless, the value $\lambda = \tilde{\lambda} = 0$ is, of course, especially relevant for the stability analysis of the front, and (4.22) is again of constant coefficients type at leading order for this special value of λ. Hence, for $\lambda = 0$ we can obtain the equivalent of Lemma 4.1 so that it follows that

$$\Delta_{\text{slow}} v_\xi|_{\lambda=0} = 2\varepsilon^2(t_2(0) - 1)\sqrt{\gamma} + \mathcal{O}(\varepsilon^2\sqrt{\varepsilon}). \tag{4.23}$$

Note that eventually it becomes clear at this point why the choice $G_1 = -\varepsilon^2\gamma$ is the most relevant scaling of G_1. With this scaling the "jumps" $\Delta_{\text{slow}} v_\xi$ and $\Delta_{\text{fast}} v_\xi$ (4.21) are of the same magnitude in ε at $\lambda = 0$. Therefore, $t_2(0, \varepsilon)$ is asymptotically close to 1 for all G_1 with $|G_1| \gg \varepsilon^2$—see Lemma 4.2 and its proof. Thus, the stability problem (3.3) can only have a double eigenvalue at 0 if $G_1 = \mathcal{O}(\varepsilon^2)$. This establishes a significant link between the stability analysis and the existence analysis of section 2, since it is clear from the analysis there that the scaling $G_1 = \mathcal{O}(\varepsilon^2)$ is also the most relevant scaling for the (superslow) existence problem (Remark 2.2). Moreover, this link is even much more explicit.

THEOREM 4.9. *Assume that $G(V) = -\varepsilon^2\gamma V$, that $H_0 - 2\tau < 0$ and $\mathcal{O}(1)$, and that $\varepsilon > 0$ is small enough. Let the front solution $(U_h(\xi;\varepsilon), V_h(\xi;\varepsilon))$ be a heteroclinic solution that corresponds to an intersection $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ as described in Theorem 2.5. The stability problem associated to the front solution has a double eigenvalue at $\lambda = 0$ if and only if the intersection $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ is nontransversal. If the intersection $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ is a second order contact, then the front bifurcates at*

$$0 < \gamma_{\text{double}} = \frac{1}{4(1+v_0)^2}\left[\int_{-\infty}^{\infty}(1 + v_0 - u_0^2)H(u_0^2, v_0)d\xi + 2\int_{-\infty}^{\infty}(1 + v_0 - u_0^2)[u_0^2\tfrac{\partial H}{\partial U^2}(u_0^2, v_0) + (1 + v_0)\tfrac{\partial H}{\partial V}(u_0^2, v_0)]d\xi\right]^2 \tag{4.24}$$

by merging with another front solution in a saddle-node bifurcation of heteroclinic orbits.

Proof. First, we recall from section 2.3 that a heteroclinic connection that corresponds to the intersection of $W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-} = \{q = \varepsilon\sqrt{\gamma}v\}$ and T_o^- is determined by (2.15). This is, of course, a leading order approximation. In the proof of this theorem we refrain from mentioning this obvious fact at several places. To determine the v_0-dependence of the right-hand side of this relation, we define $w_0(\xi)$ as the (monotonically increasing) heteroclinic solution of $\ddot{w} + (1-w^2)w = 0$. It follows that

$$u_0(\xi; v_0) = \sqrt{1+v_0}\, w_0(\sqrt{1+v_0}\,\xi), \quad w_0(t) = \tanh\sqrt{\frac{1}{2}}t; \tag{4.25}$$

see (2.3). Replacing $u_0(\xi; v_0)$ by $w_0(t)$ in (2.15) yields

$$\sqrt{\gamma}v_0 = \frac{1}{2}\sqrt{1+v_0}\int_{-\infty}^{\infty}(1-w_0^2)H((1+v_0)w_0^2, v_0)dt. \tag{4.26}$$

Thus, $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ is nontransversal if (2.15) holds and

(4.27)

$$\begin{aligned}
\sqrt{\gamma} = {} & \tfrac{1}{2}\tfrac{\partial}{\partial v_0}\left\{\sqrt{1+v_0}\int_{-\infty}^{\infty}(1-w_0^2)H((1+v_0)w_0^2, v_0)dt\right\} \\
= {} & \tfrac{1}{4\sqrt{1+v_0}}\int_{-\infty}^{\infty}(1-w_0^2)H((1+v_0)w_0^2, v_0)dt \\
& +\tfrac{1}{2}\sqrt{1+v_0}\int_{-\infty}^{\infty}(1-w_0^2)[w_0^2\tfrac{\partial H}{\partial U^2}((1+v_0)w_0^2, v_0) + \tfrac{\partial H}{\partial V}((1+v_0)w_0^2, v_0)]dt \\
= {} & \tfrac{1}{2(1+v_0)}\int_{-\infty}^{\infty}(1+v_0-u_0^2)H(u_0^2, v_0)d\xi \\
& \tfrac{1}{1+v_0}\int_{-\infty}^{\infty}(1+v_0-u_0^2)[u_0^2\tfrac{\partial H}{\partial U^2}(u_0^2, v_0) + (1+v_0)\tfrac{\partial H}{\partial V}(u_0^2, v_0)]d\xi
\end{aligned}$$

by reintroducing $u_0(\xi; v_0)$. Note that (4.24) follows from this equation. The expression for $t_2(0,\varepsilon)$ is determined by (4.21), (4.23), and (4.18):

$$t_2(0,\varepsilon) = 1 - \frac{\mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3}{2\sqrt{\gamma}(1+v_0)} + \mathcal{O}(\sqrt{\varepsilon}),$$

where

$$\begin{aligned}
\mathcal{I}_1 &= \int_{-\infty}^{\infty}(1+v_0-u_0^2)H(u_0^2, v_0)d\xi, \\
\mathcal{I}_2 &= \int_{-\infty}^{\infty}(1+v_0-u_0^2)[u_0^2\tfrac{\partial H}{\partial U^2}(u_0^2, v_0) + (1+v_0)\tfrac{\partial H}{\partial V}(u_0^2, v_0)]d\xi, \\
\mathcal{I}_3 &= \int_{-\infty}^{\infty}[(1+v_0-u_0^2)\tfrac{\partial H}{\partial U^2}(u_0^2, v_0) - H(u_0^2, v_0)]\xi u_0 u_{0,\xi}d\xi.
\end{aligned} \tag{4.28}$$

We find by partial integration that

$$\mathcal{I}_3 = \int_{-\infty}^{\infty}\frac{1}{2}\xi\frac{\partial}{\partial\xi}[(1+v_0-u_0^2)H(u_0^2, v_0)]d\xi = -\frac{1}{2}\mathcal{I}_1,$$

which implies that

$$t_2(0,\varepsilon) = 1 - \frac{\mathcal{I}_1 + 2\mathcal{I}_2}{4\sqrt{\gamma}(1+v_0)} + \mathcal{O}(\sqrt{\varepsilon}),$$

so that we can conclude by (4.28) that $t_2(0,\varepsilon) = 0$ is equivalent to the nontransversality condition (4.27). Hence, a double eigenvalue of (3.3) coincides with a saddle-node bifurcation of heteroclinic orbits, unless the tangency between T_o^- and $W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ is degenerate. □

Finally, we can turn to the question about the character of the destabilization of the regular front solution, which has been studied in sections 2.1 and 4.1, as G_1 approaches 0. In order to do so, we first note that the existence problem for the regular case can be recovered from that of the singular limit by reintroducing $G_1 = -\gamma\varepsilon^2$ in the existence condition (2.15). This implies that v_0 must become $\mathcal{O}(\varepsilon)$ and that

$$\sqrt{-G_1}v_0 = \varepsilon\frac{1}{2}\int_{-\infty}^{\infty}(1-u_0^2)H(u_0^2,0)d\xi + \mathcal{O}(\varepsilon\sqrt{\varepsilon}), \tag{4.29}$$

which is equivalent to (2.4) in Theorem 2.1. Thus, the structure of the front $(U_h(\xi), V_h(\xi))$ as a function of $G_1 \uparrow 0$ can be determined by tracing the intersection $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ in the superslow limit as $W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-} = \{q = \varepsilon\sqrt{\gamma}v\}$ goes down from being almost vertical ($G_1 = \mathcal{O}(1)$, $\gamma = \mathcal{O}(1/\varepsilon^2)$) to horizontal ($G_1 = \gamma = 0$). Note that this process determines a unique "regular" element in the intersection $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$; all other elements of $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ do not persist in the regular limit $\gamma = \mathcal{O}(1/\varepsilon^2)$ (here, we do not pay attention to possible heteroclinic connections that have $v_0 \gg 1$ as $\gamma \gg 1$). It depends on the sign of $\frac{1}{2}\int_{-\infty}^{\infty}(1-u_0^2)H(u_0^2,0)d\xi$ whether v_0 will be positive or negative (4.29), i.e., whether the regular intersection $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ travels through the first or through the third quadrant of the (v,q)-plane as γ decreases. Since $H(U^2,V)$ is smooth, we can make a distinction between two different types of behavior:

Type D: The regular element of $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ merges at a certain critical value of $G_1 = -\varepsilon^2\gamma < 0$ with another element of $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ in a saddle-node bifurcation of heteroclinic orbits.

Type E: The regular element of $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ exists up to the limit $G_1 = 0$.

Note that T_o^- approaches $(-1,0)$ as $v_0 \downarrow -1$ (4.26) so that an element of $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ can only reach the singular region $\{v_0 \le -1\}$ at $\gamma = 0$, which indeed implies that there can only be orbits of Type D and E in the third quadrant. We can now describe the destabilization of the regular fronts as G_1 approaches 0.

THEOREM 4.10. *Assume that $G(V) = -\varepsilon^2\gamma V$, that $H_0 - 2\tau < 0$ and $\mathcal{O}(1)$, and that $\varepsilon > 0$ is small enough. Consider the heteroclinic front solution $(U_h(\xi), V_h(\xi))$ determined in Theorem 2.1 for $G_1 < 0$ and $\mathcal{O}(1)$ and in Theorem 2.5 for $G_1 = \mathcal{O}(\varepsilon^2)$. If the front is of Type D as G_1 becomes $\mathcal{O}(\varepsilon^2)$, then it is asymptotically stable up to $G_1 = -\varepsilon^2\gamma_{\text{double}} < 0$ (4.24) and is destabilized by a (discrete) eigenvalue through a saddle-node bifurcation of heteroclinic orbits. A front solution of Type E is stable up to $G_1 = 0$ and is destabilized by the essential spectrum.*

Thus, the destabilization of a regular front solution in the limit $G_1 \uparrow 0$ is completely determined by the geometrical structure of $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ in the superslow limit. Note that Figure 2.3 presents examples of Type D and Type E behavior.

Proof. The proof of this theorem is a bit more subtle than a priori might be expected, since in general we do not have control over the eigenvalues of (3.3) near the tip of σ_{ess} (see also Remark 4.11), except that these eigenvalues must be $\mathcal{O}(\varepsilon^2)$ close to σ_{ess} (see also below). Thus, for instance, the following scenario for a Type D orbit might be possible as γ decreases to γ_{double}: two eigenvalues bifurcate (subsequently)

from σ_{ess} (as real eigenvalues), merge, and become a pair of complex eigenvalues. This pair crosses through the $\text{Re}(\lambda) = 0$ axis at $\gamma_{\text{Hopf}} > \gamma_{\text{double}}$ and touches down again on the real axis. At γ_{double} one of these eigenvalues returns to $\text{Re}(\lambda) = 0$. Thus, in this scenario, there already exists an unstable eigenvalue at $\gamma = \gamma_{\text{double}}$; moreover, the front destabilizes by a Hopf bifurcation at $\gamma_{\text{Hopf}} > \gamma_{\text{double}}$.

Let us first note that a destabilization by a Hopf bifurcation is the only alternative to the statements of the theorem, since eigenvalues move through either 0 or (in pairs) through the $\text{Re}(\lambda) = 0$ axis. If we can show that a Hopf bifurcation cannot occur for $\gamma > \gamma_{\text{double}}$, then it is clear that for Type D orbits $\lambda_{\text{edge}} < 0$ for $\gamma > \gamma_{\text{double}}$ and that there is no unstable spectrum at $\gamma = \gamma_{\text{double}}$ (this follows from Theorem 4.3: if γ is $\gg \mathcal{O}(1/\varepsilon)$, all nontrivial eigenvalues must be in $\{\text{Re}(\lambda) < -\varepsilon\}$; hence, by decreasing γ, there is one eigenvalue, λ_{edge}, that is the first to reach 0; this happens at the saddle-node bifurcation (Theorem 4.9), i.e., at $\gamma = \gamma_{\text{double}}$). Thus, the front is stable for $\gamma > \gamma_{\text{double}}$. The same argument can be used to establish the nonexistence of unstable spectrum for Type E orbits if there are no Hopf bifurcations possible.

To show that there cannot be Hopf bifurcations (for $H_0 - 2\tau < 0$ and $\mathcal{O}(1)$, see section 5), we first ascertain that λ must be $\mathcal{O}(\varepsilon^2)$, i.e., that (4.12) is the correct scaling. This follows by the same arguments as in the proof of Lemma 4.2. If $\Delta_{\text{slow}} v_\xi \gg \Delta_{\text{fast}} v_\xi$, then there cannot be an eigenvalue. Thus, it follows from (4.11) that $|\lambda|$ must indeed be $\mathcal{O}(\varepsilon^2)$ near $\lambda^+(0)$. Hence, even if there is a Hopf bifurcation, it will be $\mathcal{O}(\varepsilon^2)$ close to 0. Next, we realize that this situation is covered by (4.21) for the jump through the fast field; thus, $\Delta_{\text{fast}} v_\xi$ is real (at leading order), independent of $\tilde{\lambda}$. However, it follows from (4.22) that $\Delta_{\text{slow}} v_\xi$ cannot be real if $\tilde{\lambda}$ is complex valued. Hence, there cannot be a Hopf bifurcation $\mathcal{O}(\varepsilon^2)$ close to $\lambda = 0$. □

Remark 4.11. By the same geometrical arguments (that are based on Theorem 4.9) we can describe the character of the bifurcations as function γ in the stability problem associated to a heteroclinic orbit that corresponds to a nonregular element of $T_o^- \cap W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$. However, it should be noted that, in general, we do not have enough information on the spectrum of (3.3) to establish the stability of such a front, since we did not determine all possible eigenvalues. In general, we cannot exclude the possibility that various eigenvalues have bifurcated from the essential spectrum for these fronts (in fact, the possible oscillatory character of a solution $v(X)$ of (4.22) strongly suggests that this can happen). Nevertheless, we may, for instance, conclude that if the regular orbit is of Type D, then it merges with a nonregular orbit at γ_{double} that is unstable for any $\gamma > \gamma_{\text{double}}$ for which it exists.

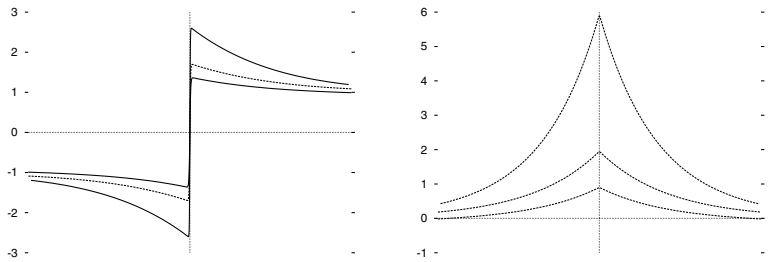
Remark 4.12. The most simple example one can consider is $H(U^2, V) \equiv H_0$. This corresponds to the case in which the function $F(U^2, V)$ in (1.1) is (the most general) linear function of U^2 and V with parameters G_1 and H_0 (i.e., $F(U^2, V) = H_0 + (H_0 + G_1)V - H_0 U^2$; recall that $F(1,0)$ must be 0). In this case, T_o^- is given by $\{q = 2\varepsilon H_0 \sqrt{1 + v_0} + \mathcal{O}(\varepsilon^2)\}$ so that $W^u(-1,0,0,0)|_{\mathcal{M}_\varepsilon^-}$ can never be tangent to T_o^-. Hence, in this case, there is a uniquely determined front solution of Type E for any $H_0 \neq 0$ and $G_1 < 0$; i.e., the front solution is stable up to $G_1 = 0$ and is destabilized by the essential spectrum.

Remark 4.13. We did not consider the degenerate case in which $H(U^2, V)$ is such that $H(1 + V, V) \equiv 0$ (section 1), i.e., functions H such that $H(U^2, V) = (1 + V - U^2)\tilde{H}(U^2, V)$ for some smooth function $\tilde{H}$. In a sense, this is a much more simple problem, for instance, since in the superslow limit, the stability problem in the slow field is automatically of constant coefficients type (at leading order); see (4.11), (4.22). Moreover, it is also clear from these same relations that we can find $\mathcal{O}(1)$ instead of

$\mathcal{O}(\varepsilon^2)$ eigenvalues in this case if $\tau = \mathcal{O}(\varepsilon^2)$. In fact, the situation is very much like the stability analysis of (homoclinic) pulses in monostable systems in [4, 5]. For instance, as in [4, 5], potential eigenvalues are no longer "slaved" to the tips of the essential spectrum or to the eigenvalues of the fast reduced limit (Lemma 4.2). Moreover, the "natural" persistence result of Lemma 4.4 is also not valid in this case, in general.

5. Simulations and discussion.

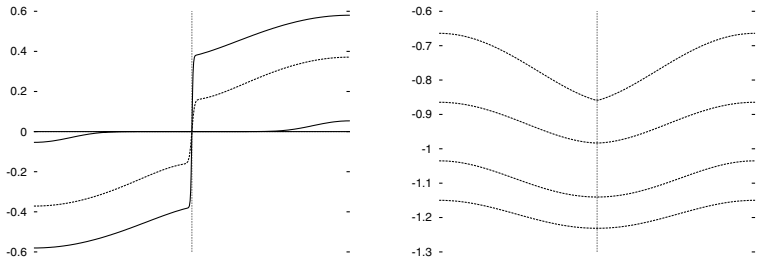
5.1. Simulations. We now examine numerically the difference between the two types of bifurcations discussed in Theorems 4.6 and 4.10. We consider the example system of sections 2.2 and 4.2 for $H_0 > 0$ (case (i), Type D) and $H_0 < 0$ (case (ii), Type E). First, we note that in both cases the simulations confirm that the fronts are asymptotically stable up to the analytically determined bifurcation values. In case (i) the front destabilizes at $\gamma < \gamma_{\text{double}}$ due to an eigenvalue in the discrete spectrum. The eigenfunction associated to this type of destabilization is localized to a neighborhood of the front, as can be seen in Figure 5.1. In this case the front becomes unstable and blows up in finite time, while the background states remain stable. In case (ii), the tip of the essential spectrum becomes positive and the background states become unstable as γ passes through 0. As can be seen in Figure 5.2, this destabilization causes the front to collapse. The U component then tends to 0 on the entire real line, and the V component grows according to $V_t = V_{xx} + \varepsilon^2|\gamma|V$. Thus, we may conclude that Type D or Type E orbits indeed exhibit significantly different behavior at the destabilization. These simulations were performed using SPMDF [2], with Neumann boundary conditions at $x = \pm 50$. The initial conditions used in Figure 5.1 are given by $U(x,0) = u_0(x,\varepsilon;v_1)$ (2.3) and $V(x,0) = v_1 e^{-\varepsilon\sqrt{|\gamma|}|x|}$ (as described in Theorem 2.3).



(a) U-component (b) V-component

FIG. 5.1. *Numerical simulation of destabilization caused by the discrete spectrum; both components blow up in finite time ($H_0 = 1$, $\gamma = 1.4$, $\tau = 1$, and $\varepsilon = 0.1$).*

5.2. Hopf bifurcations. As we have seen in section 4.1, in general there can be (complex) eigenvalues near the endpoints $\lambda^{\pm}(0)$ of σ_{ess}. Thus, if we keep $G_1 < 0$ fixed at an $\mathcal{O}(1)$ value and increase H_0 such that $H_0 + G_1 - 2\tau$ approaches 0, we encounter a similar issue as was studied in the previous section: Will the front be destabilized by σ_{ess} at $H_0 = 2\tau - G_1$, or (just) before that, by an eigenvalue? In this case, the bifurcation is of Hopf type, and it is not associated to the existence problem. This problem can in principle be analyzed by the methods developed here, i.e., by determining $t_2(\lambda,\varepsilon)$ through $\Delta_{\text{slow}}v_\xi$ and $\Delta_{\text{fast}}v_\xi$. We have already mentioned



(a) U-component (b) V-component

FIG. 5.2. *Numerical simulation of destabilization caused by the essential spectrum; $U \to 0$ and $|V|$ grows slowly and exponentially ($H_0 = -1$, $\gamma = -0.1$, $\tau = 1$, and $\varepsilon = 0.1$).*

the new features of the measuring the slow "jump" $\Delta_{\text{slow}} v_\xi$ in section 4.3. Moreover, since the bifurcation does not occur near $\lambda = 0$, we do not have an explicit formula for $u_{\text{in}}(\xi)$, like (4.18), and it is thus not immediately clear whether it is possible to determine $\Delta_{\text{fast}} v_\xi$. Note that this latter issue is solvable with the hypergeometric functions method developed in [3, 5]. Nevertheless, we do not go deeper into this subject here.

5.3. Planar fronts and stripes. A next step in the study of (planar) stripes, as mentioned in the introduction, is the stability analysis of planar fronts, i.e., the analysis of the stability of the fronts $(U_h(\xi), V_h(\xi))$ with respect to two-dimensional perturbations (thus, $(U_h(\xi), V_h(\xi))$ represents a planar front that has a trivial structure in the y-direction). The methods developed here can be used to study this problem (as is also suggested by [7] in which a similar problem has been studied in a monostable Gierer–Meinhardt context). It should be noted here that there are several papers in the literature that consider the question of the (non-)persistence of the stability of one-dimensional fronts as two-dimensional planar fronts (see, for instance, [17, 20, 13, 16]). The analysis in [20, 16] of a class of singularly perturbed bistable systems shows that the planar fronts considered there cannot be stable, while it is shown that planar fronts can be stable in a more regular context in [13]. Thus, this is a nontrivial issue. Preliminary analysis of the front solutions considered in this paper indicates that these solutions remain stable as planar fronts in the regular case (i.e., as long as $G_1 < 0$ and $\mathcal{O}(1)$). The analysis of the planar fronts and their spatially periodic counterparts, the stripe patterns, is the subject of a work in progress.

REFERENCES

[1] J. ALEXANDER, R. A. GARDNER, AND C. K. R. T. JONES, *A topological invariant arising in the stability analysis of travelling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.

[2] J. G. BLOM AND P. A. ZEGELING, *Algorithm* 731: *A moving-grid interface for systems of one-dimensional time-dependent partial differential equations*, ACM Trans. Math. Software, 20 (1994), pp. 194–214.

[3] A. DOELMAN, R. A. GARDNER, AND T. J. KAPER, *Stability analysis of singular patterns in the* 1*-D Gray–Scott model: A matched asymptotics approach*, Phys. D, 122 (1998), pp. 1–36.

[4] A. DOELMAN, R. A. GARDNER, AND T. J. KAPER, *A stability index analysis of* 1*-D patterns of the Gray–Scott model*, Mem. Amer. Math. Soc., 155 (2002).

[5] A. Doelman, R. A. Gardner, and T. J. Kaper, *Large stable pulse solutions in reaction-diffusion equations*, Indiana Univ. Math. J., 50 (2001), pp. 443–507.
[6] A. Doelman, T. J. Kaper, and P. Zegeling, *Pattern formation in the one-dimensional Gray-Scott model*, Nonlinearity, 10 (1997), pp. 523–563.
[7] A. Doelman and H. van der Ploeg, *Homoclinic stripe patterns*, SIAM J. Appl. Dyn. Syst., 1 (2002), pp. 65–104.
[8] W. Eckhaus, *Asymptotic Analysis of Singular Perturbations*, North–Holland, Amsterdam, 1979.
[9] N. Fenichel, *Geometrical singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.
[10] R. A. Gardner and C. K. R. T. Jones, *Stability of the travelling wave solutions of diffusive predator-prey systems*, Trans. Amer. Math. Soc., 327 (1991), pp. 465–524.
[11] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, 1981.
[12] C. K. R. T. Jones, *Geometric singular perturbation theory*, in Dynamical Systems, Lecture Notes in Math. 1609, R. Johnson ed., Springer-Verlag, Berlin, 1995, pp. 44–118.
[13] T. Kapitula, *Multidimensional stability of planar travelling waves*, Trans. Amer. Math. Soc., 349 (1997), pp. 257–269.
[14] T. Kapitula, *The Evans function and generalized Melnikov integrals*, SIAM J. Math. Anal., 30 (1998), pp. 273–297.
[15] T. Kapitula and B. Sandstede, *Edge bifurcations for near integrable systems via Evans function techniques*, SIAM J. Math. Anal., 33 (2002), pp. 1117–1143.
[16] Y. Nishiura and H. Suzuki, *Nonexistence of higher dimensional stable Turing patterns in the singular limit*, SIAM J. Math. Anal., 29 (1998), pp. 1087–1105.
[17] T. Ohta, M. Mimura, and R. Kobayashi, *Higher-dimensional localized patterns in excitable media*, Phys. D, 34 (1989), pp. 115–144.
[18] R. L. Pego and M. I. Weinstein, *Eigenvalues, and instabilities of solitary waves*, Philos. Trans. Roy. Soc. London Ser. A, 340 (1992), pp. 47–94.
[19] C. Robinson, *Sustained resonance for a nonlinear system with slowly-varying coefficients*, SIAM J. Math. Anal., 14 (1983), pp. 847–860.
[20] M. Taniguchi and Y. Nishiura, *Instability of planar interfaces in reaction-diffusion systems*, SIAM J. Math. Anal., 25 (1994), pp. 99–134.
[21] E. C. Titchmarsh, *Eigenfunction Expansions Associated with Second-Order Differential Equations*, 2nd ed., Oxford University Press, Oxford, UK, 1962.

SIAM J. MATH. ANAL.
Vol. 35, No. 6, pp. 1451–1485

REGULARITY OF SOLUTIONS TO THE NAVIER–STOKES SYSTEM FOR COMPRESSIBLE FLOWS ON A POLYGON*

JAE RYONG KWEON† AND R. BRUCE KELLOGG‡

Abstract. The steady-state nonlinear compressible viscous Navier–Stokes system with nonzero boundary conditions is considered on a polygon D. It is shown that the leading corner singularities for the velocity are the same as those of the Lamé system and the leading corner singularity for the temperature is the same as that of the Laplacian. If P is a concave vertex of D with interior angle ω, the velocity $\mathbf{u}$ and temperature σ can be split into singular and regular parts near the vertex P. The regular functions are $\mathbf{u}_R = \mathbf{u} - \chi[C_1 r^{\lambda_1}\mathcal{T}_1(\theta) + C_2 r^{\lambda_2}\mathcal{T}_2(\theta)] \in \mathbf{H}^{2,q}$ and $\sigma_R = \sigma - \chi C_3 r^{\pi/\omega}\sin[(\pi/\omega)\theta] \in \mathrm{H}^{2,q}$ with $2 < q < 1/(1-\lambda_1)$, where the numbers λ_i $(i = 1, 2)$ satisfy $\frac{1}{2} < \lambda_1 < \pi/\omega < \lambda_2 < 1$, the $\mathcal{T}_i$ are trigonometric vector functions, χ is a cutoff function, C_i $(i = 1, 3)$ are constants, and r is the distance to the vertex. If D is convex, $[\mathbf{u}, \sigma] \in \mathbf{H}^{2,q} \times \mathrm{H}^{2,q}$.

Key words. compressible flows, corner singularities, nonlinearity

AMS subject classifications. 35M10, 35B25

DOI. 10.1137/S0036141002418066

1. Introduction and main results.

Our concern is with the compressible Navier–Stokes system. Among the many open problems associated with this system is to give a description of the singularities of a solution caused by a corner or edge of the boundary. In addition to the intrinsic mathematical interest in these singularities, such information might have application to certain physical problems. For instance, imagine a high speed flow over a body where a wall of the body is turned downward (or upward) at the corner through a deflection angle [2]. If the flow is in high speed, say, supersonic or hypersonic, then at the corner the flow properties may change drastically. In addition, when a solution domain is composed of different materials, corner singularities may occur at the intersections of their internal interfaces. A further understanding of fluid behavior at such corners may be an essential ingredient in such related physical situations.

The purpose of this paper is to study the steady-state compressible Navier–Stokes system in two dimensions in a polygonal domain with nonzero boundary conditions. In particular there is given a decomposition of the velocity and the temperature into singular and regular parts near concave vertices of the polygon.

The equations to be considered are

$$\begin{cases} -\mu\Delta\mathbf{u} - (\mu+\nu)\nabla\mathrm{div}\mathbf{u} + \rho(\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = 0 \quad \text{in} \quad D, \\ \mathrm{div}(\rho\mathbf{u}) = 0 \quad \text{in} \quad D, \\ -\gamma\Delta\sigma + c_v\rho\mathbf{u}\cdot\nabla\sigma + \sigma p_\sigma \mathrm{div}\mathbf{u} = \psi(\mathbf{u},\mathbf{u}) \quad \text{in} \quad D, \\ \mathbf{u} = \mathbf{u}_0, \quad \sigma = \sigma_0 \quad \text{on} \quad \partial D, \\ p = p_0 \quad \text{on} \quad \partial D_{in}, \end{cases} \tag{1.1}$$

where D is an open bounded domain in the plane with polygonal boundary ∂D,

*Received by the editors November 19, 2002; accepted for publication (in revised form) August 15, 2003; published electronically February 18, 2004.
http://www.siam.org/journals/sima/35-6/41806.html

†Department of Mathematics, Pohang University of Science and Technology, Pohang 790–784, Korea (kweon@postech.ac.kr). This work was supported by Korea Research Foundation grant KRF–2001–015–DS0002.

‡Department of Mathematics, University of South Carolina, Columbia, SC 29208 (kellogg@ipst.umd.edu).

$\mathbf{u} = [u_1, u_2]$ is the unknown velocity vector, and p is the unknown pressure; σ is the unknown temperature; $\rho = \rho(p, \sigma)$ is a given positive function, strictly increasing in the first variable, that provides density as a function of pressure and temperature; $\mathbf{u}_0 = [u_0, v_0]$ with $u_0 > 0$, p_0, and $\sigma_0 > 0$ are the given boundary values; μ, ν are the constant coefficients of viscosity with $\mu > 0$ and $\mu + \nu > 0$, γ is the constant coefficient of conductivity with $\gamma > 0$, and c_v is a positive constant. The number ν, the "bulk" viscosity, is often taken to be $-\frac{2}{3}\mu$ [3, Chapter 3] as suggested by Stokes (see [2, Chapter 15]). The three differential equations in (1.1) represent the conservation of momentum, mass, and energy, respectively. In the energy equation

$$\psi(\mathbf{u}, \mathbf{u}) = \gamma_0 \sum_{i,j=1}^{2} \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right)^2 + \gamma_1 (\mathrm{div}\mathbf{u})^2, \tag{1.2}$$

where γ_0 and γ_1 are positive constants, it is assumed that the boundary data $\mathbf{u}_0$, ρ_0, and σ_0 are given by functions defined on all of $\bar{D}$. The inflow and outflow boundaries, ∂D_{in} and ∂D_{out}, are defined by

$$\begin{aligned} \partial D_{in} &= \{(x, y) \in \partial D : \mathbf{u}_0 \cdot \mathbf{n} < 0\}, \\ \partial D_{out} &= \{(x, y) \in \partial D : \mathbf{u}_0 \cdot \mathbf{n} \geq 0\}, \end{aligned} \tag{1.3}$$

where $\mathbf{n}$ denotes the unit outward pointing normal to ∂D. The function $\rho(p, \sigma)$ is assumed to have Lipschitz continuous first order partial derivatives ρ_p, ρ_σ, with gradient

$$\nabla \rho = \rho_p \nabla p + \rho_\sigma \nabla \sigma. \tag{1.4}$$

In [14] the stationary barotropic compressible Navier–Stokes equations were investigated in a plane domain with corners of angle less than π. In [13] we analyzed the barotropic compressible Navier–Stokes system on a nonconvex polygonal domain with the further simplification that $\nu = -\mu$. We showed that the lowest order corner singularity comes from the Laplace operator; the continuity equation does not have an effect on this lowest order singularity. In this paper a similar result is obtained; the lowest order singularity in the velocity comes from the Lamé system, and the lowest order singularity in the temperature comes from the Laplace operator. The analysis follows that of [13], but complications arise from the presence of the Lamé system in the momentum equations and the presence of the energy equation in the system.

Let $\bar{\mathbf{u}} = \mathbf{u} - \mathbf{u}_0$, $\bar{p} = p - p_0$, and $\bar{\sigma} = \sigma - \sigma_0$. Inserting $[\mathbf{u}, p, \sigma] = [\bar{\mathbf{u}} + \mathbf{u}_0, \bar{p} + p_0, \bar{\sigma} + \sigma_0]$ into system (1.1), rearranging the resulting system for the unknown variable $[\bar{\mathbf{u}}, \bar{p}, \bar{\sigma}]$ and setting $[\bar{\mathbf{u}}, \bar{p}, \bar{\sigma}] = [\mathbf{u}, p, \sigma]$ again, system (1.1) becomes

$$\begin{aligned} -\mu \Delta \mathbf{u} - (\mu + \nu) \nabla \mathrm{div}\mathbf{u} + \rho(\mathbf{u} + \mathbf{u}_0) \cdot \nabla \mathbf{u} + \nabla p &= \mathbf{f}, \\ \mathrm{div}\mathbf{u} + \kappa_1 (\mathbf{u} + \mathbf{u}_0) \cdot \nabla p + \kappa_2 (\mathbf{u} + \mathbf{u}_0) \cdot \nabla \sigma &= g, \\ -\gamma \Delta \sigma + c_v \rho (\mathbf{u} + \mathbf{u}_0) \cdot \nabla \sigma + \sigma p_\sigma \mathrm{div}\mathbf{u} &= h, \end{aligned} \tag{1.5}$$

with $[\mathbf{u}, p, \sigma]$ satisfying zero boundary conditions, where $\kappa_1 = \rho_p/\rho$, $\kappa_2 = \rho_\sigma/\rho$, and

$$\begin{aligned} \mathbf{f}_0 &:= \mu \Delta \mathbf{u}_0 + (\mu + \nu) \nabla \mathrm{div}\mathbf{u}_0 - \rho \mathbf{u}_0 \cdot \nabla \mathbf{u}_0 - \nabla p_0, \\ g_0 &:= -\mathrm{div}\mathbf{u}_0 - \kappa_1 \mathbf{u}_0 \cdot \nabla p_0 - \kappa_2 \mathbf{u}_0 \cdot \nabla \sigma_0, \\ h_0 &:= \gamma \Delta \sigma_0 - c_v \rho \mathbf{u}_0 \cdot \nabla \sigma_0 - \sigma_0 p_\sigma \mathrm{div}\mathbf{u}_0, \\ \mathbf{f}(\mathbf{u}, p, \sigma) &= \mathbf{f}_0 - \rho \mathbf{u} \cdot \nabla \mathbf{u}_0, \\ g(\mathbf{u}, p, \sigma) &= g_0 - \kappa_1 \mathbf{u} \cdot \nabla p_0 - \kappa_2 \mathbf{u} \cdot \nabla \sigma_0, \\ h(\mathbf{u}, p, \sigma) &= h_0 - c_v \rho \mathbf{u} \cdot \nabla \sigma_0 + p_\sigma (\sigma_0 \mathrm{div}\mathbf{u} - \sigma \mathrm{div}\mathbf{u}_0) \\ &\quad + \psi(\mathbf{u} + \mathbf{u}_0, \mathbf{u} + \mathbf{u}_0). \end{aligned} \tag{1.6}$$

Note that system (1.5) with zero boundary condition is equivalent to (1.1).

In order to state the main result of this paper we give some notation concerning the polygon D and some numbers concerning the singular functions. Let the vertices of D be denoted by P_n, $n = 1, \dots, N$. Let ω_n be the interior angle of the vertex P_n. Let $\lambda_{1,n}$ and $\lambda_{2,n}$ be the first and second leading singular exponents for the Lamé system (1.12) corresponding to each concave vertex P_n so that $\frac{1}{2} < \lambda_{1,n} < \lambda_{2,n} < 1$ (see [9]). Let $q_2^* = \min_n\{2/(2-\lambda_{1,n}) : P_n \text{ is a concave vertex}\}$, and let $\mathcal{I}^* = \{n : 2/(2-\lambda_{1,n}) \geq q_2^*\}$. We define

$$q_1^* = \min_{n\in\mathcal{I}^*}\left\{\frac{2}{1-\lambda_{1,n}} : P_n \text{ is a concave vertex}\right\}.$$

Our main result, which will be shown in section 5, is the following theorem.

THEOREM 1.1. *Suppose that μ and γ are sufficiently large. Let D be concave, and let $2 < q < \frac{1}{2}q_1^*$. Let $\mathbf{u}_0 \in \mathbf{H}^{2,q}(D)$, $p_0 \in H^{1,q}(D)$, and $\sigma_0 \in H^{2,q}(D)$. Suppose the vector field $\mathbf{u}_0$ satisfies the conditions* (A2)–(A4) *stated in section 2. For any constant K_1, there is a constant K_2 such that if $\|[\mathbf{u}_0, \sigma_0]\|_{2,q,D} + \|p_0\|_{1,q,D} \leq K_1$ and $\|[\nabla\mathbf{u}_0, \nabla\sigma_0]\|_{1,q,D} + \|\nabla p_0\|_{1,q,D} + |\sigma_0|_\infty \leq K_2$, there is a unique solution $[\mathbf{u}, p, \sigma] \in \mathbf{H}^{1,q}(D) \times \mathrm{H}^{1,q}(D) \times \mathrm{H}^{1,q}(D)$ of system* (1.1). *Furthermore, the solution has the following properties. For each $n \in \mathcal{I}^*$ let (r_n, θ_n) be the polar coordinates based on the concave vertex P_n, arranged so that $\theta_0 = 0$ on one side of P_n, and let $\mathbf{u}_{n,s} = \chi_n(C_{1,n}\Phi_{1,n} + C_{2,n}\Phi_{2,n})$ and $\sigma_{n,s} = \chi_n C_{3,n}\phi_n$, where $C_{1,n}$, $C_{2,n}$, and $C_{3,n}$ are the constants given in* (3.31), (3.62), *and* (3.51), *respectively, $\Phi_{i,n}$ and ϕ_n are given in* (2.11) *and* (2.4), *respectively, and χ_n is a smooth cutoff function which is* 1 *near P_n and zero outside a small neighborhood of P_n. Then the solution $[\mathbf{u}, \sigma]$ may be split into singular and regular parts*

$$[\mathbf{u}, \sigma] = \sum_{n\in\mathcal{I}^*}[\mathbf{u}_{n,s}, \sigma_{n,s}] + [\mathbf{u}_R, \sigma_R], \quad [\mathbf{u}_R, \sigma_R] := [\mathbf{u}, \sigma] - \sum_{n\in\mathcal{I}^*}[\mathbf{u}_{n,s}, \sigma_{n,s}]$$

with the property $[\mathbf{u}_R, p, \sigma_R] \in \mathbf{H}^{2,q}(D) \times \mathrm{H}^{1,q}(D) \times \mathrm{H}^{2,q}(D)$. Also there is a constant $K_3 = C(K_1, K_2, D)$ such that

$$\|[\mathbf{u}_R - \mathbf{u}_0,\, \sigma_R - \sigma_0]\|_{2,q,D} + \|[\mathbf{u} - \mathbf{u}_0,\, \sigma - \sigma_0]\|_{1,q,D} + \|p - p_0\|_{1,q,D} \leq K_3.$$

If D is convex and $q > 2$, then $[\mathbf{u}, \sigma] = [\mathbf{u}_R, \sigma_R]$ satisfies the above inequality.

Note that in Theorem 1.1, the large condition on the conductivity γ is only needed for defining the solution operator $\mathcal{E}$ (see (1.14)).

In order to prove Theorem 1.2 we linearize system (1.5) and analyze it on D. To do this, let $\mathbf{w}$, η, and τ be given functions with $\mathbf{w} = 0$ on ∂D, $\eta = 0$ on ∂D_{in}, and $\tau = 0$ on ∂D, respectively. Let $\mathbf{U} = \mathbf{w} + \mathbf{u}_0$, $\mathbf{f} = \mathbf{f}(\mathbf{w}, \eta, \tau)$, $g = g(\mathbf{w}, \eta, \tau)$, and $h = h(\mathbf{w}, \eta, \tau)$. Our linearized system for (1.5) is

$$\left\{\begin{array}{l} -\mu\Delta\mathbf{u} - (\mu+\nu)\nabla\mathrm{div}\mathbf{u} + \rho\,(\mathbf{U}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } D, \\ \quad \mathrm{div}\mathbf{u} + \kappa_1\mathbf{U}\cdot\nabla p + \kappa_2\mathbf{U}\cdot\nabla\sigma = g \quad \text{in } D, \\ -\gamma\Delta\sigma + \tilde{\rho}\,\mathbf{U}\cdot\nabla\sigma + \tilde{\tau}\,\mathrm{div}\mathbf{u} = h \quad \text{in } D, \\ \quad \mathbf{u} = 0,\ \sigma = 0 \quad \text{on } \partial D, \\ \quad p = 0 \quad \text{on } \partial D_{in}, \end{array}\right. \tag{1.7}$$

where $\rho = \rho(\eta + p_0, \tau + \sigma_0)$, $\kappa_1 = \rho_p(\eta + p_0, \tau + \sigma_0)/\rho$, $\kappa_2 = \rho_\sigma(\eta + p_0, \tau + \sigma_0)/\rho$, $\tilde{\rho} = \rho c_v$, and $\tilde{\tau} = \tau p_\sigma$ with $p_\sigma = p_\sigma(\rho + \rho_0, \tau + \sigma_0)$. In studying (1.7) it is assumed that $\mathbf{U} = [U, V]$, ρ, κ_1, τ are given functions.

To analyze (1.7) on the bounded polygon D and for an ease of description, some restrictions are needed on the vector field $\mathbf{U}$ and the polygon D (see Assumption A in section 2). From Assumption A, it is possible to construct a finite open covering $\{\Omega_i\}$ such that each open set Ω_i is a polygon having at most one vertex of D, the remaining vertices being convex (see section 4). Also, Ω_i satisfies (A4). Thus it suffices to analyze the behavior of the solution in a polygon Ω having one concave vertex P. Without loss of generality assume that the vertex P is placed at the origin $(0,0)$. Let $\chi \in C_0^\infty(R^2)$ be a smooth cutoff function which is identically 1 near the origin $(0,0)$ and which satisfies

$$\chi(x,y) \equiv 0 \text{ outside a neighborhood of } (0,0). \tag{1.8}$$

Using (1.7) and (1.8), the behavior of the solution near the origin $(0,0)$ can be investigated by considering the following generalized compressible Stokes system:

$$\begin{cases} -\mu\,\Delta(\chi\mathbf{u}) - (\mu+\nu)\nabla\mathrm{div}(\chi\mathbf{u}) + \rho(\mathbf{U}\cdot\nabla)(\chi\mathbf{u}) + \nabla(\chi p) \\ = \chi\,\mathbf{f} - 2\,\mu\nabla\chi\cdot\nabla\mathbf{u} - (\mu+\nu)(\nabla\chi\mathrm{div}\mathbf{u} + \nabla\chi\nabla\mathbf{u}) \\ \quad + \mathbf{u}(-\mu\,\Delta\chi - (\mu+\nu)\nabla^2\chi + \rho\mathbf{U}\cdot\nabla\chi) + p\nabla\chi \ \text{ in } \Omega, \\ \mathrm{div}(\chi\mathbf{u}) + \kappa_1\,\mathbf{U}\cdot\nabla(\chi p) + \kappa_2\,\mathbf{U}\cdot\nabla(\chi\sigma) \\ = \chi\,g + \mathbf{u}\cdot\nabla\chi + (p\kappa_1 + \sigma\kappa_2)\mathbf{U}\cdot\nabla\chi \ \text{ in } \Omega, \\ -\gamma\,\Delta(\chi\sigma) + \tilde{\rho}(\mathbf{U}\cdot\nabla)(\chi\sigma) + \tilde{\tau}\mathrm{div}(\chi\mathbf{u}) \\ = \chi h - 2\,\gamma\nabla\chi\cdot\nabla\sigma + \sigma\tilde{\rho}\,\mathbf{U}\cdot\nabla\chi + \tilde{\tau}\nabla\chi\cdot\mathbf{u} \ \text{ in } \Omega, \\ \chi\mathbf{u} = 0 \ \text{ on } \Gamma, \ \chi\sigma = 0 \ \text{ on } \Gamma, \\ \chi p = 0 \ \text{ on } \Gamma_{in}, \end{cases} \tag{1.9}$$

where Γ is the boundary of Ω and Γ_{in} is the inflow boundary of Γ. One finds that $[\chi\mathbf{u}, \chi p, \chi\sigma]$ is a weak solution of (1.9). Applying this to each set Ω_i in the open cover $\{\Omega_i\}$, we will see that the solution $[\mathbf{u}, p, \sigma]$ of (1.7) may be expressed in a sum of the local functions:

$$\mathbf{u} = \sum_{n=1}^{N} \mathbf{u}_n, \quad p = \sum_{n=1}^{N} p_n, \quad \sigma = \sum_{n=1}^{N} \sigma_n. \tag{1.10}$$

Here the triple $[\mathbf{u}_n, p_n, \sigma_n]$, $1 \le n \le N$, is the weak solution of (1.9) corresponding to the vertex P_n of D. Hence, from (1.9) it suffices to consider the equations

$$\begin{cases} -\mu\,\Delta\mathbf{u} - (\mu+\nu)\nabla\mathrm{div}\mathbf{u} + \rho\,(\mathbf{U}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{f} \ \text{ in } \Omega, \\ \mathrm{div}\mathbf{u} + \kappa_1\,\mathbf{U}\cdot\nabla p + \kappa_2\,\mathbf{U}\cdot\nabla\sigma = g \ \text{ in } \Omega, \\ -\gamma\Delta\sigma + \tilde{\rho}\,\mathbf{U}\cdot\nabla\sigma + \tilde{\tau}\mathrm{div}\mathbf{u} = h \ \text{ in } \Omega, \\ \mathbf{u} = 0, \ \sigma = 0 \ \text{ on } \Gamma, \\ p = 0 \ \text{ on } \Gamma_{in}. \end{cases} \tag{1.11}$$

The problem (1.11) will be discussed in section 3. For the existence of a solution, see Lemma 2.9. In section 3 we will show that the velocity and the temperature of the solution $[\mathbf{u}, p, \sigma]$ of (1.11) may be decomposed into singular and regular parts near the concave vertex.

We next state a result for the linearized problem (1.7), which will be shown in section 4. In doing so, we give some notation. Set $q_2^\lambda = \min_n\{2/(s-\lambda_{1,n}) : P_n \text{ is a concave vertex}\}$ and define a set

$$\mathcal{I}_s^* = \{n : 2/(s-\lambda_{1,n}) \ge q_2^\lambda\}.$$

THEOREM 1.2. *Let D be a concave polygon. Suppose Assumption A given in section 2 holds. Let $q > 2$ be sufficiently close to 2 and $s \geq 1$. Assume that $[\mathbf{f}, g, h] \in \mathbf{H}^{s-2,q}(D) \times \mathrm{H}^{s-1,q}(D) \times \mathrm{H}^{s-2,q}(D)$. Suppose that μ and γ are sufficiently large.*

(a) *If*

$$s < \min_{n \in \mathcal{I}_s^*} \{\lambda_{1,n}\} + 1 + 2/q,$$

there is a unique solution $[\mathbf{u}, p, \sigma] \in \mathbf{H}^{s-1,q}(D) \times \mathrm{H}^{s-2,q}(D) \times \mathrm{H}^{s-1,q}(D)$ of (1.7).

(b) *Let $\mathcal{I}_s^* \neq \emptyset$. If s is given with*

$$\max_{n \in \mathcal{I}_s^*} \{\lambda_{2,n}\} + 2/q < s \leq 2,$$

then the solution $[\mathbf{u}, \sigma]$ may be split into singular and regular parts

$$[\mathbf{u}, \sigma] = \sum_{n \in \mathcal{I}_s^*} [\mathbf{u}_{n,s}, \sigma_{n,s}] + [\mathbf{u}_R, \sigma_R], \quad [\mathbf{u}_R, \sigma_R] = [\mathbf{u}, \sigma] - \sum_{n \in \mathcal{I}_s^*} [\mathbf{u}_{n,s}, \sigma_{n,s}]$$

with $[\mathbf{u}_R, p, \sigma_R] \in \mathbf{H}^{s,q}(D) \times \mathrm{H}^{s-1,q}(D) \times \mathrm{H}^{s,q}(D)$ and

$$[\mathbf{u}_{n,s},\, \sigma_{n,s}] = [C_{1,n}\Phi_{1,n} + C_{2,n}\Phi_{2,n},\, C_{3,n}\phi_n],$$

where $\mathbf{C}_n = [C_{1,n}, C_{2,n}, C_{3,n}]$ is constructed in Steps 1, 2, and 3 in section 3. Furthermore, there is a constant $K = C(C_0, \|\mathbf{U}_R\|_{2,q,D} + \sum_{n \in \mathcal{I}_s^*} |\mathbf{d}_n|)$ with a given constant vector $\mathbf{d}_n = [d_{1,n}, d_{2,n}, d_{3,n}]$ such that

$$\begin{aligned} &\|[\mathbf{u}_R, \sigma_R]\|_{s,q,D} + \sum_{n \in \mathcal{I}_s^*} |\mathbf{C}_n| + \|p\|_{s-1,q,D} \\ &\leq K\big(\|[\mathbf{f}, h]\|_{s-2,q,D} + \|B(\kappa^{-1} g)\|_{s-1,q,D}\big). \end{aligned}$$

(c) *If D is convex, then $[\mathbf{u}, \sigma] = [\mathbf{u}_R, \sigma_R]$ satisfies the above inequality.*

To apply to our problem (1.7) known results for elliptic problems on the polygonal domain we define some solution operators corresponding to differential equations as follows. We will consider the Lamé system

$$\begin{aligned} \mathbf{L}\mathbf{u} := -\Delta \mathbf{u} - \nu_1 \nabla \mathrm{div} \mathbf{u} = \mathbf{f} \ \text{ in } D, \\ \mathbf{u} = 0 \ \text{ on } \partial D, \end{aligned} \tag{1.12}$$

where the parameter $\nu_1 > 0$. We define $\mathbf{A}_{\nu_1} : \mathbf{f} \mapsto \mathbf{u}$ to be the solution operator to this system. An energy argument shows that if $\mathbf{f} \in \mathbf{L}^2(D)$ (or $\mathbf{H}^{-1}(D)$), there is a unique solution $\mathbf{u} = \mathbf{A}_{\nu_1}\mathbf{f} \in \mathbf{H}_0^1(D)$ to (1.12). Using the operator $\mathbf{A}_{\nu_1}$ and setting $\nu_1 = 1 + \nu/\mu$, we define an operator $\mathcal{M}$ as follows:

$$\mathcal{M} = \Big(I + \mu^{-1}\mathbf{A}_{\nu_1}(\rho\, \mathbf{U} \cdot \nabla)\Big)^{-1} \mathbf{A}_{\nu_1}. \tag{1.13}$$

The operator $\mathcal{M}$ will be used in analyzing the momentum equations in (1.7). If $\mathbf{u} = \mu^{-1}\mathcal{M}\mathbf{F}$, then $\mathbf{u}$ satisfies the following problem: $\mathbf{L}_\beta \mathbf{u} := -\mu\Delta\mathbf{u} - (\mu+\nu)\nabla \mathrm{div}\mathbf{u} + \rho\, \mathbf{U} \cdot \nabla \mathbf{u} = \mathbf{F}$ in D, $\mathbf{u} = 0$ on ∂D.

In a similar manner, we define $A : \mathrm{L}^2(D)$ (or $\mathrm{H}^{-1}(D)$) $\longrightarrow \mathrm{H}_0^1(D)$ by $\sigma := Ah$, where σ is the solution of $-\Delta\sigma = h$ in D, $\sigma = 0$ on ∂D. Using A, we define an operator $\mathcal{E}$ as follows:

$$\mathcal{E} = \Big(I + \gamma^{-1} A(\tilde{\rho}\, \mathbf{U} \cdot \nabla)\Big)^{-1} A. \tag{1.14}$$

The operator $\mathcal{E}$ will be used in analyzing the energy equation in (1.7). If $\sigma = \gamma^{-1}\mathcal{E}H$, then σ satisfies $E\sigma := -\gamma\Delta\sigma + \tilde{\rho}\,\mathbf{U}\cdot\nabla\sigma = H$ in D, $\sigma = 0$ on ∂D.

For $1 < q < \infty$ and $1 \le s \le 2$ we let

$$\mathrm{Q}^{s,q}(D) := \{\chi \in \mathrm{H}^{s-1,q}(D) : \|\chi\|_{\mathrm{Q}^{s,q}(D)} < \infty\}$$

with $\|\chi\|_{\mathrm{Q}^{s,q}(D)} := \|\chi\|_{s-1,q,D} + \|\mathbf{U}\cdot\nabla\chi\|_{s-1,q,D}$ and $Q = Q^{s,q}$ if $s = 1$. We define the solution operator $B : \mathrm{H}^{s-1,q}(D) \longrightarrow \mathrm{Q}^{s,q}(D)$ by $\chi := BG$, where χ is the solution of

$$\begin{cases} \mathbf{U}\cdot\nabla\chi = G & \text{in } D, \\ \chi = 0 & \text{on } \partial D_{in}, \end{cases} \tag{1.15}$$

where the inflow boundary is $\partial D_{in} = \{(x,y) \in \Gamma : \mathbf{U}\cdot\mathbf{n} < 0\}$.

In considering the Lamé system (1.12) on a sector S, we recall the following. Let S be a sector in the plane with angle ω and placed at the origin. Consider

$$\mathbf{L}\mathbf{u} = \mathbf{f} \quad \text{in } S, \quad \mathbf{u} = 0 \quad \text{on } \Gamma := \partial S. \tag{1.16}$$

Let us consider the transcendental equations in [9, (3.1.22)–(3.1.23)]:

$$(1 + 2\nu_1^{-1})\sin(\lambda\omega) - \lambda\sin\omega = 0, \tag{1.17}$$

$$(1 + 2\nu_1^{-1})\sin(\lambda\omega) + \lambda\sin\omega = 0. \tag{1.18}$$

Note that the roots to (1.17) and (1.18) are, in general, complex, but the first several ones are real, depending on the angle of the vertex (for details, see section 2 and [9, Theorem 3.1.2]). From [9] we can see that if $\lambda := \lambda(\omega, \nu_1)$ solves either (1.17) or (1.18), then there are vector valued trigonometric functions $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$ (see [9, (3.128)–(3.129)]) such that (i) if $0 < \omega < \pi$, then $1 < \lambda_1 < \pi/\omega$, where λ_1 is the unique simple real root of (1.17), and

$$\Phi = r^{\lambda_1}\mathcal{T}_1(\theta) \tag{1.19}$$

satisfies $\mathbf{L}\Phi = 0$ and $\mathcal{T}_1 = 0$ at $\theta = 0, \omega$, and (ii) if $\pi < \omega < 2\pi$, then $1/2 < \lambda_1 < \pi/\omega < \lambda_2 < 1 < \lambda_3 < 2\pi/\omega$, where λ_1 is the real root of (1.18) and λ_2, λ_3 are the real roots of (1.17), and

$$\Phi_i = r^{\lambda_i}\mathcal{T}_i(\theta) \tag{1.20}$$

satisfies $\mathbf{L}\Phi_i = 0$ and $\mathcal{T}_i = 0$ at $\theta = 0, \omega$ for $i = 1, \ldots, 3$. Note that if $\nu_1 = 0$, then $\lambda_j = j\pi/\omega$ and $\Phi_j = r^{\lambda_j}\sin(\lambda_j\theta)\mathbf{C}$ with $\mathbf{C} = [1, 1]$.

Considering the above facts we can give a result for the Lamé system (1.16) on the concave sector S, which can be derived from [5, 9]: Let $q \ge 2$. (a) For each $i = 1, \ldots, 3$, there is a linear functional Λ_{i,ν_1} such that Λ_{i,ν_1} is bounded on $\mathbf{H}^{s-2,q}(S)$ for $s > \lambda_i + 2/q$ but not for $s \le \lambda_i + 2/q$. (b) If $\mathbf{f} \in \mathbf{H}^{s-2,q}(S)$, and $\mathbf{u} = \mathbf{A}_{\nu_1}\mathbf{f} = \mathbf{L}^{-1}\mathbf{f}$ is the solution of (1.16) with $\mathbf{u} \equiv 0$ for $r = \sqrt{x^2 + y^2} > 1$, then, for $i = 1, \ldots, 3$, if $\lambda_i + 2/q < s < \lambda_{i+1} + 2/q$ with $\lambda_4 = 2\pi/\omega$, then

$$\mathbf{u}_{R,i} := \mathbf{u}_{R,i-1} - \chi\,\Lambda_{i,\nu_1}(\mathbf{f})\Phi_i \in \mathbf{H}^{s,q}(S)$$

with $\|\mathbf{u}_{R,i}\|_{s,q,S} \le C\|\mathbf{f}\|_{s-2,q,S}$, where $\mathbf{u}_{R,0} = \mathbf{u}$.

In what follows, we denote by $\mathrm{L}^q(D)$ the space of all measurable functions u defined on D for which $\|u\|_{0,q,D} := (\int_D |u(\mathbf{x})|^q\,d\mathbf{x})^{1/q} < \infty$. If $q = 2$, we let $\|u\|_{0,D}$

denote the norm in $\mathrm{L}^2(D)$. If $q = \infty$, we define the norm of $\mathrm{L}^\infty(D)$ by $\|u\|_\infty := \sup\{|u(\mathbf{x})| : \mathbf{x} \in D\}$. For k a positive integer, the Sobolev space is defined as follows:

$$\mathrm{H}^{k,q}(D) = \left\{ v \in \mathrm{L}^q(D) : \|v\|_{k,q,D} := \left(\sum_{|\alpha| \le k} \|\nabla^\alpha v\|_{0,q,D}^q \right)^{1/q} < \infty \right\}.$$

If $q = 2$, we denote by $\mathrm{H}^{k,q}(D) = \mathrm{H}^k(D)$ and write $\|v\|_{k,q,D} = \|v\|_{k,D}$. For $s \ge 0$ we denote by $\mathrm{H}^{s,q}(\Omega)$ the space of all functions u defined in Ω such that $\|u\|_{s,q,\Omega} < \infty$, where $s = l + \sigma$ with $l = \lfloor s \rfloor$ and $0 \le \sigma < 1$. The norm is defined by

$$\|u\|_{s,q,\Omega} = \left\{ \|u\|_{l,q,\Omega}^q + \sum_{|\eta|=l} \iint_{\Omega\times\Omega} \frac{|D^\eta u(\mathbf{x}) - D^\eta u(\mathbf{y})|^q}{|\mathbf{x}-\mathbf{y}|^{2+q\sigma}} \,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \right\}^{1/q}.$$

We set $\mathrm{H}_0^{s,q}(D) = \mathrm{H}^{s,q}(D) \cap \mathrm{H}_0^1(D)$. We denote by $\mathrm{H}^{-1,q}(D)$ the dual space of $\mathrm{H}_0^{1,q'}(D)$ with norm

$$\|f\|_{-1,q,D} = \sup_{0 \ne v \in H_0^{1,q'}(D)} \frac{\langle f, v\rangle}{\|v\|_{1,q',D}},$$

where $\langle\ ,\ \rangle$ denotes the duality pairing. In a similar manner, for $0 < s < 1$ we denote by $\mathrm{H}^{-s,q}(D)$ the dual space of $\mathrm{H}_0^{s,q'}(D)$ with norm

$$\|f\|_{-s,q,D} = \sup_{0 \ne v \in \mathrm{H}_0^{s,q'}(D)} \frac{\langle f, v\rangle}{\|v\|_{s,q',D}}.$$

In this paper we will use the Sobolev imbedding theorems, $\mathrm{H}^{1,q}(D) \hookrightarrow \mathrm{L}^\infty(D)$ $(q > 2)$ and $\mathrm{H}^1(D) \hookrightarrow \mathrm{L}^4(D)$, and the trace theorem, $\mathrm{H}^{1,q}(D) \hookrightarrow \mathrm{L}^q(\partial D)$ $(1 \le q < \infty)$. Finally, we shall denote by the boldface $\mathbf{H}^{s,q}(D) = \mathrm{H}^{s,q}(D) \times \mathrm{H}^{s,q}(D)$ and often use the following notation: $\|[\mathbf{w}, \chi]\|_{s,q,D} = \|\mathbf{w}\|_{s,q,D} + \|\chi\|_{s,q,D}$.

2. Preliminary results. Our purpose in this section is to give some preliminary results concerning the elliptic and hyperbolic parts of the system (1.7). We first give some basic information concerning corner singularities for the Lamé operator, the Laplace operator, and the corresponding convection diffusion operators of (1.7). We then formulate some hypotheses concerning the vector field $\mathbf{U}$, we construct the resulting streamlines of $\mathbf{U}$, and we discuss the solution of the second equation in (1.7). Finally, we give a basic existence result for the system (1.7).

We need some notation concerning the geometry of D. Let $\{P_n\}$, $n = 1, \ldots, N$, denote the N vertices of D, with $P_n = (x_n, y_n)$. Let $r_n = [(x - x_n)^2 + (y - y_n)^2]^{\frac{1}{2}}$ denote the distance of a point to P_n. With the vertex P_n we associate two numbers $\omega_{n,1}$ and $\omega_{n,2}$, satisfying $\omega_{n,1} < \omega_{n,2} < \omega_{n,1} + 2\pi$. The two sides of D at P_n lie along the rays $(x_n + t\cos\omega_{n,1},\, y_n + t\sin\omega_{n,1})$ and $(x_n + t\cos\omega_{n,2},\, y_n + t\sin\omega_{n,2})$, $t \ge 0$. Let $\Gamma_{n,l}$, $l = 1, 2$, denote these two sides, and let $\Gamma_n := \Gamma_{n,1} \cup \Gamma_{n,2}$. Thus $\partial D = \bigcup_n \Gamma_n$ is the boundary of D. For $l = 1, 2$, let $\mathbf{n}_{n,l}$ denote the outward pointing normal to the side $\Gamma_{n,l}$. Thus, $\mathbf{n}_{n,l} = (-1)^l[-\sin\omega_{n,l}, \cos\omega_{n,l}]^T$. The interior angle of D at P_n is $\omega_n = \omega_{n,2} - \omega_{n,1}$. We define the following notation: for $s > 1$,

$$q_1(t) = \begin{cases} \frac{2}{s-1-t} & \text{if } t < s-1, \\ \infty & \text{if } t \ge s-1 \end{cases} \quad \text{and} \quad q_2(t) = \begin{cases} \frac{2}{s-t} & \text{if } t < s, \\ \infty & \text{if } t \ge s. \end{cases} \tag{2.1}$$

We shall need some information concerning the corner singularity expansion of the two elliptic boundary value problems that are imbedded in the system (1.7). We first give some notation which will also be used later. Let χ_n be a suitable smooth cutoff function which is 1 near the vertex P_n and which vanishes outside a small neighborhood of P_n. Let

$$\alpha_n = \pi/\omega_n, \quad \alpha_{i,n} = i\alpha_n, \quad \bar{s}_{i,n} = i\alpha_n + 2/q,$$
$$q_1^\alpha = \min_n\{q_1(\alpha_n)\}, \quad q_2^\alpha = \min_n\{q_2(\alpha_n)\},$$
$$\mathcal{J} = \{(i,n) : q_2(\alpha_{i,n}) \geq q_2^\alpha\}.$$

We start with the following result for the Laplace problem, which is proved in [5].

LEMMA 2.1. *Suppose $1 < q < q_1^\alpha$. Let $s \geq 1$ and $h \in \mathrm{H}^{s-2,q}(D)$. Let $\sigma = Ah$.*

(a) *If $1 \leq s < \min\{\bar{s}_{1,n}\}$, then A is a bounded operator from $\mathrm{H}^{s-2,q}(D)$ to $\mathrm{H}^{s,q}(D)$ and $\|\sigma\|_{s,q,D} \leq C\|h\|_{s-2,q,D}$.*

(b) *If $\mathcal{J} \neq \emptyset$, there is a bounded linear functional $\Lambda_{i,n}$ on $\mathrm{H}^{s-2,q}(D)$, $s > \bar{s}_{i,n}$ and a singular function $\phi_{i,n} \notin \mathrm{H}^{\bar{s}_{i,n},q}(D)$ such that*

$$\sigma_R := \sigma - \sum_{(i,n)\in\mathcal{J}} \Lambda_{i,n}(h)\phi_{i,n} \in \mathrm{H}^{s,q}(D) \tag{2.2}$$

with $\|\sigma_R\|_{s,q,D} \leq C\|h\|_{s-2,q,D}$ and

$$\sum_{(i,n)\in\mathcal{J}} |\Lambda_{i,n}(h)| \leq C\|h\|_{s-2,q,D}. \tag{2.3}$$

The singular function $\phi_{i,n}$ is given by the formula

$$\phi_{i,n}(x,y) = \chi_n\, r_n^{i\alpha_n} \sin[i\alpha_n(\theta - \omega_{n,1})]. \tag{2.4}$$

(c) *If $\mathcal{J} = \emptyset$, then $\sigma = \sigma_R \in \mathrm{H}^{s,q}(D)$ and satisfies $\|\sigma\|_{s,q,D} \leq C\|h\|_{s-2,q,D}$.*

We next consider the convection diffusion problem in the energy equation of (1.7):

$$\begin{aligned} E\sigma := -\gamma\Delta\sigma + \tilde{\rho}\,\mathbf{U}\cdot\nabla\sigma = H \ \text{ in } \ D,\\ \sigma = 0 \ \text{ on } \partial D. \end{aligned} \tag{2.5}$$

Using the operator $\mathcal{E}$ defined in (1.14), the solution σ of (2.5) is given by $\sigma = \gamma^{-1}\mathcal{E}\,H$. We associate with (2.5) the linear functional

$$\Lambda_{i,n,E}(H) = \gamma^{-1}\Lambda_{i,n}(H - \tilde{\rho}\,\mathbf{U}\cdot\nabla\sigma). \tag{2.6}$$

If the operator $\mathcal{E}$ is well defined, then in the right-hand side of (2.6) σ may be replaced by $\gamma^{-1}\mathcal{E}\,H$ so $\Lambda_{i,n,E}$ is indeed a function of H. The problem (2.5) is a special case of the system [13, (2.3)], and the following result is contained in [13, Theorem 2.1].

THEOREM 2.2. *Let $1 < q < q_1^\alpha$. Suppose that either γ is large or $|\tilde{\rho}\,\mathbf{U}|_\infty$ is small.*

(a) *If $1 \leq s < \min\{\bar{s}_{1,n}\}$ (or if $q < q_2^\alpha$), then $\mathcal{E}$ is a bounded operator from $\mathrm{H}^{s-2,q}(D)$ to $\mathrm{H}^{s,q}(D)$ and $\sigma = \gamma^{-1}\mathcal{E}H \in H^{s,q}(D)$ satisfies $\|\sigma\|_{s,q,D} \leq C\gamma^{-1} \times \|H\|_{s-2,q,D}$.*

(b) *If $\mathcal{J} \neq \emptyset$, the linear functional $\Lambda_{i,n,E}$ defined by (2.6) is bounded on $\mathrm{H}^{s-2,q}(D)$ for $s > \bar{s}_{i,n}$, the singular function $\phi_{i,n}$ given by (2.4) satisfies $\phi_{i,n} \notin \mathrm{H}^{\bar{s}_{i,n},q}(D)$, and*

$$\sigma_R := \sigma - \sum_{(i,n)\in\mathcal{J}} \Lambda_{i,n,E}(H)\phi_{i,n} \in \mathrm{H}^{s,q}(D) \tag{2.7}$$

with $\|\sigma_R\|_{s,q,D} \le C\gamma^{-1}\|H\|_{s-2,q,D}$ *and*

$$\sum_{(i,n)\in\mathcal{J}} |\Lambda_{i,n,E}(H)| \le C\|H\|_{s-2,q,D}. \tag{2.8}$$

(c) *If* $\mathcal{J} = \emptyset$, *then* $\sigma = \sigma_R \in \mathrm{H}^{s,q}(D)$ *and satisfies* $\|\sigma\|_{s,q,D} \le C\gamma^{-1}\|H\|_{s-2,q,D}$.

Next we describe the formulas of the singular functions for the Lamé system (1.12) and define numbers for the regularities of the singular functions. To describe these formulas, consider the following transcendental equations given in [9, (3.1.22)–(3.1.23)]:

$$(1 + 2\nu_1^{-1})\sin(\lambda\omega_n) - \lambda \sin\omega_n = 0, \tag{2.9}$$

$$(1 + 2\nu_1^{-1})\sin(\lambda\omega_n) + \lambda \sin\omega_n = 0. \tag{2.10}$$

Equations (2.9) and (2.10) have an infinite number of complex solutions. Ordering these solutions with nondecreasing real part, we get a nondecreasing sequence of numbers $\lambda_{i,n}$, $i = 1, 2, \ldots$. The numbers $s_{i,n}$ are given by $s_{i,n} = \mathcal{R}e\lambda_{i,n} + 2/q$, $i = 1, 2, \ldots$. The singular function $\Phi_{i,n}$ has the form

$$\Phi_{i,n} = \chi_n\, r_n^{\lambda_{i,n}}\, \mathcal{T}_{i,n}(\theta), \tag{2.11}$$

where $\mathcal{T}_{i,n}(\theta)$ is a vector of trigonometric functions.

Some more information is available concerning the numbers $\lambda_{i,n}$ [9]. If the vertex P_n is convex, so $\omega_n < \pi$, then $\lambda_{1,n}$ is real and $1 < \lambda_{1,n} < \alpha_n$. Hence if $q > 2$ and q is sufficiently close to 2, then $s_{1,n} > 2$. If the vertex P_n is concave, so $\omega_n > \pi$, then the first 3 roots, $\lambda_{i,n}$ for $i = 1, 2, 3$, are real and satisfy the inequalities $\frac{1}{2} < \lambda_{1,n} < \alpha_n < \lambda_{2,n} < 1 < \lambda_{3,n} < 2\alpha_n$. Hence if $q > 2$, $\frac{3}{2} < s_{1,n} < \bar{s}_{1,n} < s_{2,n} < 2$. Furthermore, if q is sufficiently close to 2, then $s_{3,n} > 2$. In this paper, our goal is to obtain enough terms of a corner singularity expansion so that the remainder is in $\mathrm{H}^{2,q}$ for some number $q > 2$. It follows that we will not need any singular functions corresponding to a convex vertex, and we will need two "velocity" singular functions and one "temperature" singular function corresponding to each concave vertex.

We now state a result for the Lamé system (1.12), which can be derived from [5, 9]. We give some notation which will also be used later:

$$q_1^\lambda = \min_n\{q_1(\lambda_{1,n})\}, \quad q_2^\lambda = \min_n\{q_2(\lambda_{1,n})\},$$
$$\mathcal{I} = \{(i,n) : q_2(\lambda_{i,n}) \ge q_2^\lambda\}.$$

LEMMA 2.3. *Suppose* $1 < q < q_1^\lambda$. *Let* $\mathbf{f} \in \mathbf{H}^{s-2,q}(D)$.

(a) *If* $s < \min\{s_{1,n}\}$, *then* $\mathbf{A}_{\nu_1}$ *is a bounded operator from* $\mathbf{H}^{s-2,q}(D)$ *to* $\mathbf{H}^{s,q}(D)$ *and* $\mathbf{u} = \mathbf{A}_{\nu_1}\mathbf{f} \in \mathbf{H}^{s,q}(D)$ *satisfies* $\|\mathbf{u}\|_{s,q,D} \le C\|\mathbf{f}\|_{s-2,q,D}$.

(b) *If* $\mathcal{I} \ne \emptyset$, *there is a bounded linear functional* Λ_{i,n,ν_1} *on* $\mathbf{H}^{s-2,q}(D)$, $s > s_{i,n}$, *and a singular function* $\Phi_{i,n} \notin \mathbf{H}^{s_{i,n},q}(D)$ *such that*

$$\mathbf{u}_R := \mathbf{u} - \sum_{(i,n)\in\mathcal{I}} \Lambda_{i,n,\nu_1}(\mathbf{f})\Phi_{i,n} \in \mathbf{H}^{s,q}(D) \tag{2.12}$$

with $\|\mathbf{u}_R\|_{s,q,D} \le C\|\mathbf{f}\|_{s-2,q,D}$.

(c) *If* $\mathcal{I} = \emptyset$, *then* $\mathbf{u} = \mathbf{u}_R \in \mathbf{H}^{s,q}(D)$ *and satisfies* $\|\mathbf{u}\|_{s,q,D} \le C\|\mathbf{f}\|_{s-2,q,D}$.

We now investigate the behavior of weak solutions of the boundary value problem

$$\begin{aligned} \mathbf{L}_\beta\mathbf{u} := -\mu\,\Delta\mathbf{u} - (\mu+\nu)\nabla\mathrm{div}\mathbf{u} + \rho\mathbf{U}\cdot\nabla\mathbf{u} = \mathbf{F} \ \text{ in } \ D,\\ \mathbf{u} = 0 \ \text{ on } \partial D. \end{aligned} \tag{2.13}$$

If $\mathbf{u}$ solves the problem (2.13) we write $\mathbf{u} = \mu^{-1}\mathcal{M}\mathbf{F}$. The next lemma shows that with some conditions on the coefficients of (2.13), the operator $\mathcal{M}$ is well defined and bounded in appropriate norms.

LEMMA 2.4. *Suppose $1 < q < q_1^\lambda$ (or $1 \le s < \lambda_{1,n}+1+2/q$). If μ is large enough or if $|\rho\,\mathbf{U}|_\infty$ is small enough, then the operator $\mathcal{M}$ is a well-defined bounded map from $\mathbf{H}^{s-3,q}(D)$ to $\mathbf{H}_0^{s-1,q}(D)$. In addition, $\|\mathcal{M}\| \le C\mu^{-1}$ uniformly for large μ.*

Proof. The equation $\mathbf{L}_\beta\mathbf{u} = \mathbf{F}$ may be written $-\Delta\mathbf{u} - (1+\mu^{-1}\nu)\mathrm{div}\mathbf{u} = \mu^{-1}\mathbf{F} - \mu^{-1}\rho\mathbf{U}\cdot\nabla\mathbf{u}$. Set $\nu_1 = 1+\mu^{-1}\nu$. If μ is large, then $0 < \nu_1 < \nu_*$ for some number ν_* and for each number ν_1 the map $\mathbf{A}_{\nu_1}$ is a bounded operator from $\mathbf{H}^{s-3,q}(D)$ to $\mathbf{H}^{s-1,q}(D)$ for the Lamé system (1.12). If $\mathbf{u} \in \mathbf{H}_0^{s-1,q}(D)$ is a weak solution of (2.13), we have $\mathbf{u} = \mu^{-1}\mathbf{A}_{\nu_1}(\mathbf{F} - \rho\,\mathbf{U}\cdot\nabla\mathbf{u})$. Hence, if $\mathbf{F} \in \mathbf{H}^{s-3,q}(D)$, then $\|\mathbf{u}\|_{s-1,q,D} \le C\mu^{-1}\|\mathbf{F}\|_{s-3,q,D} + C\mu^{-1}|\rho\,\mathbf{U}|_\infty\|\mathbf{u}\|_{s-2,q,D}$. If μ is large enough, or if $|\rho\,\mathbf{U}|_\infty$ is small enough, we obtain a bound for $\|\mathbf{u}\|_{s-1,q,D}$. A fixed point argument then gives the existence of a weak solution for any $\mathbf{F}$. ∎

We associate with (2.13) the following linear functionals:

$$\Lambda_{i,n,L}(\mathbf{F}) = \mu^{-1}\Lambda_{i,n,\nu_1}(\mathbf{F} - \rho\,\mathbf{U}\cdot\nabla\mathbf{u}), \tag{2.14}$$

where $\nu_1 = 1+\mu^{-1}\nu$. In the formula for $\Lambda_{i,n,L}(\mathbf{F})$, $\mathbf{u}$ is regarded as a solution of (2.13), and in this way the functional is regarded as a linear function on $\mathbf{F}$. It is used in the corner singularity expansion associated with the problem (2.13).

THEOREM 2.5. *Assume that the coefficients of $\mathbf{L}_\beta$ satisfy the hypotheses of Lemma 2.4. Suppose $1 < q < q_1^\lambda$. Let $\mathbf{F} \in \mathbf{H}^{s-2,q}(D)$.*

(a) *If $1 \le s < \min\{s_{1,n}\}$, then $\mathcal{M}$ is a bounded operator from $\mathbf{H}^{s-2,q}(D)$ to $\mathbf{H}^{s,q}(D)$ and $\mathbf{u} = \mu^{-1}\mathcal{M}\mathbf{F} \in \mathbf{H}^{s,q}(D)$ satisfies $\|\mathbf{u}\|_{s,q,D} \le C\mu^{-1}\|\mathbf{F}\|_{s-2,q,D}$.*

(b) *If $\mathcal{I} \ne \emptyset$, there is a bounded linear functional $\Lambda_{i,n,L}$ on $\mathbf{H}^{s-2,q}(D)$, $s > s_{i,n}$, and a singular function $\Phi_{i,n} \notin \mathbf{H}^{s_{i,n},q}(D)$ such that*

$$\mathbf{u}_R := \mathbf{u} - \sum_{(i,n)\in\mathcal{I}} \Lambda_{i,n,L}(\mathbf{F})\Phi_{i,n} \in \mathbf{H}^{s,q}(D) \tag{2.15}$$

with $\|\mathbf{u}_R\|_{s,q,D} \le C\mu^{-1}\|\mathbf{F}\|_{s-2,q,D}$ and

$$\sum_{(i,n)\in\mathcal{I}} |\Lambda_{i,n,L}(\mathbf{F})| \le C\|\mathbf{F}\|_{s-2,q,D}. \tag{2.16}$$

(c) *If $\mathcal{I} = \emptyset$, then $\mathbf{u} = \mathbf{u}_R \in \mathbf{H}^{s,q}(D)$ and satisfies $\|\mathbf{u}\|_{s,q,D} \le C\mu^{-1}\|\mathbf{F}\|_{s-2,q,D}$.*

Proof. If $s < \min\{s_{1,n}\}$, then $\mathbf{u} \in \mathbf{H}^{s,q}(D)$, so $\mathbf{f} = \mu^{-1}(\mathbf{F} - \rho\,\mathbf{U}\cdot\nabla\mathbf{u})$ belongs to $\mathbf{H}^{s-2,q}(D)$. Since $\mathbf{u} = \mathbf{A}_{\nu_1}\mathbf{f}$, the result follows from Lemma 2.4. ∎

By a streamline of the vector field $\mathbf{U}$ we mean a curve $(x, k(x))$, where the function k satisfies $k'(x) = U(x,k(x))^{-1}V(x,k(x))$. We now make some assumptions on the vector field $\mathbf{U}$ that will be used in the linear analysis.

Assumption A.

(A1) The vector field $\mathbf{U} = [U, V]$ is of the form $\mathbf{U} = \sum_{(i,n)\in\mathcal{I}} d_{i,n}\Phi_{i,n} + \mathbf{U}_R$, with $\mathbf{U}_R \in \mathbf{H}^{2,q}(D)$ and $d_{i,n}$ given numbers, and is continuous on $\bar{D}$, Lipschitz continuous at each point of $\bar{D}$ with the exception of the vertices.

(A2) There is a constant $C_0 > 0$ such that $U > C_0$.

(A3) For each $i = 1, \dots, N$ and $l = 1, 2$, the quantity

$$(-1)^l[-U(r_i\cos\omega_{i,l},\, r_i\sin\omega_{i,l})\sin\omega_{i,l} + V(r_i\cos\omega_{i,l},\, r_i\sin\omega_{i,l})\cos\omega_{i,l}]$$

is nonnegative and in absolute value bounded below by C_0.

(A4) Each streamline generated by $\mathbf{U}$ intersects the boundary of D at only two points, and at most one of these points is a vertex.

Because of (A1), the streamlines of $\mathbf{U}$ are well-defined curves. (A2) implies that the streamlines may be parametrized by x. (A3) means that the vector field $\mathbf{U}$ is either always tangent to $\Gamma_{i,l}$ or never tangent to $\Gamma_{i,l}$, $l = 1, 2$. If the quantity in (A3) is *negative* for all $r_i > 0$, we say that $\Gamma_{i,l}$ is an incoming side. We write ∂D_{in} for the union of the incoming sides of D. For convenience we will assume that the curve ∂D_{in} is given by a piecewise linear function $x = \delta(y)$ with δ an increasing function of y.

We study the first order partial differential equation

$$\begin{cases} \mathbf{U} \cdot \nabla p = G \ \text{ in } D, \\ p = 0 \ \text{ on } \partial D_{in}. \end{cases} \tag{2.17}$$

The solution to problem (2.17) is obtained by integrating along the streamlines of the vector field $\mathbf{U} = [U, V]$. These streamlines may be written $(x, k(x, \bar{y}))$, where k satisfies the differential equation

$$k_x(x, \bar{y}) = U^{-1}V(x, k) \quad \text{and} \quad k(\delta(\bar{y}), \bar{y}) = \bar{y}. \tag{2.18}$$

Thus $(x, k(x, \bar{y}))$ gives the streamline emanating from the point $(\delta(\bar{y}), \bar{y})$ on ∂D_{in}. From the theory of differential equations, if the functions U and V are Lipschitz continuous, the solution k to (2.18) exists, is unique, and is continuously differentiable in x and continuous in $\bar{y}$. Also by (A3), the vector field $\mathbf{U}$ is not tangent to the curve $x = \delta(y)$, so k is a strictly monotone function of $\bar{y}$. Hence for fixed x the equation $y = k(x, \bar{y})$ has a well-defined solution, which we write $\bar{y} = \xi(x, y)$.

In the situation considered in this paper, the vector field $\mathbf{U}$ comes from the vector field $\mathbf{u}$, and therefore the Lipschitz continuity may be a concern. In particular, we expect that $\mathbf{u}$ may have a singularity at the vertices. However, from Lemma 2.3 we expect or hope that this singularity has a well-defined character. This is the reason for the particular form of $\mathbf{U}$ that is specified in (A1). The next lemma gives some properties of the streamlines under what will turn out to be appropriate hypotheses on the vector field. The proof is given in [13, Lemma 2.3].

LEMMA 2.6. *If $1 < q < q_2^\lambda$, let $\mathbf{U} \in \mathbf{H}^{2,q}(D)$. If $\mathcal{I} \neq \emptyset$, let $\mathbf{U} = \sum_{(i,n)\in\mathcal{I}} \mathbf{U}_{i,n} + \mathbf{U}_R$, with $\mathbf{U}_{i,n} = d_{i,n}\Phi_{i,n}$, where $\mathbf{U}_R \in \mathbf{H}^{2,q}(D)$ $(2 < q < q_1^\lambda)$, $d_{i,n}$ is given, and $\Phi_{i,n}$ is given in* (2.11). *Then the functions k and ξ are well defined and continuously differentiable on the polygon D. Furthermore, k, ξ, and the first derivatives of k and ξ are bounded by a constant that depends only on D, C_0, and $\|\mathbf{U}_R\|_{2,q,D} + \sum_{(i,n)\in\mathcal{I}} |d_{i,n}|$.*

Using the function $k(x, \bar{y})$ and its inverse function $\xi(x, y)$ the solution of (2.17) is given by the formula

$$p(x, y) = \int_{\delta(\xi(x,y))}^{x} \bar{G}(s, k(s, \xi))\, ds, \tag{2.19}$$

where $\bar{G} = U^{-1}G$. The formula (2.19) defines the solution operator $B : G \mapsto BG$ of (2.17). In the next lemma the map B is shown to be bounded on $H^{s,q}(\Omega)$ for $0 \le s \le 1$.

LEMMA 2.7. *Let $1 \le q < \infty$ and $0 \le s \le 1$. Assume $G \in \mathrm{H}^{s,q}(D)$. Then $p = BG$ satisfies*

$$\|BG\|_{s,q,D} \le C\|G\|_{s,q,D}, \tag{2.20}$$

where $C = C(D, C_0, \|\mathbf{U}_R\|_{2,q,D} + \sum_{(i,n)\in\mathcal{I}} |d_{i,n}|)$.

Proof. This result is given in [13, Lemma 2.4] in the cases $s = 0$ and $s = 1$. For the intermediate values of s it follows by interpolation. ∎

LEMMA 2.8. *Suppose* $\mathcal{I} \neq \emptyset$. *Set* $\mathbf{u}_s = \sum_{(i,n)\in\mathcal{I}} C_{i,n}\Phi_{i,n}$ *with given numbers* $C_{i,n}$. *Suppose the regularity exponent* s *satisfies*

$$s \le \min\Big\{2,\ 1/q + 1 + \min_n\{\lambda_{1,n}\}\Big\}.$$

Then $\nabla B(\kappa_1^{-1}\nabla\mathbf{u}_s) \in \mathrm{H}^{s-2,q}(D)$ *and satisfies the following inequality:*

$$\|B(\kappa_1^{-1}\nabla\mathbf{u}_s)\|_{s-1,q,D} \le K \sum_{(i,n)\in\mathcal{I}} |C_{i,n}|, \tag{2.21}$$

where $K = C(C_0, \|\kappa_1^{-1}\|_{1,q,D}, \|\mathbf{U}_R\|_{2,q,D} + \sum_{(i,n)\in\mathcal{I}} |d_{i,n}|)$.

Proof. Set $g_s = \nabla\mathbf{u}_s$ and $G_s = \kappa_1^{-1}\nabla\mathbf{u}_s$ for simplicity. Using (1.15) and (2.18) we have

$$(BG_s)(x,y) = \int_{\delta(\xi(x,y))}^{x} U^{-1}G_s(s,k(s,\xi))\,ds. \tag{2.22}$$

Obviously,

$$\|BG_s\|_{0,q,D} \le C|\kappa_1^{-1}|_\infty \sum_{(i,n)\in\mathcal{I}} |C_{i,n}|.$$

Next the derivative of the function BG_s with respect to the y variable is

$$\begin{aligned}(\nabla_y BG_s)(x,y) = &\int_{\delta(\xi)}^{x} \big[\nabla_y(\kappa_1^{-1})\,U^{-1}g_s(s,k(s,\xi))\\ &\qquad + \kappa_1^{-1}\nabla_y(U^{-1})\,g_s(s,k(s,\xi)) + (\kappa_1 U)^{-1}\nabla_y g_s(s,k(s,\xi))\big]\,ds\\ &+ (\kappa_1 U)^{-1}g_s(\delta(\xi),\xi)\,\delta'(\xi)\,\xi_y(x,y)\\ &= I + II + III + IV. \end{aligned} \tag{2.23}$$

One has to be careful in estimating $\|I\|_{0,q,D}$. So

$$|I| \le C_0\bigg(\int_{\delta(\xi)}^{x} |\nabla_y\kappa_1^{-1}(s,k(s,\xi))|^q ds\bigg)^{\frac{1}{q}}\bigg(\int_{\delta(\xi)}^{x} |g_s(s,k(s,\xi))|^{q'} ds\bigg)^{\frac{1}{q'}}. \tag{2.24}$$

If we set $\bar{y} = \xi(x,y)$ and $t = s/\bar{y}$, then, near each vertex $P_n = (x_n, y_n)$,

$$\begin{aligned}\int_{\delta(\xi)}^{x} |\nabla_y\Phi_{1,n}(s,h(s,\xi))|^{q'} ds &\le C\int_{\delta(\bar{y})}^{x} \big(s^2 + k(s,\bar{y})^2\big)^{\frac{q'(\lambda_{1,n}-1)}{2}}\,ds\\ &\le C\,\bar{y}^{q'(\lambda_{1,n}-1)+1}\int_{\delta(\bar{y})/\bar{y}}^{x/\bar{y}} \big(t^2 + k(t\bar{y},\bar{y})^2/\bar{y}^2\big)^{\frac{q'(\lambda_{1,n}-1)}{2}}\,dt\\ &\le C\,\bar{y}^{q'(\lambda_{1,n}-1)+1}\int_{\delta(\bar{y})/\bar{y}}^{x/\bar{y}} t^{q'(\lambda_{1,n}-1)}\,dt\\ &\le C\big(x^{q'(\lambda_{1,n}-1)+1} - \delta(\bar{y})^{q'(\lambda_{1,n}-1)+1}\big) < \infty,\end{aligned}$$

because $q'(\lambda_{1,n}-1)+1>0$. So integrating both sides of (2.24), we have

$$\|I\|_{0,q,D} \le C\|\nabla_y \kappa_1^{-1}\|_{0,q,D} \sum_{(i,n)\in\mathcal{I}} |C_{i,n}|. \tag{2.25}$$

Similarly,

$$\|II\|_{0,q,D} \le C|\kappa_1^{-1}|_{0,\infty}\left(\|\mathbf{U}_R\|_{2,q,D} + \sum_{(i,n)\in\mathcal{I}} |d_{i,n}|\right) \sum_{(i,n)\in\mathcal{I}} |C_{i,n}|. \tag{2.26}$$

Applying the same procedures as given in [12, Lemma 2.5] near each vertex P_n, we have

$$|III| \le C|\kappa_1^{-1}|_\infty \sum_{(i,n)\in\mathcal{I}} |C_{i,n}|(|x|^{\lambda_{1,n}-1} + |\xi_n(x,y)|^{\lambda_{1,n}-1}). \tag{2.27}$$

Integrating both sides of (2.27) on D, we have

$$\begin{aligned}
\|III\|_{0,q,D} &\le C \sum_{(i,n)\in\mathcal{I}} |C_{i,n}|\left[|\Omega_n|^{1/q} + \left(\iint_{\Omega_n} |\xi_n(x,y)|^{(\lambda_{1,n}-1)q}\,dxdy\right)^{1/q}\right] \\
&\quad \text{(letting } \bar{y}=\xi_n(x,y) \text{ and } dy = k_{n,\bar{y}}\,d\bar{y}) \\
&\le C \sum_{(i,n)\in\mathcal{I}} |C_{i,n}|\left[|\Omega_n|^{1/q} + \left(\int_{\bar{y}=0}^{1} |\bar{y}|^{(\lambda_{1,n}-1)q}\,d\bar{y}\right)^{1/q}\right] \\
&\le C \sum_{(i,n)\in\mathcal{I}} |C_{i,n}|, \qquad (2.28)
\end{aligned}$$

where $C = C(\|\kappa_1^{-1}\|_{1,q,D},\ \|\mathbf{U}_R\|_{2,q,D} + \sum_{(i,n)\in\mathcal{I}} |d_{i,n}|)$. It is clear that $\|IV\|_{0,q,D} \le C\sum_{(i,n)\in\mathcal{I}} |C_{i,n}|$. A similar argument is valid for the quantity $\|\nabla_x(BG_s)\|_{0,q,D}$. Consequently, $\|\nabla BG_s\|_{0,q,D}$ is estimated by the right-hand side of (2.28). Since $\|BG_s\|_{s-1,q,D} \le C\|BG_s\|_{1,q,D}$, (2.21) follows. □

Finally, we discuss system (1.7) on the bounded (convex or nonconvex) polygonal domain D. To do this, first we define bilinear forms $a(\mathbf{v},\mathbf{w})$ on $\mathbf{H}_0^1 \times \mathbf{H}_0^1$ and $e(\sigma,\eta)$ on $\mathrm{H}_0^1 \times \mathrm{H}_0^1$ and bilinear forms $b(\chi,\mathbf{v})$ and $\tilde{b}(\chi,\mathbf{v})$ on $\mathrm{L}^2 \times \mathbf{H}_0^1$ as follows:

$$\begin{aligned}
a(\mathbf{u},\mathbf{v}) &= \int_D \mu\nabla\mathbf{u}\cdot\nabla\mathbf{v} + (\mu+\nu)\mathrm{div}\mathbf{u}\,\mathrm{div}\mathbf{v}\,\mathrm{d}\mathbf{x} + \int_D \rho\,\mathbf{U}\cdot\nabla\mathbf{u}\,\mathbf{v}\,\mathrm{d}\mathbf{x}, \\
b(\chi,\mathbf{v}) &= -\int_D \chi\,\mathrm{div}\mathbf{v}\,\mathrm{d}\mathbf{x}, \\
e(\sigma,\eta) &= \int_D \gamma\nabla\sigma\cdot\nabla\eta\,\mathrm{d}\mathbf{x} + \int_D \tilde{\rho}\,\mathbf{U}\cdot\nabla\sigma\,\eta\,\mathrm{d}\mathbf{x}, \\
\tilde{b}(\chi,\mathbf{v}) &= \int_D \tilde{\tau}\,\chi\,\mathrm{div}\mathbf{v}\,\mathrm{d}\mathbf{x}.
\end{aligned}$$

Using these forms, the first and third equations in (1.7) imply

$$a(\mathbf{u},\mathbf{v}) + b(p,\mathbf{v}) = \langle \mathbf{f},\mathbf{v}\rangle \quad \forall \mathbf{v}\in\mathbf{H}_0^1(D), \tag{2.29}$$

$$e(\sigma,\eta) + \tilde{b}(\eta,\mathbf{u}) = \langle h,\eta\rangle \quad \forall \eta\in\mathrm{H}_0^1(D). \tag{2.30}$$

Second, using the solution operator B, the second equation in (1.7) implies

$$p = B\bar{g} - B(\kappa_1^{-1}\mathrm{div}\mathbf{u}) - B(\bar{\mathbf{U}}\cdot\nabla\sigma), \tag{2.31}$$

where $\bar{g} = \kappa_1^{-1}g$ and $\bar{\mathbf{U}} = \kappa_1^{-1}\kappa_2\mathbf{U}$. In addition, using the solution operator $\mathcal{E}$ defined in (1.14), the solution σ of (2.30) is

$$\sigma = \gamma^{-1}\mathcal{E}(h - \tilde{\tau}\,\mathrm{div}\mathbf{u}). \tag{2.32}$$

Combining (2.31) and (2.32) we have

$$p = Bg_* - B(\kappa_1^{-1}\mathrm{div}\mathbf{u}) + \gamma^{-1}B[\bar{\mathbf{U}}\cdot\nabla\mathcal{E}(\tilde{\tau}\mathrm{div}\mathbf{u})], \tag{2.33}$$

where $g_* = \bar{g} - \gamma^{-1}\bar{\mathbf{U}}\cdot\nabla\mathcal{E}h$. This formula may be used to eliminate p from (2.29). We obtain

$$\begin{aligned}\tilde{a}(\mathbf{u},\mathbf{v}) &:= a(\mathbf{u},\mathbf{v}) - b(B(\kappa_1^{-1}\mathrm{div}\mathbf{u}),\mathbf{v}) + \gamma^{-1}b(B\tilde{\mathbf{u}},\mathbf{v})\\ &= \langle\mathbf{f},\mathbf{v}\rangle - b(Bg_*,\mathbf{v}),\end{aligned} \tag{2.34}$$

where $\tilde{\mathbf{u}} := \bar{\mathbf{U}}\cdot\nabla\mathcal{E}(\tilde{\tau}\mathrm{div}\mathbf{u})$. We define a weak solution $[\mathbf{u},p,\sigma]$ to the problem (1.7) to be a function $\mathbf{u}\in\mathbf{H}_0^1(D)$ satisfying

$$\tilde{a}(\mathbf{u},\mathbf{v}) = \langle\mathbf{f},\mathbf{v}\rangle - b(Bg_*,\mathbf{v}) \quad \forall\mathbf{v}\in\mathbf{H}_0^1(D) \tag{2.35}$$

and a function $p\in\mathrm{L}^2(D)$ satisfying (2.33) and a function $\sigma\in\mathrm{H}_0^1(D)$ satisfying (2.32). Note that if $[\mathbf{u},p,\sigma]$ is a weak solution of (1.7), then the second equation of (1.7) holds.

In the next lemma we establish the unique solvability of the problem (1.7), provided μ is large enough. Furthermore, we show that the solution has a regularity dictated by the largest angle of the vertices of D and the ratio $\mu^{-1}\nu$, and we give an inequality corresponding to this regularity.

LEMMA 2.9. *Let $1\le s<\min\{s_{1,n}\}$ and $s\le 2$. Suppose that μ and γ are sufficiently large. Then there is a unique solution $[\mathbf{u},p,\sigma]\in\mathbf{H}^{s,q}(D)\times\mathrm{H}^{s-1,q}(D)\times\mathrm{H}^{s,q}(D)$ of (1.7). Furthermore, there exists a constant K remaining finite for large μ such that if $[\mathbf{f},h]\in\mathbf{H}^{s-2,q}(D)\times\mathrm{H}^{s-2,q}(D)$ and $g\in\mathrm{H}^{s-1,q}(D)$, then*

$$\|[\mathbf{u},\sigma]\|_{s,q,D} + \|p\|_{s-1,q,D} \le K\big(\|[\mathbf{f},h]\|_{s-2,q,D} + \|B(\kappa_1^{-1}g)\|_{s-1,q,D}\big). \tag{2.36}$$

Proof. Using the operator $\mathcal{M}$, the solution $\mathbf{u}$ of (2.29) is given by

$$\mathbf{u} = \mu^{-1}\mathcal{M}(\mathbf{f} - \nabla p).$$

Replacing the function p by (2.33) we have

$$(I - \mu^{-1}\mathcal{M}_1)\mathbf{u} = \mu^{-1}\mathcal{M}(\mathbf{f} - \nabla Bg_*), \tag{2.37}$$

where $g_* = \bar{g} - \gamma^{-1}\bar{\mathbf{U}}\cdot\nabla\mathcal{E}h$ and

$$\mathcal{M}_1 := \mathcal{M}\nabla\big[B(\kappa_1^{-1}\nabla\cdot) - \gamma^{-1}B(\bar{\mathbf{U}}\cdot\nabla\mathcal{E}(\tilde{\tau}\nabla\cdot))\big]. \tag{2.38}$$

Using Lemma 2.7, Theorem 2.2, and Theorem 2.5 it follows that the mappings in the following two sequences are bounded, provided $s<\min\{s_{1,n}\}$:

$$\mathbf{H}^{s,q}\xrightarrow{\tilde{\tau}\nabla\cdot}\mathrm{H}^{s-1,q}\xrightarrow{\mathcal{E}}\mathrm{H}^{s,q}\xrightarrow{\bar{\mathbf{U}}\cdot\nabla}\mathrm{H}^{s-1,q}\xrightarrow{B}\mathrm{H}^{s-1,q}\xrightarrow{\nabla}\mathbf{H}^{s-2,q}\xrightarrow{\mathcal{M}}\mathbf{H}^{s,q},$$
$$\mathbf{H}^{s,q}\xrightarrow{\kappa_1^{-1}\nabla\cdot}\mathrm{H}^{s-1,q}\xrightarrow{B}\mathrm{H}^{s-1,q}\xrightarrow{\nabla}\mathbf{H}^{s-2,q}\xrightarrow{\mathcal{M}}\mathbf{H}^{s,q}.$$

So $\mathcal{M}_1$ is a bounded mapping from $\mathbf{H}^{s,q}$ to $\mathbf{H}^{s,q}$. If μ is large enough, then $I-\mu^{-1}\mathcal{M}_1$ is invertible and using (2.32) we have

$$\begin{aligned}\mathbf{u} &= \mu^{-1}\mathcal{M}_1^*\mathcal{M}(\mathbf{f}-\nabla Bg_*)\\ &= \mu^{-1}\mathcal{M}_1^*\mathcal{M}[\mathbf{f}-\nabla B\bar{g}-\gamma^{-1}\nabla B(\bar{\mathbf{U}}\cdot\nabla\mathcal{E})h],\end{aligned}\tag{2.39}$$

which belongs to $\mathbf{H}^{1,q}(D)$, where $\mathcal{M}_1^* := (I-\mu^{-1}\mathcal{M}_1)^{-1}$. Using this, the functions p and σ are determined by (2.33) and (2.32), respectively. On the other hand, since

$$\|Bg_*\|_{s-1,q,D} \le \|B(\kappa_1^{-1}g)\|_{s-1,q,D} + \gamma^{-1}\|B\bar{\mathbf{U}}\cdot\nabla\mathcal{E}\|\|h\|_{s-2,q,D},$$

we have

$$\begin{aligned}\mu\|\mathbf{u}\|_{s,q,D} &\le c_1(\|\mathbf{f}\|_{s-2,q,D}+\|B(k_1^{-1}g)\|_{s-1,q,D})+c_2\|h\|_{s-2,q,D}\\ &\le \max\{c_1,c_2\}\big(\|[\mathbf{f},h]\|_{s-2,q,D}+\|B(k_1^{-1}g)\|_{s-1,q,D}\big),\end{aligned}\tag{2.40}$$

where $c_1=\max\{\|\mathcal{M}_1^*\mathcal{M}\|,\|\mathcal{M}_1^*\mathcal{M}\nabla\|\}$ and $c_2=\gamma^{-1}c_1\|B\bar{\mathbf{U}}\cdot\nabla\mathcal{E}\|$. Also it follows from (2.32)–(2.33) that

$$\|p\|_{s-1,q,D} \le c_3\|\mathrm{div}\mathbf{u}\|_{s-1,q,D}+\|B\bar{g}\|_{s-1,q,D}+c_4\|h\|_{s-2,q,D},\tag{2.41}$$

$$\begin{aligned}\|\sigma\|_{s,q,D} &\le c_5\|\tilde{\tau}\mathrm{div}\mathbf{u}\|_{s-2,q,D}+\gamma^{-1}\|\mathcal{E}\|\|h\|_{s-2,q,D}\\ &\le c_5\|\tilde{\tau}\mathrm{div}\mathbf{u}\|_{0,q,D}+\gamma^{-1}\|\mathcal{E}\|\|h\|_{s-2,q,D}\\ &\le c_6\|\mathbf{u}\|_{1,q,D}+\gamma^{-1}\|\mathcal{E}\|\|h\|_{s-2,q,D}\\ &\le c_6\|\mathbf{u}\|_{s,q,D}+\gamma^{-1}\|\mathcal{E}\|\|h\|_{s-2,q,D},\end{aligned}\tag{2.42}$$

where $c_3=C(\|\kappa_1^{-1}\|_{1,q,D}\|B\|+\gamma^{-1}|\tilde{\tau}|_\infty\|B\bar{\mathbf{U}}\cdot\nabla\mathcal{E}\|)$, $c_4=\gamma^{-1}\|B\bar{\mathbf{U}}\cdot\nabla\mathcal{E}\|$, $c_5=C\gamma^{-1}\|\mathcal{E}\|$, and $c_6=C\gamma^{-1}\|\mathcal{E}\||\tilde{\tau}|_\infty$. Combining (2.40)–(2.42) the inequality (2.36) follows. □

In Lemma 2.9, the condition $s\le 2$ is automatically satisfied if there is at least one concave vertex.

3. A polygon with one concave vertex. In this paper, our goal is to obtain enough terms of a corner singularity expansion so that the remainder is in $\mathrm{H}^{2,q}$ for some number $q>2$ and close to 2. If the polygon is convex, Lemma 2.9 shows that the velocity and temperature are already in $\mathrm{H}^{2,q}$, so no corner singularity expansion is needed. Our discussion of the numbers $s_{1,n}$ and $\bar{s}_{1,n}$ suggests that for each concave vertex, two velocity singular functions and one temperature singular function must be subtracted from the solution in order to achieve a remainder with the desired regularity. In this section we discuss a polygon with exactly one concave vertex. We denote the concave vertex P_n and suppose without loss of generality that $P_n=(0,0)$. On the basis of Theorems 2.2 and 2.5 we split the solution of (1.7) into singular and regular parts and investigate its behavior in a neighborhood of the vertex P_n. The general polygon is considered in the next section.

In this section we refer to λ_1, λ_2, and λ_3 instead of $\lambda_{1,n}$, $\lambda_{2,n}$, and $\lambda_{3,n}$, respectively. Note that they satisfy

$$\frac{1}{2}<\lambda_1<\alpha<\lambda_2<1<\lambda_3<2\alpha.$$

We consider three steps to subtract the leading singular functions of the Lamé system and the Laplace equation from the exact solution $[\mathbf{u},p,\sigma]$.

Step 1. $\lambda_1 + 2/q < s < \alpha + 2/q$. In this step we split the first leading singular function of the Lamé system from the velocity $\mathbf{u}$ as follows:

$$\begin{cases} \mathbf{u} = C_1\Phi_1 + \mathbf{u}_{R,1}, \quad \Phi_1 = \chi\, r^{\lambda_1}\mathcal{T}_1(\theta), \\ p = p_{s,1} + p_{R,1}, \end{cases} \tag{3.1}$$

where $\chi = \chi_n$, $\mathcal{T}_1 = \mathcal{T}_{1,n}$, C_1 is a parameter to be constructed later, and $p_{s,1}$ will be constructed shortly.

We first define the pressure singular function $p_{s,1}$ corresponding to the velocity singular functions $C_1\Phi_1$:

$$\begin{cases} \kappa_1\mathbf{U}\cdot\nabla p_{s,1} = -C_1 \mathrm{div}\Phi_1 \ \text{ in } \Omega, \\ p_{s,1} = 0 \ \text{ on } \Gamma_{in}. \end{cases} \tag{3.2}$$

The function $p_{s,1}$ is given by $p_{s,1} = -C_1 B(\kappa_1^{-1}\mathrm{div}\Phi_1)$, which belongs to $\mathrm{H}^{s-1,q}(\Omega)$ for $s < \lambda_1 + 1 + 1/q$ by Lemma 2.8. With C_1 given and $\mathbf{u}_{R,1}$ defined by (3.1), $[\mathbf{u}_{R,1}, p, \sigma]$ is the solution of the problem

$$\begin{cases} -\mu\Delta\mathbf{u}_{R,1} - (\mu+\nu)\nabla\mathrm{div}\mathbf{u}_{R,1} + \rho\mathbf{U}\cdot\nabla\mathbf{u}_{R,1} + \nabla p = \mathbf{f} + \mathbf{f}_{s,1} \ \text{ in } \Omega, \\ \mathrm{div}\mathbf{u}_{R,1} + \kappa_1\mathbf{U}\cdot\nabla p + \kappa_2\mathbf{U}\cdot\nabla\sigma = g + g_{s,1} \ \text{ in } \Omega, \\ -\gamma\Delta\sigma + \tilde{\rho}\,\mathbf{U}\cdot\nabla\sigma + \tilde{\tau}\mathrm{div}\mathbf{u}_{R,1} = h + h_{s,1}, \\ \mathbf{u}_{R,1} = 0,\ \sigma = 0 \ \text{ on } \Gamma = \partial\Omega, \\ p = 0 \ \text{ on } \Gamma_{in}, \end{cases} \tag{3.3}$$

where

$$\begin{aligned} \mathbf{f}_{s,1} &= C_1(\mu\Delta\Phi_1 + (\mu+\nu)\nabla\mathrm{div}\Phi_1 - \rho\mathbf{U}\cdot\nabla\Phi_1), \\ g_{s,1} &= -C_1\mathrm{div}\Phi_1, \\ h_{s,1} &= -C_1\tilde{\tau}\mathrm{div}\Phi_1. \end{aligned} \tag{3.4}$$

We want to select C_1 so that $\mathbf{u}_{R,1} \in \mathbf{H}^{s,q}(\Omega)$ and $\sigma \in \mathrm{H}^{s,q}(\Omega)$. If $g \in \mathrm{H}^{s-1,q}(\Omega)$, then using Lemmas 2.7 and 2.8,

$$\begin{aligned} \|p\|_{s-1,q,\Omega} &\le \big\|B[\kappa_1^{-1}(g - \mathrm{div}\mathbf{u}_{R,1} - \kappa_2\mathbf{U}\cdot\nabla\sigma)]\big\|_{s-1,q,\Omega} + \|B\bar{g}_{s,1}\|_{s-1,q,\Omega} \\ &\le C\big(\|[\mathbf{u}_{R,1},\sigma]\|_{s,q,\Omega} + \|B\bar{g}\|_{s-1,q,\Omega}\big) + K_1|C_1|, \end{aligned}$$

where $\bar{g}_{s,1} = \kappa_1^{-1}g_{s,1}$, $\bar{g} = \kappa_1^{-1}g$, and $K_1 = C|\kappa_1^{-1}|_\infty\|B\|$.

We use the operators $\mathcal{M}$, B, and $\mathcal{E}$ to express the solution $[\mathbf{u}_{R,1}, p, \sigma]$ of (3.3) as follows:

$$\mathbf{u}_{R,1} = \mu^{-1}\mathcal{M}(\mathbf{f} + \mathbf{f}_{s,1} - \nabla p) \in \mathbf{H}^{s,q}(\Omega), \tag{3.5}$$

$$p = B[\kappa_1^{-1}(g + g_{s,1} - \mathrm{div}\mathbf{u}_{R,1} - \kappa_2\mathbf{U}\cdot\nabla\sigma)] \in \mathrm{H}^{s-1,q}(\Omega), \tag{3.6}$$

$$\sigma = \gamma^{-1}\mathcal{E}(h + h_{s,1} - \tilde{\tau}\mathrm{div}\mathbf{u}_{R,1}) \in \mathrm{H}^{s,q}(\Omega). \tag{3.7}$$

Considering the momentum equation in (3.3) and the point of view of Theorem 2.5, to achieve an increased regularity for $\mathbf{u}_{R,1}$ we must pick the unknown constant C_1 so that

$$\Lambda_{1,\nu_1}(\mathbf{f} + \mathbf{f}_{s,1} - \rho\,\mathbf{U}\cdot\nabla\mathbf{u}_{R,1} - \nabla p) = 0, \tag{3.8}$$

where $\Lambda_{1,\nu_1} := \Lambda_{1,n,\nu_1}$ with $\nu_1 = 1+\mu^{-1}\nu$ is the linear functional defined in Lemma 2.3. We will show that (3.8) gives a well-defined algebraic equation for the parameter C_1 and that this equation has a solution (see Lemma 3.4 below).

Applying the divergence operator to $\mathbf{u}_{R,1}$ given in the formula (3.5) and inserting it into the formula (3.6), the solution p is expressed in the form

$$(I-\mu^{-1}R)p=\frac{1}{\mu}B(\mu\,\bar{g}_{s,1}-\Pi_1\mathrm{div}\mathcal{M}\mathbf{f}_{s,1}-\mu\gamma^{-1}\bar{\mathbf{U}}\cdot\nabla\mathcal{E}h_{s,1}+G_1), \tag{3.9}$$

where $\Pi_1:=\kappa_1^{-1}-\gamma^{-1}\bar{\mathbf{U}}\cdot\nabla\mathcal{E}\tilde{\tau}$ and

$$\begin{aligned}R&:=B[\kappa_1^{-1}\mathrm{div}\,\mathcal{M}\nabla-\gamma^{-1}(\bar{\mathbf{U}}\cdot\nabla)\mathcal{E}\tilde{\tau}\mathrm{div}\mathcal{M}\nabla],\\ G_1&:=\mu\bar{g}-\Pi_1\mathrm{div}\mathcal{M}\mathbf{f}-\mu\gamma^{-1}\bar{\mathbf{U}}\cdot\nabla\mathcal{E}h.\end{aligned}$$

Then

$$p=\mu^{-1}B^*B(\mu\bar{g}_{s,1}-\Pi_1\mathrm{div}\mathcal{M}\mathbf{f}_{s,1}-\mu\gamma^{-1}\bar{\mathbf{U}}\cdot\nabla\mathcal{E}h_{s,1}+G_1), \tag{3.10}$$

where B^* is defined by

$$B^*=(I-\mu^{-1}\,R)^{-1}, \tag{3.11}$$

which exists and is bounded (to be shown in Lemma 3.1), and the solution formula p of (3.10) is well defined. On the other hand, substituting the function p of the formula (3.6) into the formula (3.5) and letting $S=\mathcal{M}\nabla B(\kappa_1^{-1}\mathrm{div}-\gamma^{-1}(\bar{\mathbf{U}}\cdot\nabla)\mathcal{E}\tilde{\tau}\mathrm{div})$, the function $\mathbf{u}_{R,1}$ is given by

$$\begin{aligned}&(I-\mu^{-1}\,S)\mathbf{u}_{R,1}\\ &=\frac{1}{\mu}\mathcal{M}\{\mathbf{f}_{s,1}-\nabla B[\bar{g}_{s,1}-\gamma^{-1}(\bar{\mathbf{U}}\cdot\nabla)\mathcal{E}h_{s,1}]+\mathbf{F}_1\},\end{aligned} \tag{3.12}$$

where $\mathbf{F}_1:=\mathbf{f}-\nabla B(\bar{g}-\gamma^{-1}(\bar{\mathbf{U}}\cdot\nabla)\mathcal{E}h)$. So

$$\mathbf{u}_{R,1}=\frac{1}{\mu}\mathcal{M}^*\mathcal{M}\{\mathbf{f}_{s,1}-\nabla B[\bar{g}_{s,1}-\gamma^{-1}(\bar{\mathbf{U}}\cdot\nabla)\mathcal{E}h_{s,1}]+\mathbf{F}_1\}, \tag{3.13}$$

where $\mathcal{M}^*=(I-\mu^{-1}\,S)^{-1}$, which exists and is bounded (to be shown in Lemma 3.1). Thus, the function $\mathbf{u}_{R,1}$ is well defined. Finally, using (3.7) and (3.13) the temperature function σ is expressed in the form

$$\sigma=\mu^{-1}[\Pi_3h_{s,1}-\Pi_2(\mathbf{f}_{s,1}-\nabla B\bar{g}_{s,1})+H_1], \tag{3.14}$$

where $H_1=\Pi_3h-\Pi_2(\mathbf{f}-\nabla Bg)$ and

$$\Pi_2=\gamma^{-1}\mathcal{E}\tilde{\tau}\mathrm{div}\mathcal{M}^*\mathcal{M}, \tag{3.15}$$

$$\Pi_3=\gamma^{-1}\mathcal{E}[\mu-\gamma^{-1}\tilde{\tau}\mathrm{div}\mathcal{M}^*\mathcal{M}\nabla B(\bar{\mathbf{U}}\cdot\nabla)\mathcal{E}]. \tag{3.16}$$

If the parameter C_1 is constructed, the solution $[\mathbf{u}_{R,1},p,\sigma]$ of (3.3) can be expressed.

LEMMA 3.1. *Let $1\le s<\lambda_1+2/q$. Let $\mathcal{M}$, B, and $\mathcal{E}$ be the solution operators defined in* (1.13)–(1.15)*, respectively.*

(a) *Then the following norms are bounded:*

$$\|\mathcal{M}\nabla B\nabla\cdot\|=\sup_{0\neq\mathbf{v}\in\mathbf{H}^{s,q}(\Omega)}\frac{\|\mathcal{M}\nabla B\nabla\cdot\mathbf{v}\|_{s,q,\Omega}}{\|\mathbf{v}\|_{s,q,\Omega}}<\infty, \tag{3.17}$$

$$\|B\nabla\cdot\mathcal{M}\nabla\|=\sup_{0\neq\chi\in\mathrm{Q}^{s,q}(\Omega)}\frac{\|B[\nabla\cdot(\mathcal{M}\nabla\chi)]\|_{\mathrm{Q}^{s,q}(\Omega)}}{\|\chi\|_{\mathrm{Q}^{s,q}(\Omega)}}<\infty. \tag{3.18}$$

(b) *If μ is large enough, then the operator $\mathcal{M}^*$ of* (3.13) *is a well-defined bounded operator on $\mathbf{H}^{s,q}(\Omega)$, and the operator B^* of* (3.11) *is a well-defined bounded operator on $\mathrm{Q}^{s,q}(\Omega)$. In addition, $\|\mathcal{M}^*\|$ and $\|B^*\|$ are bounded uniformly in μ for large μ. Furthermore, the following formulas give the solution of* (3.3)*:*

$$\mathbf{u}_{R,1} = \frac{1}{\mu}\mathcal{M}^*\mathcal{M}\left\{\mathbf{f}_{s,1} - \nabla B\left[\bar{g}_{s,1} - \frac{1}{\gamma}(\bar{\mathbf{U}}\cdot\nabla)\mathcal{E}h_{s,1}\right] + \mathbf{F}_1\right\}, \tag{3.19}$$

$$p = \frac{1}{\mu}B^*B\left\{\mu\bar{g}_{s,1} - \Pi_1\nabla\cdot\mathcal{M}\mathbf{f}_{s,1} - \frac{\mu}{\gamma}(\bar{\mathbf{U}}\cdot\nabla)\mathcal{E}h_{s,1} + G_1\right\}, \tag{3.20}$$

$$\sigma = \frac{1}{\mu}[\Pi_3 h_{s,1} - \Pi_2(\mathbf{f}_{s,1} - \nabla B g_{s,1}) + H_1], \tag{3.21}$$

where $\mathbf{F}_1$, G_1, and H_1 are defined in (3.12), (3.9), *and* (3.14), *respectively.*

Proof. First, (3.17)–(3.18) follow from the following two diagrams:

$$\mathbf{H}^{s,q} \xrightarrow{\mathrm{div}} \mathrm{H}^{s-1,q} \xrightarrow{B} \mathrm{Q}^{s,q} \xrightarrow{\nabla} \mathbf{H}^{s-2,q'} \xrightarrow{\mathcal{M}} \mathbf{H}^{s,q},$$
$$\mathrm{Q}^{s,q} \xrightarrow{\nabla} \mathbf{H}^{s-2,q'} \xrightarrow{\mathcal{M}} \mathbf{H}^{s,q} \xrightarrow{\mathrm{div}} \mathrm{H}^{s-1,q} \xrightarrow{B} \mathrm{Q}^{s,q}.$$

Next we claim that the operator B^* exists and is also bounded. For any function $g \in C_0^\infty(\Omega)$ we let $[\mathbf{u}, p, \sigma]$ solve

$$\begin{cases} -\mu\Delta\mathbf{u} - (\mu+\nu)\nabla\mathrm{div}\mathbf{u} + \rho\,\mathbf{U}\cdot\nabla\mathbf{u} + \nabla p = 0 \ \text{ in } \Omega, \\ \mathrm{div}\mathbf{u} + \kappa_1\,\mathbf{U}\cdot\nabla p + \kappa_2\,\mathbf{U}\cdot\nabla\sigma = \kappa_1\,\mathbf{U}\cdot\nabla g \ \text{ in } \Omega, \\ -\gamma\Delta\sigma + \tilde{\rho}\,\mathbf{U}\cdot\nabla\sigma + \tilde{\tau}\mathrm{div}\mathbf{u} = 0 \ \text{ in } \Omega, \\ \mathbf{u} = 0,\ \sigma = 0 \ \text{ on } \Gamma, \\ p = 0 \ \text{ on } \Gamma_{in}. \end{cases} \tag{3.22}$$

From Lemma 2.9 and the second equation of (3.22) one easily has

$$\|p\|_{Q^{s,q}(\Omega)} \le C\|g\|_{Q^{s,q}(\Omega)}. \tag{3.23}$$

From the first and third equations of (3.22) and using the operators $\mathcal{M}$ and $\mathcal{E}$, one has $\mathbf{u} = -\mu^{-1}\mathcal{M}\nabla p$ and $\sigma = -\gamma^{-1}\mathcal{E}\tilde{\tau}\mathrm{div}\mathbf{u}$. Considering this and the second equation of (3.22),

$$\begin{aligned} p - g &= -B(\kappa_1^{-1}\mathrm{div}\mathbf{u} + \bar{\mathbf{U}}\cdot\nabla\sigma) \\ &= \mu^{-1}B[\kappa_1^{-1}\mathrm{div}\mathcal{M}\nabla - \gamma^{-1}(\bar{\mathbf{U}}\cdot\nabla)\mathcal{E}\tilde{\tau}\mathrm{div}\mathcal{M}\nabla]p. \end{aligned} \tag{3.24}$$

From the definition of the operator R we have $p - \mu^{-1}Rp = g$, so $p = B^*g$. Since $C_0^\infty(\Omega)$ is dense in the space $\mathrm{Q}^{s,q}(\Omega)$, the operator B^* is a well-defined bounded operator on $\mathrm{Q}^{s,q}(\Omega)$ (by (3.23)). Similarly, it can be shown that the operator $\mathcal{M}^*$ exists and is bounded on $\mathrm{H}^{s,q}(\Omega)$. ∎

We now go back to the linear equation (3.8). Using (3.19)–(3.20) we have

$$\rho\mathbf{U}\cdot\nabla\mathbf{u}_{R,1} + \nabla p = \mu^{-1}(\mathbf{K}\mathbf{f}_{s,1} + \mathbf{L}\bar{g}_{s,1} + \mathbf{J}h_{s,1}) + \mu^{-1}\mathbf{Z}, \tag{3.25}$$

where

$$\begin{aligned} \mathbf{K} &= \rho\mathbf{U}\cdot\nabla\mathcal{M}^*\mathcal{M} - \nabla B^*B(\Pi_1\mathrm{div}\mathcal{M}), \\ \mathbf{L} &= \mu\nabla B^*B - \rho\mathbf{U}\cdot\nabla\mathcal{M}^*\mathcal{M}\nabla B, \\ \mathbf{J} &= \gamma^{-1}\rho\mathbf{U}\cdot\nabla\mathcal{M}^*\mathcal{M}\nabla B(\bar{\mathbf{U}}\cdot\nabla)\mathcal{E} - \mu\gamma^{-1}\nabla B^*B\bar{\mathbf{U}}\cdot\nabla\mathcal{E}, \\ \mathbf{Z} &= \rho\mathbf{U}\cdot\nabla\mathcal{M}^*\mathcal{M}\mathbf{F}_1 + \nabla B^*BG_1. \end{aligned} \tag{3.26}$$

So

$$\begin{aligned}&\mathbf{f}_{s,1}-\rho\mathbf{U}\cdot\nabla\mathbf{u}_{R,1}-\nabla p\\&=\mathbf{f}_{s,1}-\mu^{-1}(\mathbf{K}\mathbf{f}_{s,1}+\mathbf{L}g_{s,1}+\mathbf{J}h_{s,1})-\mu^{-1}\mathbf{Z}\\&=(\mathbf{I}-\mu^{-1}\mathbf{K})\mathbf{f}_{s,1}-\mu^{-1}(\mathbf{L}g_{s,1}+\mathbf{J}h_{s,1})-\mu^{-1}\mathbf{Z}.\end{aligned}\tag{3.27}$$

Using (3.4), (3.27), and (3.8), we obtain an algebraic equation for the unknown C_1:

$$X_1C_1=Y_1,\tag{3.28}$$

where

$$\begin{aligned}&X_1=\Lambda_{1,\nu_1}[(\mathbf{I}-\mu^{-1}\mathbf{K})\alpha_1]+\mu^{-1}\Lambda_{1,\nu_1}[\mathbf{L}(\mathrm{div}\Phi_1)+\mathbf{J}(\tilde{\tau}\mathrm{div}\Phi_1)],\\&Y_1=-\Lambda_{1,\nu_1}(\mathbf{f}-\mu^{-1}\mathbf{Z}),\\&\alpha_1=\mu\Delta\Phi_1+\nu\nabla\mathrm{div}\Phi_1-\rho\mathbf{U}\cdot\nabla\Phi_1.\end{aligned}\tag{3.29}$$

In order to show that the coefficient X_1 is well defined, it is enough to show the following lemma.

LEMMA 3.2. *Let $\Phi=\Phi_1$(or Φ_2 given in Step 3) be given above. Let $1\le s\le\min\{2,\lambda_1+1+2/q\}$. Then the values of the operators ∇B^*B (or ∇B), $\nabla B^*B(\bar{\mathbf{U}}\cdot\nabla)\mathcal{E}$, $(\mathbf{U}\cdot\nabla)\mathcal{M}\mathcal{M}^*\nabla B$ evaluated at the functions*

$$\begin{aligned}&\alpha_0:=\mu\Delta\Phi+(\mu+\nu)\nabla\nabla\cdot\Phi,\quad \Pi_1\nabla\cdot\mathcal{M}\alpha_0,\\&\Pi_1\nabla\cdot\mathcal{M}(\mathbf{U}\cdot\nabla)\Phi,\quad \nabla\cdot\Phi,\end{aligned}\tag{3.30}$$

belong to $\mathbf{H}^{s-2,q}(\Omega)$.

Proof. We have $\Phi=\chi\, r^\lambda\mathcal{T}(\theta)$, where $\lambda=\lambda_1$ (or λ_2) and $\mathcal{T}=\mathcal{T}_1$ (or $\mathcal{T}_2$). Hence we have $|\nabla\Phi(x,y)|\le Cr^{\lambda-1}$ for all $(x,y)\in\Omega$. From Lemma 2.8, $B\nabla(\kappa^{-1}\Phi)\in\mathrm{H}^{s-1,q}(\Omega)$ and $\nabla B\nabla(\kappa^{-1}\Phi)\in\mathrm{H}^{s-2,q}(\Omega)$. So we have $\mathcal{M}\nabla B(\kappa^{-1}\nabla\Phi)\in\mathbf{H}^{s-1,q}(\Omega)$. Since M^* and B^* are bounded operators by Lemma 3.1, the required property follows. Furthermore, the function $\nabla\, B^*B[\kappa^{-1}\nabla\,\mathcal{M}(\mu\Delta\Phi+(\mu+\nu)\nabla\mathrm{div}\Phi)]$ belongs to $\mathrm{H}^{s-1,q}(\Omega)$ because $\mu\Delta\Phi+(\mu+\nu)\nabla\mathrm{div}\Phi\equiv0$ near the origin. ∎

Using Lemma 3.2, we show that X_1 and Y_1 are well-defined coefficients.

LEMMA 3.3. (a) *The number $\nu_1:=1+\mu^{-1}\nu>0$, which is close to 1 for a large value μ and a fixed value ν.*

(b) *If μ is large enough, then $X_1\neq0$ and is finite.*

(c) *If $[\mathbf{f},h]\in\mathbf{H}^{s-2,q}\times\mathrm{H}^{s-2,q}$ and $g\in\mathrm{H}^{s-1,q}$, then Y_1 is finite and estimated by*

$$|Y_1|\le C\big(\|[\mathbf{f},h]\|_{s-2,q,\Omega}+\|B(\kappa^{-1}g)\|_{s-1,q,\Omega}\big).$$

Proof. (a) follows from $\mu>0$ and $\mu+\nu>0$. (b) The finiteness of X_1 follows from Lemma 3.2. Using Lemmas 2.4 and 3.1, it is seen that the operators $(\mathbf{U}\cdot\nabla)\mathcal{M}^*\mathcal{M}\nabla B$ and ∇B^*B are bounded maps from $\mathrm{H}^{s-1,q}(\Omega)$ to $\mathrm{H}^{s-2,q}(\Omega)$, and $(\mathbf{U}\cdot\nabla)\mathcal{M}^*\mathcal{M}$ is a bounded map from $\mathrm{H}^{s-2,q}(\Omega)$ to $\mathrm{H}^{s-2,q}(\Omega)$. To prove the finiteness of Y_1 it is enough to show that the mapping $\nabla B^*B(\kappa_1^{-1}\mathrm{div}\mathcal{M})$, which is one of the main difficult terms, is bounded from $\mathbf{H}^{s-2,q}(\Omega)$ to $\mathbf{H}^{s-2,q}(\Omega)$. For this, let $\mathbf{f}\in\mathbf{H}^{s-2,q}(\Omega)$. For $s<s_{1,n}=\lambda_1+2/q$, $\mathcal{M}\mathbf{f}\in\mathbf{H}^{s,q}(\Omega)$ with $\|\mathcal{M}\mathbf{f}\|_{s,q,\Omega}\le C\|\mathbf{f}\|_{s-2,q,\Omega}$. Then $\mathrm{div}\mathcal{M}\mathbf{f}\in\mathrm{H}^{s-1,q}(\Omega)$. So, from Lemma 2.7 $B(\kappa_1^{-1}\mathrm{div}\mathcal{M}\mathbf{f})\in\mathrm{H}^{s-1,q}(\Omega)$ with the corresponding norm inequality. For $\lambda_1+2/q<s<\alpha+2/q$, Theorem 2.5 gives $\mathcal{M}\mathbf{f}=\Lambda_{1,\nu_1}(\mathbf{f})\Phi_1+\mathbf{u}_{R,1}$ with $\mathbf{u}_{R,1}\in\mathbf{H}^{s,q}(\Omega)$ and $|\Lambda_{1,\nu_1}(\mathbf{f})|+\|\mathbf{u}_{R,1}\|_{s,q,\Omega}\le C\|\mathbf{f}\|_{s-2,q,\Omega}$. Then $B(\kappa_1^{-1}\mathrm{div}\mathcal{M}\mathbf{f})=\Lambda_{1,\nu_1}(\mathbf{f})B(\kappa_1^{-1}\mathrm{div}\Phi_1)+B(\kappa_1^{-1}\mathrm{div}\mathbf{u}_{R,1})$. Applying Lemmas 2.4 and 2.7, we obtain

$B(\kappa_1^{-1}\mathrm{div}\mathcal{M}\mathbf{f}) \in \mathrm{H}^{s-1,q}(\Omega)$ with the corresponding norm inequality. Using Lemma 3.1(b), $B^*B(\kappa_1^{-1}\mathrm{div}\mathcal{M}\mathbf{f}) \in \mathrm{H}^{s-1,q}(\Omega)$ and $\nabla B^*B(\Pi_1\mathrm{div}\mathcal{M}\mathbf{f}) \in \mathrm{H}^{s-2,q}(\Omega)$. In a similar way the other mappings can be handled. Thus the assertion follows.

Suppose μ is large enough. Then $X_1 \neq 0$. Indeed, note that $\|\mathcal{M}^*\|$ and $\|B^*\|$ are bounded uniformly in μ by Lemma 3.1(b) and $\|\mathcal{M}\| \le C\mu^{-1}$ for large μ by Lemma 2.4. Also note that B is independent of μ. So from (3.26), we see that $\|K\| \le C$ and $\mu^{-1}\|L\| \le C$. On the other hand, letting $\nu_1 = 1+\mu^{-1}\nu$, $\alpha_{\nu_1} := \Delta\Phi_1 + \nu_1\nabla\mathrm{div}\Phi_1$ and $\Phi_{1,\beta} := \rho\mathbf{U}\cdot\nabla\Phi_1$, we have

$$\begin{aligned} X_1 &= \mu\,\Lambda_{1,\nu_1}(\alpha_{\nu_1}) - \Lambda_{1,\nu_1}(K\alpha_{\nu_1}) + \Lambda_{1,\nu_1}[\Phi_{1,\beta} - \mu^{-1}K\Phi_{1,\beta}] \\ &\quad + \mu^{-1}\Lambda_{1,\nu_1}[\mathbf{L}(\mathrm{div}\Phi_1) + \mathbf{J}(\tilde{\tau}\mathrm{div}\Phi_1)] \\ &= O(\mu) + O(\mu^{-1}) + O(1), \quad \text{if } \Lambda_{1,\nu_1}(\alpha_{\nu_1}) \neq 0. \end{aligned}$$

From Lemma 2.3, we know that $\Lambda_{1,\nu_1}(\alpha_{\nu_1}) \neq 0$ for any number $\nu_1 > 0$ (note that $\nu_1 = 1/3$ for $\nu = -2\mu/3$). So $\Lambda_{1,\nu_1}(\alpha_{\nu_1}) \neq 0$ for any number $\nu_1 = 1+\mu^{-1}\nu > 0$. So if μ is large enough, $X_1 \neq 0$.

(c) Recall that $\mathbf{F}_1$ and G_1 are defined in (3.12) and (3.9), respectively. Since $\|\mathbf{Z}\|_{s-2,q,\Omega} \le C(\|[\mathbf{f},h]\|_{s-2,q,\Omega} + \|B\bar{g}\|_{s-1,q,\Omega})$, and using (3.30), we have

$$\begin{aligned} |Y_1| &\le C\|\mathbf{f} - \mu^{-1}\mathbf{Z}\|_{s-2,q,\Omega} \\ &\le C\big(\|[\mathbf{f},h]\|_{s-2,q,\Omega} + \|B\bar{g}\|_{s-1,q,\Omega}\big), \end{aligned}$$

where $C = C(\|\kappa_1^{-1}\|_\infty)$. Hence $|C_1| \le C(\|[\mathbf{f},h]\|_{s-2,q,\Omega} + \|B\bar{g}\|_{s-1,q,\Omega})$. □

We are now ready to determine the constant parameter C_1 in (3.1) so that the velocity $\mathbf{u}$ of (1.7) is split into singular and regular parts. Hence using the algebraic equation (3.28) and Lemma 3.3(c) we deduce the following lemma.

LEMMA 3.4. *Equation* (3.8) *holds if and only if the number* C_1 *in* (3.1) *is determined so that*

$$C_1 = Y_1/X_1. \tag{3.31}$$

With this choice,

$$|C_1| \le C\Big(\|[\mathbf{f},h]\|_{s-2,q,\Omega} + \|B\bar{g}\|_{s-1,q,\Omega}\Big). \tag{3.32}$$

Finally, with respect to Step 1 we establish a regularity result for the regular part $\mathbf{u}_{R,1}$ of the velocity function and the corresponding pressure and temperature functions.

THEOREM 3.5. *Let* $s \in (\lambda_1 + 2/q,\, \alpha + 2/q)$. *Let* $[\mathbf{f},g,h] \in \mathbf{H}^{s-2,q}(\Omega) \times \mathrm{H}^{s-1,q}(\Omega) \times \mathrm{H}^{s-2,q}(\Omega)$. *Suppose that* μ *and* γ *are sufficiently large. Let* $[\mathbf{u},p,\sigma]$ *be, in the sense of* (2.35), (2.32)–(2.33), *a weak solution of* (1.7). *Let* C_1 *be given by* (3.31). *Then the pair* $[\mathbf{u}_{R,1},p,\sigma]$ *is the solution of* (3.3). *Furthermore, if* $\mu_* := \mu - C\|\mathcal{M}\|\|B\|(1+\gamma^{-1}\|\mathcal{E}\|) > 0$, *then* $[\mathbf{u}_{R,1},p,\sigma] \in \mathbf{H}^{s,q}(\Omega) \times \mathrm{H}^{s-1,q}(\Omega) \times \mathrm{H}^{s,q}(\Omega)$ *and satisfies*

$$\|[\mu\mathbf{u}_{R,1},\sigma]\|_{s,q,\Omega} + \|p\|_{s-1,q,\Omega} \le K\big(\|[\mathbf{f},h]\|_{s-2,q,\Omega} + \|B\bar{g}\|_{s-1,q,\Omega}\big), \tag{3.33}$$

where $K = C(1+\mu_*^{-1})$.

Proof. It is obvious that the pair $[\mathbf{u}_{R,1},p,\sigma]$ given in (3.19)–(3.21) satisfies (3.3). Using Lemma 2.8, we have $\|B(\kappa_1^{-1}g_{s,1})\|_{s-1,q,\Omega} \le C|C_1|$ for $s < \lambda_{1,n} + 1 + 1/q$, and using Lemma 2.7, we have $\|B(\bar{\mathbf{U}}\cdot\nabla\sigma)\|_{s-1,q,\Omega} \le C\|B\|\|\sigma\|_{s,q,\Omega}$. Using (3.6),

$$\|p\|_{s-1,q,\Omega} \le C\|B\|\|[\mathbf{u}_{R,1},\sigma]\|_{s,q,\Omega} + \text{r.h.s.}, \tag{3.34}$$

where $C = C(\|\kappa_1^{-1}\|_{1,q,D})$ and r.h.s. $= C(\|[\mathbf{f}, h]\|_{s-2,q} + \|B\bar{g}\|_{s-1,q})$. Also note that the inequality $\|h_{s,1}\|_{s-2,q,\Omega} \le C|C_1|$ is true for $s < \lambda_{1,n} + 1 + 2/q$. Using (3.7) and (3.32),

$$\|\sigma\|_{s,q,\Omega} \le C\gamma^{-1}\|\mathcal{E}\|\|\mathbf{u}_{R,1}\|_{s-1,q,\Omega} + \text{r.h.s.} \tag{3.35}$$

for $C = C(|\tilde{\tau}|_\infty)$. Combining (3.34) and (3.35),

$$\|p\|_{s-1,q,\Omega} \le C\|B\|(1 + \gamma^{-1}\|\mathcal{E}\|)\|\mathbf{u}_{R,1}\|_{s,q,\Omega} + \text{r.h.s.}, \tag{3.36}$$

where $C = C(\|\kappa_1^{-1}\|_{1,q,\Omega}, |\tilde{\tau}|_\infty)$. Since $\|\mathbf{f}_{s,1}\|_{s-2,q,\Omega} \le C|C_1|$ for $s < \lambda_{1,n} + 1 + 2/q$, and using (3.5) and (3.32),

$$\|\mathbf{u}_{R,1}\|_{s,q,\Omega} \le \mu^{-1}\|\mathcal{M}\|\|p\|_{s-1,q} + \text{r.h.s.} \tag{3.37}$$

Combining (3.37) with (3.36) and using the condition $\mu_* > 0$,

$$\mu\|\mathbf{u}_{R,1}\|_{s,q,\Omega} \le C\mu_*^{-1}\text{r.h.s.} \tag{3.38}$$

Using (3.38), (3.35), and (3.36), we show (3.33). ∎

Step 2. $\alpha + 2/q < s < \lambda_2 + 2/q$. In this second step, we split the first leading singular function for the Laplace problem from the temperature σ of (1.7) as follows:

$$\begin{cases} \sigma = \sigma_s + \sigma_R, \quad \sigma_s = C_*\phi, \\ p = p_{s,1} + p_{s,2} + p_{R,2}, \end{cases} \tag{3.39}$$

where $\phi = \chi\, r^\alpha \sin[\alpha(\theta - \omega_1)]$ with $\chi = \chi_n$, $\alpha = \pi/\omega$, C_* will be determined later, and $p_{s,2}$ will be constructed shortly. As with the function $p_{s,1}$ of Step 1, let $p_{s,2} = -B(\bar{\mathbf{U}}\cdot\nabla\sigma_s) \in \mathrm{H}^{s-1,q}(\Omega)$ with $\bar{\mathbf{U}} = \kappa_1^{-1}\kappa_2\mathbf{U}$. With C_* given and $\sigma_R = \sigma - \sigma_s$ defined above, $[\mathbf{u}_{R,1}, p, \sigma_R]$ is the solution of the problem

$$\begin{cases} -\mu\Delta\mathbf{u}_{R,1} - (\mu+\nu)\nabla\mathrm{div}\mathbf{u}_{R,1} + \rho\,\mathbf{U}\cdot\nabla\mathbf{u}_{R,1} + \nabla p = \mathbf{f} + \mathbf{f}_{s,1} \ \text{ in } \Omega, \\ \quad \mathrm{div}\mathbf{u}_{R,1} + \kappa_1\mathbf{U}\cdot\nabla p + \kappa_2\mathbf{U}\cdot\nabla\sigma_R = g + g_{s,2} \ \text{ in } \Omega, \\ -\gamma\Delta\sigma_R + \tilde{\rho}\,\mathbf{U}\cdot\nabla\sigma_R + \tilde{\tau}\mathrm{div}\mathbf{u}_{R,1} = h + h_{s,2}, \\ \quad \mathbf{u}_{R,1} = 0,\ \sigma_R = 0 \ \text{ on } \Gamma = \partial\Omega, \\ \quad p = 0 \ \text{ on } \Gamma_{in}, \end{cases} \tag{3.40}$$

where $g_{s,2} = g_{s,1} - C_*\kappa_2\mathbf{U}\cdot\nabla\phi$ and $h_{s,2} = h_{s,1} + C_*(\gamma\Delta\phi - \tilde{\rho}\,\mathbf{U}\cdot\nabla\phi)$. Using the energy equation in (3.40) and Theorem 2.2 to obtain an increased regularity for σ_R, we need to pick the parameter C_* so that

$$\Lambda(h + h_{s,2} - \tilde{\rho}\,\mathbf{U}\cdot\nabla\sigma_R - \tilde{\tau}\mathrm{div}\mathbf{u}_{R,1}) = 0, \tag{3.41}$$

where $\Lambda := \Lambda_{1,n}$ is defined in Lemma 2.1. As in Step 1, using (3.41), we derive a well-defined algebraic equation for C_* and show that it is solvable. As in (3.19)–(3.21) one can derive the solution formula for (3.40) as follows:

$$\mathbf{u}_{R,1} = \frac{1}{\mu}\mathcal{M}^*\mathcal{M}\left\{\mathbf{f}_{s,1} - \nabla B\left[\bar{g}_{s,2} - \frac{1}{\gamma}(\bar{\mathbf{U}}\cdot\nabla)\mathcal{E}h_{s,2}\right] + \mathbf{F}_1\right\}, \tag{3.42}$$

$$p = \frac{1}{\mu}B^*B\left\{\mu\bar{g}_{s,2} - \Pi_1\nabla\cdot\mathcal{M}\mathbf{f}_{s,1} - \frac{\mu}{\gamma}(\bar{\mathbf{U}}\cdot\nabla)\mathcal{E}h_{s,2} + G_1\right\}, \tag{3.43}$$

$$\sigma_R = \frac{1}{\mu}[\Pi_3 h_{s,2} - \Pi_2(\mathbf{f}_{s,1} - \nabla B g_{s,2}) + H_1], \tag{3.44}$$

where $\bar{g}_{s,2} = \kappa_1^{-1} g_{s,2}$ and F_1, G_1, H_1 are defined in (3.12), (3.9), and (3.14), respectively. Furthermore,

$$\begin{aligned}
&\tilde{\rho}\,\mathbf{U}\cdot\nabla\sigma_R + \tilde{\tau}\mathrm{div}\mathbf{u}_{R,1}\\
&\quad = \gamma^{-1}\tilde{\rho}\,\mathbf{U}\cdot\nabla\mathcal{E}(h+h_{s,2}) - (\tilde{\rho}\gamma^{-1}\mathbf{U}\cdot\nabla\mathcal{E} - I)\tilde{\tau}\mathrm{div}\mathbf{u}_{R,1}\\
&\quad = \mu^{-1}[\bar{K}\mathbf{f}_{s,1} + \bar{L}g_{s,2} + \bar{J}h_{s,2}] + \mu^{-1}\bar{Z},
\end{aligned} \tag{3.45}$$

where

$$\begin{aligned}
\bar{K} &= (I - \gamma^{-1}\tilde{\rho}\,\mathbf{U}\cdot\nabla\mathcal{E})\tilde{\tau}\mathrm{div}\mathcal{M}^*\mathcal{M},\\
\bar{L} &= (\gamma^{-1}\tilde{\rho}\,\mathbf{U}\cdot\nabla\mathcal{E} - I)\tilde{\tau}\mathrm{div}\mathcal{M}^*\mathcal{M}\nabla B\kappa_1^{-1},\\
\bar{J} &= \mu\gamma^{-1}\tilde{\rho}\,\mathbf{U}\cdot\nabla\mathcal{E} - \gamma^{-1}(\gamma^{-1}\tilde{\rho}\,\mathbf{U}\cdot\nabla\mathcal{E} - I)\tilde{\tau}\mathrm{div}\mathcal{M}^*\mathcal{M}\nabla B(\bar{\mathbf{U}}\cdot\nabla)\mathcal{E},\\
\bar{Z} &= \gamma^{-1}\tilde{\rho}\,\mathbf{U}\cdot\nabla\mathcal{E}h - \mu^{-1}(\gamma^{-1}\tilde{\rho}\,\mathbf{U}\cdot\nabla\mathcal{E} - I)\tilde{\tau}\mathrm{div}\mathcal{M}^*\mathcal{M}\mathbf{F}_1.
\end{aligned}$$

So

$$\begin{aligned}
&h_{s,2} - \tilde{\rho}\,\mathbf{U}\cdot\nabla\sigma_R - \tilde{\tau}\mathrm{div}\mathbf{u}_{R,1}\\
&= h_{s,2} - \mu^{-1}[\bar{K}\mathbf{f}_{s,1} + \bar{L}g_{s,2} + \bar{J}h_{s,2}] - \mu^{-1}\bar{Z}\\
&= -\mu^{-1}[\bar{K}\mathbf{f}_{s,1} + \bar{L}g_{s,2}] + (I - \mu^{-1}\bar{J})h_{s,2} - \mu^{-1}\bar{Z}\\
&= C_*\Big\{\mu^{-1}[\bar{L}(\kappa_2\mathbf{U}\cdot\nabla\phi)] + (I - \mu^{-1}\bar{J})(\gamma\Delta\phi - \tilde{\rho}\,\mathbf{U}\cdot\nabla\phi)\Big\}\\
&\quad - \mu^{-1}[\bar{K}\mathbf{f}_{s,1} + \bar{L}g_{s,1}] + (I - \mu^{-1}\bar{J})h_{s,1} - \mu^{-1}\bar{Z}.
\end{aligned} \tag{3.46}$$

Using (3.41) the algebraic equation for the parameter C_* is given by

$$X_*C_* = Y_*, \tag{3.47}$$

where

$$X_* = \Lambda\Big[(I - \mu^{-1}\bar{J})\alpha_* + \mu^{-1}[\bar{L}(\kappa_2\mathbf{U}\cdot\nabla\phi)]\Big], \tag{3.48}$$

$$Y_* = \Lambda\Big[\mu^{-1}(\bar{K}\mathbf{f}_{s,1} - \bar{L}g_{s,1}) - (I - \mu^{-1}\bar{J})h_{s,1} + \mu^{-1}\bar{Z} - h\Big], \tag{3.49}$$

$$\alpha_* := \gamma\Delta\phi - \tilde{\rho}\,\mathbf{U}\cdot\nabla\phi. \tag{3.50}$$

We next show that the coefficients X_* and Y_* are well defined and C_* is also determined.

LEMMA 3.6. (a) *If μ is large enough, then $X_* \neq 0$ and is finite.*

(b) *If $[\mathbf{f}, h] \in \mathbf{H}^{s-2,q} \times \mathrm{H}^{s-2,q}$ and $g \in \mathrm{H}^{s-1,q}$, then Y_* is finite and estimated by*

$$|Y_*| \le C\big(\|[\mathbf{f}, h]\|_{s-2,q,\Omega} + \|B(\kappa_1^{-1}g)\|_{s-1,q,\Omega}\big).$$

(c) *The constant is given by*

$$C_* = Y_*/X_* \tag{3.51}$$

and satisfies

$$|C_*| \le C\Big(\|[\mathbf{f}, h]\|_{s-2,q,\Omega} + \|B(\kappa_1^{-1}g)\|_{s-1,q,\Omega}\Big), \tag{3.52}$$

where $C = C(X_*)$.

Proof. The proof is similar to that of Lemmas 3.3 and 3.4. ∎

As in Step 1, with respect to Step 2 a regularity result can be given for the regular part $[\mathbf{u}_{R,1}, \sigma_{R,1}]$ of the velocity and temperature functions and the corresponding pressure function p.

THEOREM 3.7. *Let $s \in (\alpha+2/q,\, \lambda_2+2/q)$. Let $[\mathbf{f}, g, h] \in \mathbf{H}^{s-2,q}(\Omega)\times \mathrm{H}^{s-1,q}(\Omega)\times \mathrm{H}^{s-2,q}(\Omega)$. Suppose that μ and γ are sufficiently large. Let $[\mathbf{u}, p, \sigma]$ be, in the sense of* (2.35), (2.32)–(2.33), *a weak solution of* (1.7). *Let C_* be given in Lemma* 3.6. *Then the pair $[\mathbf{u}_{R,1}, p, \sigma_R]$ given in* (3.42)–(3.44) *is the solution of* (3.40) *and if the number μ_* given in Theorem* 3.5 *is positive, then $[\mathbf{u}_{R,1}, p, \sigma_R] \in \mathbf{H}^{s,q}(\Omega)\times \mathrm{H}^{s-1,q}(\Omega)\times \mathrm{H}^{s,q}(\Omega)$, satisfying*

$$\|[\mu\,\mathbf{u}_{R,1},\, \sigma_R]\|_{s,q,\Omega} + \|p\|_{s-1,q,\Omega} \le K\big(\|[\mathbf{f}, h]\|_{s-2,q,\Omega} + \|B\bar{g}\|_{s-1,q,\Omega}\big), \tag{3.53}$$

where $K = C(1+\mu_*^{-1})$.

Proof. The proof is similar to the proof of Theorem 3.5. □

Step 3. $\lambda_2 + 2/q < s \le 2$. In the third step we split the second leading singular function of the Lamé system from the solution $\mathbf{u}_{R,1}$ of (3.3) as follows:

$$\begin{cases} \mathbf{u} = C_1\Phi_1 + C_2\Phi_2 + \mathbf{u}_R, \quad \Phi_2 = \chi\, r^{\lambda_2}\mathcal{T}_2(\theta), \\ p = p_{s,1} + p_{s,2} + p_{s,3} + p_R, \end{cases} \tag{3.54}$$

where $\lambda_2 = \lambda_{2,n}$, $\mathcal{T}_2 = \mathcal{T}_{2,n}$, C_1 is given by (3.31), C_2 is the unknown parameter, and $p_{s,3}$ is constructed shortly. Likewise, the function $p_{s,3}$ is given by $p_{s,3} = -C_2 B(\kappa_1^{-1}\mathrm{div}\Phi_2)$, which is in $\mathrm{H}^{s-1,q}(\Omega)$ for $s < \lambda_3 + 1 + 1/q$. Then $[\mathbf{u}_R, p, \sigma_R]$ is the solution of the problem

$$\begin{cases} -\mu\Delta\mathbf{u}_R - (\mu+\nu)\nabla\mathrm{div}\mathbf{u}_R + \rho\,\mathbf{U}\cdot\nabla\mathbf{u}_R + \nabla p = \mathbf{f} + \mathbf{f}_s \ \text{ in } \Omega, \\ \mathrm{div}\mathbf{u}_R + \kappa_1\,\mathbf{U}\cdot\nabla p + \kappa_2\,\mathbf{U}\cdot\nabla\sigma_R = g + g_s \ \text{ in } \Omega, \\ -\gamma\Delta\sigma_R + \tilde{\rho}\,\mathbf{U}\cdot\nabla\sigma_R + \tilde{\tau}\mathrm{div}\mathbf{u}_R = h + h_s \ \text{ in } \Omega, \\ \mathbf{u}_R = 0,\ \sigma_R = 0 \ \text{ on } \Gamma = \partial\Omega, \\ p = 0 \ \text{ on } \Gamma_{in}, \end{cases} \tag{3.55}$$

where

$$\begin{aligned} \mathbf{f}_s &= \mathbf{f}_{s,1} + C_2(\mu\Delta\Phi_2 + (\mu+\nu)\nabla\mathrm{div}\Phi_2 - \rho\mathbf{U}\cdot\nabla\Phi_2), \\ g_s &= g_{s,2} - C_2\mathrm{div}\Phi_2, \\ h_s &= h_{s,2} - C_2\tilde{\tau}\mathrm{div}\Phi_2. \end{aligned} \tag{3.56}$$

As with (2.19)–(2.21) the solution $[\mathbf{u}_R, p, \sigma_R]$ of (3.55) is given by

$$\begin{aligned} \mathbf{u}_R &= \mu^{-1}\,\mathcal{M}^*\mathcal{M}\{\mathbf{f}_s - \nabla B[\bar{g}_s - \gamma^{-1}(\bar{\mathbf{U}}\cdot\nabla)\mathcal{E}h_s] + \mathbf{F}_1\}, \\ p &= \mu^{-1}B^*B[\mu\bar{g}_s - \Pi_1\nabla\cdot\mathcal{M}\mathbf{f}_s - \mu\gamma^{-1}(\bar{\mathbf{U}}\cdot\nabla)\mathcal{E}h_s + G_1], \\ \sigma_R &= \mu^{-1}[\Pi_3 h_s - \Pi_2(\mathbf{f}_s - \nabla B g_s) + H_1], \end{aligned} \tag{3.57}$$

where $\mathbf{F}_1$, G_1, and H_1 are defined in (3.12), (3.9), and (3.14), respectively.

It would seem natural in Step 3 to use the inequality $s < \lambda_3 + 2/q$ instead of the inequality $s \le 2$, because $s_{3,n} = \lambda_3 + 2/q$ is the value of the third velocity regularity index. If $q > 2$ is sufficiently close to 2, then $s_{3,n} > 2$. However, we have only established the boundedness of B on $\mathrm{H}^{t,q}(\Omega)$ for $t \le 1$, so we cannot establish the regularity of $\mathbf{u}_R$ beyond $\mathbf{H}^{2,q}(\Omega)$.

Considering the momentum equation in (3.55) and Theorem 2.5, to achieve an increased regularity for $\mathbf{u}_R$ we must pick the parameter C_2 so that

$$\Lambda_{2,\nu_1}(\mathbf{f} + \mathbf{f}_s - \rho\mathbf{U}\cdot\nabla\mathbf{u}_R - \nabla p) = 0, \tag{3.58}$$

where $\Lambda_{2,\nu_1} := \Lambda_{2,n,\nu_1}$ with $\nu_1 = 1 + \mu^{-1}\nu$. Using (3.57),

$$\rho\mathbf{U}\cdot\nabla\mathbf{u}_R + \nabla p = \mu^{-1}(\mathbf{K}\mathbf{f}_s + \mathbf{L}\bar{g}_s + \mathbf{J}h_s) + \mu^{-1}\mathbf{Z}, \tag{3.59}$$

where $\mathbf{K}$, $\mathbf{L}$, $\mathbf{J}$, and $\mathbf{Z}$ are given in (3.26). So

$$\begin{aligned}
&\mathbf{f}_s - \rho\mathbf{U}\cdot\nabla\mathbf{u}_R - \nabla p\\
&= \mathbf{f}_s - \mu^{-1}(\mathbf{K}\mathbf{f}_s + \mathbf{L}g_s + \mathbf{J}h_s) - \mu^{-1}\mathbf{Z}\\
&= (\mathbf{I} - \mu^{-1}\mathbf{K})\mathbf{f}_s - \mu^{-1}(\mathbf{L}g_s + \mathbf{J}h_s) - \mu^{-1}\mathbf{Z}\\
&= C_2\big\{(\mathbf{I} - \mu^{-1}\mathbf{K})\alpha_2 + \mu^{-1}[\mathbf{L}(\mathrm{div}\Phi_2) + \mathbf{J}(\tilde{\tau}\mathrm{div}\Phi_2)]\big\}\\
&\quad + (\mathbf{I} - \mu^{-1}\mathbf{K})\mathbf{f}_{s,1} - \mu^{-1}(\mathbf{L}g_{s,2} + \mathbf{J}h_{s,2}) - \mu^{-1}\mathbf{Z}.
\end{aligned} \tag{3.60}$$

From (3.58) and (3.60) the algebraic equation for the parameter C_2 is

$$X_2C_2 = Y_2, \tag{3.61}$$

where $\alpha_2 = \mu\Delta\Phi_2 + (\mu+\nu)\nabla\mathrm{div}\Phi_2 - \rho\mathbf{U}\cdot\nabla\Phi_2$ and

$$\begin{aligned}
X_2 &= \Lambda_{2,\nu_1}[(\mathbf{I} - \mu^{-1}\mathbf{K})\alpha_2] + \mu^{-1}\Lambda_{2,\nu_1}[\mathbf{L}(\mathrm{div}\Phi_2) + \mathbf{J}(\tilde{\tau}\mathrm{div}\Phi_2)],\\
Y_2 &= \Lambda_{2,\nu_1}[-(\mathbf{I} - \mu^{-1}\mathbf{K})\mathbf{f}_{s,1} + \mu^{-1}(\mathbf{L}g_{s,2} + \mathbf{J}h_{s,2}) + \mu^{-1}\mathbf{Z}]\\
&\quad - \Lambda_{2,\nu_1}(\mathbf{f}).
\end{aligned}$$

The next lemma shows that X_2 and Y_2 are well-defined coefficients and C_2 is also determined.

LEMMA 3.8. (a) *If μ is large enough, then $X_2 \neq 0$ and is finite.*

(b) *If $[\mathbf{f}, h] \in \mathbf{H}^{s-2,q} \times \mathrm{H}^{s-2,q}$ and $g \in \mathrm{H}^{s-1,q}$, then Y_2 is finite and estimated by $C(\|[\mathbf{f}, h]\|_{s-2,q,\Omega} + \|B(\kappa_1^{-1}g)\|_{s-1,q,\Omega})$.*

(c) *The unknown number C_2 is determined by*

$$C_2 = Y_2/X_2, \tag{3.62}$$

satisfying

$$|C_2| \le C\Big(\|[\mathbf{f}, h]\|_{s-2,q,\Omega} + \|B(\kappa_1^{-1}g)\|_{s-1,q,\Omega}\Big), \tag{3.63}$$

where $C = C(X_2)$.

Proof. The proof is similar to that of Lemmas 3.3 and 3.4. □

As in Step 2, using the same procedures given in the proof of Theorem 3.5, one can obtain a regularity result for the regular part $[\mathbf{u}_R, \sigma_R]$ of the velocity function, the temperature function, and the corresponding pressure function.

THEOREM 3.9. *Let $s \in (\lambda_2 + 2/q, 2]$. Let $[\mathbf{f}, g, h] \in \mathbf{H}^{s-2,q}(\Omega) \times \mathrm{H}^{s-1,q}(\Omega) \times \mathrm{H}^{s-2,q}(\Omega)$. Suppose that μ and γ are sufficiently large. Let $[\mathbf{u}, p, \sigma]$ be, in a sense of* (2.35), (2.32)–(2.33), *a weak solution of* (1.7). *Let C_2 be given in* (3.62). *Then the function $[\mathbf{u}_R, p, \sigma_R]$ given in* (3.57) *is the solution of* (3.55). *Moreover, if the number μ_* given in Theorem* 3.5 *is positive, then $[\mathbf{u}_R, p, \sigma_R] \in \mathbf{H}^{s,q}(\Omega) \times \mathrm{H}^{s-1,q}(\Omega) \times \mathrm{H}^{s,q}(\Omega)$ with*

$$\|[\mu\mathbf{u}_R,\, \sigma_R]\|_{s,q,\Omega} + \|p\|_{s-1,q,\Omega} \le K\big(\|[\mathbf{f}, h]\|_{s-2,q,\Omega} + \|B\bar{g}\|_{s-1,q,\Omega}\big), \tag{3.64}$$

where $K = C(1 + \mu_*^{-1})$.

Proof. The same procedures as used in the proof of Theorem 3.5 can be applied. ∎

Combining Theorems 3.5, 3.7, and 3.9 we obtain the following result.

THEOREM 3.10. *Let Ω be a polygon with one concave vertex. Let the vector field $\mathbf{U}$ satisfy Assumption A. Suppose the concave vertex is placed at the origin and has interior angle ω. Let $q > 2$ be sufficiently close to 2. Let $s \in (\lambda_2 + 2/q, 2]$. Let $[\mathbf{f}, g, h] \in \mathbf{H}^{s-2,q}(\Omega) \times \mathrm{H}^{s-1,q}(\Omega) \times \mathrm{H}^{s-2,q}(\Omega)$. Suppose that μ and γ are sufficiently large. Let $[\mathbf{u}, p, \sigma]$ be the solution of (1.7) in the sense of (2.35), (2.32)–(2.33). Then the pair $[\mathbf{u}, \sigma]$ can be split into the singular and regular parts, $[\mathbf{u}_s, \sigma_s]$ and $[\mathbf{u}_R, \sigma_R]$ near the origin:*

$$\begin{aligned} &\mathbf{u} = \mathbf{u}_s + \mathbf{u}_R, \quad \mathbf{u}_s = C_1\Phi_1 + C_2\Phi_2, \quad \mathbf{u}_R := \mathbf{u} - \mathbf{u}_s, \\ &\sigma = \sigma_s + \sigma_R, \quad \sigma_s = C_3\phi, \quad \sigma_R := \sigma - \sigma_s, \end{aligned}$$

and $[\mathbf{u}_R, p, \sigma_R] \in \mathbf{H}^{s,q}(\Omega) \times \mathrm{H}^{s-1,q}(\Omega) \times \mathrm{H}^{s,q}(\Omega)$, where $\mathbf{C} = [C_1, C_2, C_3]$ has been constructed in Steps 1, 2, and 3, with $C_3 = C_*$. In addition, there is a constant $C = C(\Omega, C_0, \|\mathbf{U}_R\|_{2,q,\Omega} + |\mathbf{d}|)$ with a given constant vector $\mathbf{d} = (d_1, d_2, d_3)$ such that

$$\begin{aligned} &|\mathbf{C}| + \|[\mu\mathbf{u}_R, \sigma_R]\|_{s,q,\Omega} + \|[\mathbf{u}, \sigma]\|_{s-1,q,\Omega} + \|p\|_{s-1,q,\Omega} \\ &\qquad \le C\big(\|[\mathbf{f}, h]\|_{s-2,q,\Omega} + \|B(\kappa^{-1}g)\|_{s-1,q,\Omega}\big). \end{aligned} \tag{3.65}$$

Proof. The proof follows from Theorems 3.5, 3.7, and 3.9. ∎

4. Bounded polygon ($q < \frac{1}{2}q_1^{\lambda}$). In this section we discuss the system (1.7) on any bounded polygon D. The vector field $\mathbf{U}$ is assumed to satisfy Assumption A. In contrast to the situation in section 3, there may be many vertices P_n such that $q_2(\lambda_{1,n}) \ge q_2^{\lambda}$. A localization and a partition of unity enable us to apply the results of sections 2 and 3. Because of the quasi-hyperbolic character of the system (1.7), the partition of unity must be constructed using the streamlines of $\mathbf{U}$. We recall that the streamlines are given by curves $(x, k(x, \bar{y}))$, where the function k satisfies (2.18).

For each vertex $P_n = (x_n, y_n)$, $n = 1, \dots, N$, we construct an open set $W_n \subset D$ and a function $\chi_n \in C_0^{\infty}(\mathbf{R}^2)$ satisfying the following conditions:

(i) the vertex P_n lies on ∂W_n and is the only vertex on ∂W_n.
(ii) $\bigcup_{n=1}^{N} W_n = D$.
(iii) $0 \le \chi_n \le 1$ on W_n, $\chi_n \equiv 0$ on $D \backslash \bar{W}_n$, $\sum_{n=1}^{N} \chi_n \equiv 1$ on D, $\mathbf{U} \cdot \nabla\chi_n \equiv 0$ on D.

We now construct the sets W_n and functions χ_n. Suppose the vertices are numbered so that $y_1 < y_2 < \cdots < y_N$. (The polygon can be rotated a little to achieve this.) For each $n = 1, \dots, N$, define a number $\bar{y}_n$ as follows. If $P_n \in \partial D_{in}$, let $\bar{y}_n = y_n$. If $P_n \notin \partial D_{in}$, let $\bar{y}_n = \xi(x_n, y_n)$. Thus, P_n lies on the streamline emanating from the point $(\delta(\bar{y}_n), \bar{y}_n)$. Define $d_n := |\bar{y}_n - \bar{y}_{n+1}|$. For each $n = 1, \dots, N$ let $y_{n*} = \bar{y}_n - \frac{3}{4}d_{n-1}$ and $y_n^* = \bar{y}_n + \frac{3}{4}d_n$. Then the open sets are constructed as follows: $W_1 := \{(x, y) \in D : y = k(x, \bar{y}),\ \bar{y} < y_1^*\}$ and $W_N := \{(x, y) \in D : y = k(x, \bar{y}),\ \bar{y} > y_{N*}\}$ and for $n = 2, \dots, N-1$,

$$W_n := \{(x, y) \in D : y = k(x, \bar{y}),\ y_{n*} < \bar{y} < y_n^*\}. \tag{4.1}$$

We see that $P_n \in \partial W_n$ and $D = \bigcup_{n=1}^{N} W_n$.

To define the functions χ_n, we start with the following two functions: $\alpha_+(y) = \exp(-y^{-2})$ if $y > 0$, $\alpha_+(y) = 0$ if $y \le 0$; and $\alpha_-(y) = 0$ if $y \ge 0$, $\alpha_-(y) = \exp(-y^{-2})$

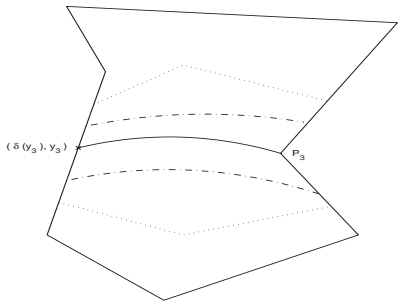


FIG. 1. *The polygon D with the region W_3 (dashed curves) and the polygon Ω_3 (dotted lines).*

if $y < 0$. Using these functions, we define the functions χ_i as follows: $\tilde{\chi}_1(x,y) := \alpha_-(y - k(x, y_1^*))$, for $n = 2, \ldots, N-1$,

$$\tilde{\chi}_n(x,y) := \begin{cases} \alpha_+(y - k(x, y_{n*})), & y > k(x, y_{n*}), \\ \alpha_-(y - k(x, y_n^*)), & y < k(x, y_n^*), \end{cases} \tag{4.2}$$

and $\tilde{\chi}_N(x,y) := \alpha_+(y - k(x, y_{N*}))$. From this construction, we see that $\mathbf{U}\cdot\nabla\tilde{\chi}_n(x,y) = 0$ for all n and $\sum_{n=1}^N \tilde{\chi}_n(x,y) \neq 0$ on D. Define $\chi_n(x,y) = \tilde{\chi}_n(x,y)/\sum_{n=1}^N \tilde{\chi}_i(x,y)$. Then the required properties for χ_n follow.

The open sets W_n are not, in general, polygonal domains. In order to apply the results in sections 3 and 4, we pick a bounded polygon Ω_n containing W_n, obtained by extending the sides of $\partial D \cap \partial W_n$ suitably, joining their end points with a broken straight line lying outside of W_n and so that each streamline by $\mathbf{U}$ intersects the boundary of Ω_n at only two points such that at most one of these points can be a vertex. Furthermore, the broken polygonal lines are constructed to have interior angles that are $< \pi$. In this way, either the polygon Ω_n is convex or it has exactly one concave vertex, the vertex P_n. The function χ_n vanishes in a neighborhood of the part of $\partial\Omega_n$ that does not lie on ∂D. Figure 1 illustrates this construction.

Let $[\mathbf{f}, g, h] \in \mathbf{H}^{s-2,q}(D) \times \mathrm{H}^{s-1,q}(D) \times \mathrm{H}^{s-2,q}(D)$, and let $[\mathbf{u}, p, \sigma] \in \mathbf{H}^{s,q}(D) \times \mathrm{H}^{s-1,q}(D) \times \mathrm{H}^{s,q}(D)$ $(s < \min\{s_{1,n}\})$ be the solution of (1.7) given by Lemma 2.9. Using the functions χ_n, set $\mathbf{u}_n = \chi_n\mathbf{u}$, $p_n = \chi_n p$, and $\sigma_n = \chi_n\sigma$. Then $\mathbf{u}_n = 0$ on $\partial\Omega_n$, $p_n = 0$ on $\partial\Omega_{n,in}$, and $\sigma_n = 0$ on $\partial\Omega_n$. So $[\mathbf{u}_n, p_n, \sigma_n]$ satisfies the problem (1.7) on the polygon Ω_n, with the right-hand sides $\mathbf{f}_n$ and g_n given by

$$\begin{aligned} \mathbf{f}_n &:= \chi_n\,\mathbf{f} - 2\,\mu\nabla\chi_n\cdot\nabla\mathbf{u} - (\mu+\nu)(\nabla\chi_n\,\mathrm{div}\mathbf{u} + \nabla\chi_n\nabla\mathbf{u}) \\ &\quad + \mathbf{u}(-\mu\,\Delta\chi_n - (\mu+\nu)\nabla\mathrm{div}\chi_n) + p\nabla\chi_n, \\ g_n &:= \chi_n\,g + \mathbf{u}\cdot\nabla\chi_n, \\ h_n &:= \chi_n h - 2\,\gamma\nabla\chi_n\cdot\nabla\sigma + \tilde{\rho}\,\sigma\mathbf{U}\cdot\nabla\chi_n + \tilde{\tau}\nabla\chi_n\cdot\mathbf{u}. \end{aligned}$$

Note that the trouble term $(\kappa_1 p + \kappa_2\sigma)\mathbf{U}\cdot\nabla\chi_n$ is not in the function g_n because $\mathbf{U}\cdot\nabla\chi_n = 0$. Using this fact and Lemma 2.9 we have the following inequality:

$$\begin{aligned} &\|[\mathbf{f}_n, h_n]\|_{s-2,q,D} + \|B(\kappa^{-1}g_n)\|_{s-1,q,D} \\ &\qquad \le C\big(\|[\mathbf{f}, h]\|_{s-2,q,D} + \|B(\kappa^{-1}g)\|_{s-1,q,D}\big). \end{aligned} \tag{4.3}$$

If P_n is a vertex with $q_2(\lambda_{1,n}) < q < \frac{1}{2}q_1(\lambda_{1,n})$, we apply Theorem 3.10. (This is possible because, by our construction, P_n is the only vertex of Ω_n with $q_2(\lambda_{1,n}) <$

$q < \frac{1}{2}q_1(\lambda_{1,n})$.) We write

$$\mathbf{u}_n = \mathbf{u}_{n,s} + \mathbf{u}_{n,R}, \quad \mathbf{u}_{n,s} = C_{1,n}\Phi_{1,n} + C_{2,n}\Phi_{2,n},$$
$$\sigma_n = \sigma_{n,s} + \sigma_{n,R}, \quad \sigma_{n,s} = C_{3,n}\phi_n,$$

where $\Phi_{i,n}$ $(i = 1, 2)$ and ϕ_n are the singular functions. Using (4.3) we have the inequality

$$|\mathbf{C}_n| + \|[\mathbf{u}_{n,R},\, \sigma_{n,R}]\|_{s,q,\Omega_n} + \|p_n\|_{s-1,q,\Omega_n} \le C\big(\|[\mathbf{f}, h]\|_{s-2,q,D} + \|B(\kappa^{-1}g)\|_{s-1,q,D}\big). \tag{4.4}$$

If P_n is a vertex with $q < q_2(\lambda_{1,n})$, we apply Lemma 2.9 to obtain

$$\|[\mathbf{u}_n, \sigma_n]\|_{s,q,\Omega_n} + \|p_n\|_{s-1,q,\Omega_n} \le C\big(\|[\mathbf{f}, h]\|_{s-2,q,D} + \|B(\kappa^{-1}g)\|_{s-1,q,D}\big). \tag{4.5}$$

To assemble these results into a result for the entire polygon it is convenient to recall the following notation: $q_2^\lambda = \min_n\{2/(s - \lambda_{1,n}) : P_n \text{ is a concave vertex}\}$ and $\mathcal{I}_s^* = \{n : 2/(s-\lambda_{1,n}) \ge q_2^\lambda\}$ for $s \ge 1$. Define the regular part $[\mathbf{u}_R, \sigma_R]$ of the solution $[\mathbf{u}, \sigma]$ by the formula

$$[\mathbf{u}_R, \sigma_R] = \sum_{n\in\mathcal{I}_s^*} [\mathbf{u}_{n,R}, \sigma_{n,R}] + \sum_{n\notin\mathcal{I}_s^*} [\mathbf{u}_n, \sigma_n].$$

Using (4.4), (4.5), and the triangle inequality we obtain the following theorem.

THEOREM 4.1. *Let D be a concave polygon. Suppose Assumption A holds. Let $q > 2$ be sufficiently close to 2 and $s \ge 1$. Assume that $[\mathbf{f}, g, h] \in \mathbf{H}^{s-2,q}(D) \times \mathrm{H}^{s-1,q}(D) \times \mathrm{H}^{s-2,q}(D)$. Suppose that μ and γ are sufficiently large.*

(a) *If*

$$s < \min_{n\in\mathcal{I}_s^*}\{\lambda_{1,n}\} + 1 + 2/q,$$

then there is a unique solution $[\mathbf{u}, p, \sigma] \in \mathbf{H}^{s-1,q}(D)\times\mathrm{H}^{s-2,q}(D)\times\mathrm{H}^{s-1,q}(D)$ of (1.7).

(b) *Let $\mathcal{I}_s^* \ne \emptyset$. If s is given with*

$$\max_{n\in\mathcal{I}_s^*}\{\lambda_{2,n}\} + 2/q < s \le 2,$$

the solution $[\mathbf{u}, \sigma]$ may be split into singular and regular parts

$$[\mathbf{u}, \sigma] = \sum_{n\in\mathcal{I}_s^*} [\mathbf{u}_{n,s}, \sigma_{n,s}] + [\mathbf{u}_R, \sigma_R] \tag{4.6}$$

with $[\mathbf{u}_R, p, \sigma_R] \in \mathbf{H}^{s,q}(D) \times \mathrm{H}^{s-1,q}(D) \times \mathrm{H}^{s,q}(D)$ and

$$[\mathbf{u}_{n,s}, \sigma_{n,s}] = [C_{1,n}\Phi_{1,n} + C_{2,n}\Phi_{2,n},\, C_{3,n}\phi_n],$$

where we denote by $\mathbf{C}_n = [C_{1,n}, C_{2,n}, C_{3,n}]$, constructed in Steps 1, 2, and 3 in section 3. Also there is a constant $K = C(C_0, \|\mathbf{U}_R\|_{2,q,D} + \sum_{n\in\mathcal{I}_s^*} |\mathbf{d}_n|)$ with a given constant vector $\mathbf{d}_n = [d_{1,n}, d_{2,n}, d_{3,n}]$ such that

$$\|[\mathbf{u}_R, \sigma_R]\|_{s,q,D} + \sum_{n\in\mathcal{I}_s^*} |\mathbf{C}_n| + \|p\|_{s-1,q,D} \le K\big(\|[\mathbf{f}, h]\|_{s-2,q,D} + \|B(\kappa^{-1}g)\|_{s-1,q,D}\big). \tag{4.7}$$

(c) *If D is convex, then $[\mathbf{u},\sigma]=[\mathbf{u}_R,\sigma_R]$ satisfies the inequality* (4.7).

Note that $1+2/q+\min_n\{\lambda_{1,n}\}>2$ for $q<4$ because $\lambda_{1,n}>\frac{1}{2}$. For solving the nonlinear problem (1.1) we will restrict the regularity exponent s to $s=2$, and we will also require $q\in(2,\frac{1}{2}q_1^*)$. As a minimal condition of the exponent q for the solution to be split, we first define the number

$$q_2^*=\min_n\left\{\frac{2}{2-\lambda_{1,n}}:P_n \text{ is a concave vertex}\right\}, \tag{4.8}$$

where the number $\lambda_{1,n}\in(\frac{1}{2},1)$ is the first leading singular exponent for the Lamé system (1.12). Let

$$\mathcal{I}^*=\left\{n:\frac{2}{2-\lambda_{1,n}}\ge q_2^*\right\}. \tag{4.9}$$

The number q_1^* is defined in this way:

$$q_1^*=\min_{n\in\mathcal{I}^*}\left\{\frac{2}{1-\lambda_{1,n}}:P_n \text{ is a concave vertex}\right\}. \tag{4.10}$$

5. Nonlinearity ($2<q<\frac{1}{2}q_1^*$). In order to solve the problem (1.1) we take $s=2$ and choose the exponent $q\in(2,\frac{1}{2}q_1^*)$. The data $[\mathbf{u}_0,p_0,\sigma_0]$ are assumed to be smooth functions. It is assumed that the vector field $\mathbf{u}_0$ satisfies the assumptions (A2)–(A4). Finally, it is assumed that the data $[\mathbf{u}_0,p_0,\sigma_0]$ are sufficiently close to constant; that is, $K_0:=\|[\nabla\mathbf{u}_0,\nabla\sigma_0]\|_{1,q,D}+\|\nabla p_0\|_{1,q,D}+|\sigma_0|_\infty$ is sufficiently small. If K_0 is small, this means that σ_0 is small, not just close to a constant.

We require some notation. Let m be the number of elements of $\mathcal{I}^*=\{n:q>\min\{q_2(\lambda_{2,n})\}\}$. Let $\Phi=[\Phi_1,\dots,\Phi_m]^t$ and $\phi=[\phi_1,\dots,\phi_m]$, where $\Phi_n=[\Phi_{1,n},\Phi_{2,n}]^t$ and ϕ_n are the leading singular functions of the Lamé system and the Laplace equation at P_n, $n\in\mathcal{I}^*$. For convenience let $\tilde\Phi=[\tilde\Phi_1,\dots,\tilde\Phi_m]$, where $\tilde\Phi_n=[\Phi_{1,n},\Phi_{2,n},\phi_n]$. We write $\mathbf{C}=[\mathbf{C}_1,\dots,\mathbf{C}_m]^t\in\mathbf{R}^{2m}$ with $\mathbf{C}_n=[C_{1,n},C_{2,n}]^t$, $\mathbf{C}_e=[C_{3,1},\dots,C_{3,m}]^t\in\mathbf{R}^m$, and $\mathbf{d}=[\mathbf{d}_1,\dots,\mathbf{d}_m]^t\in\mathbf{R}^{2m}$ with $\mathbf{d}_n=[d_{1,n},d_{2,n}]^t$, $\mathbf{d}_e=[d_{3,n},\dots,d_{3,m}]^t\in\mathbf{R}^m$. For convenience let $\tilde{\mathbf{C}}=[\tilde{\mathbf{C}}_1,\dots,\tilde{\mathbf{C}}_m]\in\mathbf{R}^{3m}$ with $\tilde{\mathbf{C}}_n=[\mathbf{C}_n,C_{3,n}]$ and $\tilde{\mathbf{d}}=[\tilde{\mathbf{d}}_1,\dots,\tilde{\mathbf{d}}_m]\in\mathbf{R}^{3m}$ with $\tilde{\mathbf{d}}_n=[\mathbf{d}_n,d_{3,n}]$. We denote the product of $\tilde{\mathbf{C}}$ and $\tilde\Phi$ by

$$\tilde{\mathbf{C}}\tilde\Phi:=[\mathbf{C}\Phi,\mathbf{C}_e\phi]=\sum_{n\in\mathcal{I}^*}[C_{1,n}\Phi_{1,n}+C_{2,n}\Phi_{2,n},\,C_{3,n}\phi_n]$$

and similarly for $\tilde{\mathbf{d}}\tilde\Phi$.

We use two Banach spaces: $\mathcal{Y}=\mathbf{H}^{1,q}(D)\times\mathrm{L}^q(D)\times\mathrm{H}^{1,q}(D)$ and $\mathcal{X}=\mathbf{H}^{2,q}(D)\times\mathrm{H}^{1,q}(D)\times\mathrm{H}^{2,q}(D)\times\mathbf{R}^{3m}$. We will use a bounded map $E:\mathcal{X}\to\mathcal{Y}$ defined as follows: $E[\mathbf{u}_R,p,\sigma_R,\tilde{\mathbf{C}}]=[\mathbf{u}_R+\mathbf{C}\Phi,p,\sigma_R+\mathbf{C}_e\phi]$. Let $B_K\subset\mathcal{X}$ be the ball of radius K, i.e.,

$$B_K=\left\{[\mathbf{w}_R,\eta,\tau_R,\tilde{\mathbf{d}}]\in\mathcal{X}:\|[\mathbf{w}_R,\tau_R]\|_{2,q,D}+\|\eta\|_{1,q,D}+\sum_{n\in\mathcal{I}^*}|\tilde{\mathbf{d}}_n|\le K\right\}.$$

For fixed data $[\mathbf{u}_0,p_0,\sigma_0]$ and K sufficiently small we define a map $T:B_K\to\mathcal{X}$ as

follows. Let $[\mathbf{w}_R, \eta, \tau_R, \tilde{\mathbf{d}}] \in B_K$. Let

$$
\begin{aligned}
&\mathbf{U} = \mathbf{w} + \mathbf{u}_0, \quad \mathbf{w} = \mathbf{w}_R + \mathbf{d}\Phi, \quad \tau = \tau_R + \mathbf{d}_e\phi,\\
&\rho = \rho(p_0 + \eta, \sigma_0 + \tau), \quad \kappa_i = \kappa_i(p_0 + \eta, \sigma_0 + \tau) \quad (i = 1, 2),\\
&\tilde{\rho} = \rho c_v, \quad \tilde{\tau} = \tau p_\sigma, \quad p_\sigma = p_\sigma(\rho + \rho_0, \tau + \sigma_0),\\
&\mathbf{f}_0 := \mu\Delta\mathbf{u}_0 + (\mu + \nu)\nabla \mathrm{div}\mathbf{u}_0 - \rho\mathbf{u}_0 \cdot \nabla\mathbf{u}_0 - \nabla p_0,\\
&g_0 := -\mathrm{div}\mathbf{u}_0 - \kappa_1\mathbf{u}_0 \cdot \nabla p_0 - \kappa_2\mathbf{u}_0 \cdot \nabla\sigma_0,\\
&h_0 := \gamma\Delta\sigma_0 - c_v\rho\mathbf{u}_0 \cdot \nabla\sigma_0 - \sigma_0 p_\sigma \mathrm{div}\mathbf{u}_0,\\
&\mathbf{f}(\mathbf{w}, \eta, \tau) = \mathbf{f}_0 - \rho\mathbf{w} \cdot \nabla\mathbf{u}_0,\\
&g(\mathbf{w}, \eta, \tau) = g_0 - \kappa_1\mathbf{w} \cdot \nabla p_0 - \kappa_2\mathbf{w} \cdot \nabla\sigma_0,\\
&h(\mathbf{w}, \eta, \tau) = h_0 - c_v\rho\mathbf{w} \cdot \nabla\sigma_0 + p_\sigma(\sigma_0 \mathrm{div}\mathbf{w} - \tau \mathrm{div}\mathbf{u}_0)\\
&\qquad\qquad + \psi(\mathbf{w} + \mathbf{u}_0, \mathbf{w} + \mathbf{u}_0).
\end{aligned}
\tag{5.1}
$$

With $[\mathbf{w}_R, \eta, \tau_R, \tilde{\mathbf{d}}] \in B_K$, the functions ρ, κ_i, p_σ are bounded functions in D with bound depending on K and K_0. If K is sufficiently small, $\mathbf{U}$ is so close to $\mathbf{u}_0$ that $\mathbf{U}$ satisfies Assumption A. With this restriction on K, let $[\mathbf{u}, p, \sigma]$ be the weak solution to the linear problem

$$
\begin{cases}
-\mu\Delta\mathbf{u} - (\mu + \nu)\nabla \mathrm{div}\mathbf{u} + \rho\,(\mathbf{U} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } D,\\
\mathrm{div}\mathbf{u} + \kappa_1\mathbf{U} \cdot \nabla p + \kappa_2\mathbf{U} \cdot \nabla\sigma = g \quad \text{in } D,\\
-\gamma\Delta\sigma + \tilde{\rho}\,\mathbf{U} \cdot \nabla\sigma + \tilde{\tau}\,\mathrm{div}\mathbf{u} = h \quad \text{in } D,\\
\mathbf{u} = 0,\ \sigma = 0 \quad \text{on } \partial D,\\
p = 0 \quad \text{on } \partial D_{in}.
\end{cases}
\tag{5.2}
$$

We apply Theorem 4.1 to obtain the decomposition $\mathbf{u} = \mathbf{C}\Phi + \mathbf{u}_R$ and $\sigma = \mathbf{C}_e\phi + \sigma_R$. From Theorem 4.1, $[\mathbf{u}_R, p, \sigma_R, \tilde{\mathbf{C}}] \in \mathcal{X}$. Set $T[\mathbf{w}_R, \eta, \tau_R, \tilde{\mathbf{d}}] = [\mathbf{u}_R, p, \sigma_R, \tilde{\mathbf{C}}]$.

LEMMA 5.1. *If K_0 and K are small enough, then $T(B_K) \subset B_K$.*

Proof. Let $[\mathbf{w}, \eta, \sigma] \in B_K$. From the formulas for $\mathbf{f}$, g, and h one sees that

$$
\begin{aligned}
\|\mathbf{f}\|_{0,q,D} &\le C(\|\nabla\mathbf{u}_0\|_{1,q,D} + \|\nabla p_0\|_{0,q,D})(1 + \|\mathbf{w}\|_{0,q,D}) \le CK_0(1 + K),\\
\|g\|_{1,q,D} &\le C\big(\|[\nabla p_0, \nabla\sigma_0]\|_{1,q,D}(\|\mathbf{w}\|_{1,q,D} + \|\nabla\mathbf{u}_0\|_{1,q,D}) + \|\nabla\mathbf{u}_0\|_{1,q,D}\big)\\
&\le CK_0(\|\mathbf{w}_R\|_{1,q,D} + |\mathbf{d}|) + C(K_0^2 + K_0)\\
&\le C(K_0K + K_0^2 + K_0).
\end{aligned}
$$

Next we compute $\|h\|_{0,q,D}$. To do this, first note that $|\psi(\mathbf{w} + \mathbf{u}_0, \mathbf{w} + \mathbf{u}_0)| \le C(\gamma_0 + \gamma_1)(|\nabla\mathbf{w}|^2 + |\nabla\mathbf{u}_0|^2)$. Second, we shall use the inequality $q < \frac{1}{2}q_1^*$ in estimating $\||\nabla\Phi|^2\|_{0,q,D}$ so that $\|\psi(\mathbf{w} + \mathbf{u}_0, \mathbf{w} + \mathbf{u}_0)\|_{0,q,D}$ can be bounded. Third, we will use the Sobolev imbedding theorem: $\mathbf{H}^{1,q}(D) \hookrightarrow \mathbf{L}^\infty(D)$. Now

$$
\begin{aligned}
\|h\|_{0,q,D} &\le \|h_0\|_{0,q,D} + C\|\mathbf{w} \cdot \nabla\sigma_0\|_{0,q,D} + |p_\sigma|_\infty\|\sigma_0 \mathrm{div}\mathbf{w} - \sigma \mathrm{div}\mathbf{u}_0\|_{0,q,D}\\
&\quad + \|\psi(\mathbf{w} + \mathbf{u}_0, \mathbf{w} + \mathbf{u}_0)\|_{0,q,D}\\
&\le CK_0(1 + \|\mathbf{w}\|_{1,q,D} + \|\sigma\|_{0,q,D}) + C\||\nabla\mathbf{w}|^2 + |\nabla\mathbf{u}_0|^2\|_{0,q,D}\\
&\le CK_0(1 + |\tilde{\mathbf{d}}| + \|\mathbf{w}_R\|_{1,q,D} + \|\sigma_R\|_{0,q,D})\\
&\quad + C(|\mathbf{d}|^2\||\nabla\Phi|^2\|_{0,q,D} + \||\nabla\mathbf{w}_R|^2 + |\nabla\mathbf{u}_0|^2\|_{0,q,D})\\
&\le CK_0(1 + |\tilde{\mathbf{d}}| + \|\mathbf{w}_R\|_{1,q,D} + \|\sigma_R\|_{0,q,D})\\
&\quad + C(|\mathbf{d}|^2 + \|\mathbf{w}_R\|_{2,q,D}^2 + \|\nabla\mathbf{u}_0\|_{1,q,D}^2)\\
&\le CK_0(1 + K) + C(K^2 + K_0^2).
\end{aligned}
$$

Hence $\|[\mathbf{f},h]\|_{0,q,D}$ and $\|g\|_{1,q,D}$ can be made arbitrarily small by making K_0 and K small enough. Using (4.7) we see that if K_0 and K are small enough, $[\mathbf{u}_R,p,\sigma_R,\tilde{\mathbf{C}}]\in B_K$. ∎

LEMMA 5.2. *If K_0 and K are small enough, there is an $a\in(0,1)$ such that for $[\mathbf{w}_R,\eta,\tau_R,\tilde{\mathbf{d}}]$ and $[\mathbf{w}_R^*,\eta^*,\tau_R^*,\tilde{\mathbf{d}}^*]$ in B_K,*

$$\begin{aligned}\|ET[\mathbf{w}_R,\eta,\tau_R,\tilde{\mathbf{d}}]-ET[\mathbf{w}_R^*,\eta^*,\tau_R^*,\tilde{\mathbf{d}}^*]\|_{\mathcal{Y}}\\ \le a\|E[\mathbf{w}_R,\eta,\tau_R,\tilde{\mathbf{d}}]-E[\mathbf{w}_R^*,\eta^*,\tau_R^*,\tilde{\mathbf{d}}^*]\|_{\mathcal{Y}}.\end{aligned}\tag{5.3}$$

Proof. For fixed data $[\mathbf{u}_0,p_0,\sigma_0]$, consider $ET[\mathbf{w}_R,\eta,\tau,\tilde{\mathbf{d}}]=[\mathbf{u},p,\sigma]$ and $ET[\mathbf{w}_R^*,\eta^*,\tau^*,\tilde{\mathbf{d}}^*]=[\mathbf{u}^*,p^*,\tau^*]$, where $[\mathbf{u},p,\sigma]$ and $[\mathbf{u}^*,p^*,\sigma^*]$ are the solutions of (5.2). Let $\rho=\rho(\eta+p_0,\tau+\sigma_0)$, $\rho^*=\rho(\eta^*+p_0,\tau^*+\sigma_0)$, $\kappa_i=\kappa_i(\eta+p_0,\tau+\sigma_0)$, $\kappa_i^*=\kappa_i(\eta^*+p_0,\tau^*+\sigma_0)$, $\tilde\tau=\tau p_\sigma(\rho+\rho_0,\tau+\sigma_0)$, and $\tilde\tau^*=\tau^*p_\sigma(\rho^*+\rho_0,\tau^*+\sigma_0)$. Then we have

$$\begin{cases}-\mu\Delta(\mathbf{u}-\mathbf{u}^*)-(\mu+\nu)\nabla\mathrm{div}(\mathbf{u}-\mathbf{u}^*)+\rho\,(\mathbf{U}\cdot\nabla)(\mathbf{u}-\mathbf{u}^*)+\nabla(p-p^*)=\mathbf{F},\\ \mathrm{div}(\mathbf{u}-\mathbf{u}^*)+\kappa_1\mathbf{U}\cdot\nabla(p-p^*)+\kappa_2\mathbf{U}\cdot\nabla(\sigma-\sigma_*)=G\quad\text{in } D,\\ -\gamma\Delta(\sigma-\sigma^*)+\tilde\rho\,\mathbf{U}\cdot\nabla(\sigma-\sigma^*)+\tilde\tau\mathrm{div}(\mathbf{u}-\mathbf{u}^*)=H,\\ \mathbf{u}-\mathbf{u}^*=0,\ \sigma-\sigma^*=0\quad\text{on }\partial D,\\ p-p^*=0\quad\text{on }\partial D_{in},\end{cases}\tag{5.4}$$

where $\mathbf{U}=\mathbf{d}\Phi+\mathbf{w}_R+\mathbf{u}_0$, $\mathbf{U}^*=\mathbf{d}^*\Phi+\mathbf{w}_R^*+\mathbf{u}_0$, $p_\sigma^*=p_\sigma(\rho^*+\rho_0,\tau^*+\sigma_0)$, and

$$\begin{aligned}\mathbf{F}&=\mathbf{f}_0-\mathbf{f}_0^*+(\rho^*-\rho)[\mathbf{U}\cdot\nabla\mathbf{u}^*+\mathbf{w}^*\cdot\nabla\mathbf{u}_0]\\&\quad+(\mathbf{w}^*-\mathbf{w})\cdot[\rho\nabla\mathbf{u}_0+\rho^*\nabla\mathbf{u}^*],\\ G&=g_0-g_0^*+\mathbf{w}^*\cdot[(\kappa_1^*-\kappa_1)\nabla p_0+(\kappa_2^*-\kappa_2)\nabla\sigma_0]\\&\quad+(\mathbf{w}^*-\mathbf{w})\cdot[\kappa_1\nabla p_0+\kappa_2\nabla\sigma_0+\kappa_1\nabla p^*+\kappa_2\nabla\sigma^*]\\&\quad+\mathbf{U}^*\cdot[(\kappa_1^*-\kappa_1)\nabla p^*+(\kappa_2^*-\kappa_2)\nabla\sigma^*],\end{aligned}\tag{5.5}$$

$$\begin{aligned}H&=h_0-h_0^*+c_v[(\rho^*-\rho)\mathbf{w}^*\cdot\nabla\sigma_0+\rho(\mathbf{w}^*-\mathbf{w})\cdot\nabla\sigma_0]\\&\quad+(p_\sigma-p_\sigma^*)[\sigma_0\mathrm{div}\mathbf{w}-\tau^*\mathrm{div}\mathbf{u}_0]+\sigma_0p_\sigma^*\mathrm{div}(\mathbf{w}-\mathbf{w}^*)\\&\quad+(\tau^*-\tau)p_\sigma\mathrm{div}\mathbf{u}_0+\psi(\mathbf{U},\mathbf{U})-\psi(\mathbf{U}^*,\mathbf{U}^*)\\&\quad+(\tilde\rho^*-\tilde\rho)\mathbf{U}\cdot\nabla\sigma^*+\tilde\rho^*(\mathbf{U}^*-\mathbf{U})\cdot\nabla\sigma^*\\&\quad+(\tilde\tau^*-\tilde\tau)\mathrm{div}\mathbf{u}^*.\end{aligned}\tag{5.6}$$

Applying to (5.4) the inequality (2.36) given in Lemma 2.9,

$$\begin{aligned}\|[\mathbf{u}-\mathbf{u}^*,\sigma-\sigma^*]\|_{1,q,D}+\|p-p^*\|_{0,q,D}\\ \le C\big(\|[\mathbf{F},H]\|_{-1,q,D}+\|B(\kappa^{-1}G)\|_{0,q,D}\big).\end{aligned}\tag{5.7}$$

We first compute $\|\mathbf{F}\|_{-1,q,D}$. In doing this, one has to be careful in estimating the following term: $\|(\rho-\rho^*)\mathbf{U}\cdot\nabla\mathbf{u}^*\|_{-1,q,D}$. For $\mathbf{u}^*=\mathbf{u}_R^*+\mathbf{C}^*\Phi$, it will be enough if one can estimate the quantity $\|(\rho-\rho^*)\nabla\Phi\|_{-1,q,D}$. Using the Hölder inequality, we have

$$\begin{aligned}\|(\rho-\rho^*)\nabla\Phi\|_{-1,q,D}&\le C\|\rho-\rho^*\|_{0,q,D}\\&\quad\cdot\sum_{n\in\mathcal{I}^*}\sup_{v\in H_0^{1,q'}}\frac{1}{\|v\|_{1,q',D}}\left\{\int_D\frac{|\chi_n\,v|^{q'}}{|r_n(x,y)|^{q'(1-\lambda_{1,n})}}\mathrm{d}\mathbf{x}\right\}^{1/q'},\end{aligned}\tag{5.8}$$

where $r_n(x,y)=\sqrt{(x-x_n)^2+(y-y_n)^2}$ is the distance function to the concave vertex P_n and χ_n is a smooth cutoff function having support near the concave vertex P_n. From the Hardy's inequality [6, Theorem 330], for $q'\neq 2$,

$$\int_0^\infty r_n^{1-q'}[\chi_n v(r_n\cos\theta_n, r_n\sin\theta_n)]^{q'}dr_n \le C\int_0^\infty |\nabla v|^{q'} r_n dr_n.$$

Integrating over θ_n,

$$\int_D \frac{(\chi_n v)^{q'}}{r_n^{q'}}\mathrm{d}\mathbf{x} \le C\|v\|^{q'}_{1,q',D}.$$

Using this result we get

$$\begin{aligned}(5.9)\quad \|(\rho-\rho^*)\nabla\Phi\|_{-1,q,D} &\le C\|\rho-\rho^*\|_{0,q,D}\\ &\le C(|\rho_p|_\infty,|\rho_\sigma|_\infty)(\|\tau-\tau^*\|_{0,q,D}+\|\eta-\eta^*\|_{0,q,D}).\end{aligned}$$

So

$$\begin{aligned}\|(\rho-\rho^*)\mathbf{U}\cdot\nabla\mathbf{u}^*\|_{-1,q,D} &\le C\big\|(\rho-\rho^*)(|\nabla\mathbf{u}_R^*|+|\mathbf{C}^*|\,|\nabla\Phi|)\big\|_{-1,q,D}\\ (5.10)\qquad &\le C(\|\mathbf{u}_R^*\|_{2,q,D}+|\mathbf{C}^*|)\|[\tau-\tau^*,\eta-\eta^*]\|_{0,q,D},\end{aligned}$$

where $C=C(\|\rho'\|_\infty,\|\mathbf{U}^*\|_\infty)$. Hence we conclude that

$$(5.11)\qquad \|\mathbf{F}\|_{-1,q,D}\le C(K+K_0)(\|\mathbf{w}-\mathbf{w}^*\|_{1,q,D}+\|[\tau-\tau^*,\eta-\eta^*]\|_{0,q,D}),$$

where, using Lemma 5.1, we note that $\|\mathbf{u}_R^*\|_{2,q,D}+|\mathbf{C}^*|+\|\mathbf{u}^*\|_{1,q,D}+\|\nabla\mathbf{u}_0\|_{1,q,D}\le C(K+K_0)$.

Second, we estimate $\|B(\kappa^{-1}G)\|_{0,q,D}$. For this, the main difficulty is to estimate $B[\kappa_1^{-1}\mathbf{U}^*\cdot[(\kappa_1-\kappa_1^*)\nabla p^*]$. Since $\mathbf{U}^*\cdot\nabla p^*=(\kappa_1^*)^{-1}(g^*-\mathbf{C}^*\mathrm{div}\Phi-\mathrm{div}\mathbf{u}_R^*-\mathbf{C}_e^*\kappa_2^*\mathbf{U}^*\cdot\nabla\phi-\kappa_2^*\mathbf{U}^*\cdot\nabla\sigma_R^*)$, it is enough to estimate

$$B[\tau_1(\kappa_1-\kappa_1^*)(\mathbf{C}^*\mathrm{div}\Phi+\mathbf{C}_e^*\kappa_2^*\mathbf{U}^*\cdot\nabla\phi)],$$

where $\tau_1:=(\kappa_1\,\kappa_1^*)^{-1}$. So

$$\begin{aligned}&|B[\tau_1\,(\kappa_1^*-\kappa_1)(\mathbf{C}^*\mathrm{div}\Phi+\mathbf{C}_e^*\mathbf{U}^*\cdot\nabla\phi)](x,y)|\\ &\quad\le C\sum_{n\in\mathcal{I}^*}|\tilde{\mathbf{C}}_n^*|\int_{\delta(\xi)}^x|\kappa_1^*-\kappa_1|\,\big|(\nabla\Phi+\nabla\phi)(s,h(s,\xi))\big|\,ds\\ &\quad\le C\sum_{n\in\mathcal{I}^*}|\tilde{\mathbf{C}}_n^*|\int_{\delta(\xi)}^x|\kappa_1^*-\kappa_1|\,[s^2+h^2(s,\xi)]^{\frac{\lambda_{1,n}-1}{2}}\,ds\\ &\quad\text{(using the Hölder inequality and } 1/q+1/q'=1)\\ &\quad\le C\sum_{n\in\mathcal{I}^*}|\tilde{\mathbf{C}}_n^*|\bigg(\int_{\delta(\xi)}^x|\kappa_1^*-\kappa_1|^q\,ds\bigg)^{\frac{1}{q}}\bigg(\int_{\delta(\xi)}^x[s^2+h^2(s,\xi)]^{(\frac{\lambda_{1,n}-1}{2})q'}\,ds\bigg)^{\frac{1}{q'}}\\ &\quad\text{(noting that } \tfrac{1}{\lambda_{1,n}}<2<q \text{ for the concave vertex } P_n)\\ (5.12)\qquad &\quad\le C|\tilde{\mathbf{C}}^*|\bigg(\int_{\delta(\xi)}^x|\kappa_1^*-\kappa_1|^q\,ds\bigg)^{\frac{1}{q}},\end{aligned}$$

where $C = C(\|\tau\|_\infty)$. Using (5.12) and $\kappa_1^* - \kappa_1 = \kappa_{1p}(\xi_1,\xi_2)(\eta-\eta^*) + k_{1\sigma}(\xi_1,\xi_2)(\tau-\tau^*)$ for some $[\xi_1,\xi_2]$ we have

$$\|B[\kappa_1^{-1}(\kappa_1-\kappa_1^*)\mathbf{U}^*\cdot\nabla p^*]\|_{0,q,D} \le CK\|[\eta-\eta^*,\tau-\tau^*]\|_{0,q,D}, \tag{5.13}$$

where, using Lemma 5.1, we note that $\|g^*\|_{1,q,D} + |\tilde{\mathbf{C}}^*| + \|[\mathbf{u}_R^*,\sigma_R^*]\|_{2,q,D} \le CK$. Similarly, the term $B[\kappa_1^{-1}\mathbf{U}^*\cdot[(\kappa_2-\kappa_2^*)\nabla\sigma^*]$ can be bounded by the right-hand side of (5.13). Using (5.5) and $\|[p^*,\sigma^*]\|_{1,q,D} + \|[\nabla p_0,\nabla\sigma_0]\|_{1,q,D} \le C(K+K_0)$ by Lemma 5.1, we have

$$\|B(\kappa_1^{-1}G)\|_{0,q,D} \le CK_1(\|\mathbf{w}-\mathbf{w}^*\|_{1,q,D} + \|[\eta-\eta^*,\tau-\tau^*]\|_{0,q,D}), \tag{5.14}$$

where $K_1 := K + K_0$ and $C = C(|\nabla\kappa_1|_\infty)$.

Third, we estimate $\|H\|_{-1,q,D}$. Since

$$|\psi(\mathbf{U},\mathbf{U}) - \psi(\mathbf{U}^*,\mathbf{U}^*)| \le C(|\nabla(\mathbf{w}+\mathbf{w}^*)| + |\nabla\mathbf{u}_0|)|\nabla(\mathbf{w}-\mathbf{w}^*)|$$

with $C = C(\gamma_0+\gamma_1)$, and replacing $\rho-\rho^*$ in (5.8) by $|\nabla(\mathbf{w}-\mathbf{w}^*)|$, following the same procedures used there, we have

$$\begin{aligned}
&\|\psi(\mathbf{U},\mathbf{U}) - \psi(\mathbf{U}^*,\mathbf{U}^*)\|_{-1,q,D}\\
&\le C(\|\nabla\mathbf{w}_R + \nabla\mathbf{w}_R^*\|_\infty + |\nabla\mathbf{u}_0|_\infty)\|\nabla(\mathbf{w}-\mathbf{w}^*)\|_{0,q,D}\\
&\quad + C(|\mathbf{d}| + |\mathbf{d}^*|)\|\nabla(\mathbf{w}-\mathbf{w}^*)\|_{0,q,D}\\
&\le C(K+K_0)\|\mathbf{w}-\mathbf{w}^*\|_{1,q,D}.
\end{aligned}$$

So the following inequality can be easily obtained:

$$\begin{aligned}
\|h-h^*\|_{-1,q,D} &\le C(K_0+K)\|[\eta-\eta^*,\tau-\tau^*]\|_{0,q,D}\\
&\quad + C(K_0 + |\sigma_0 p_\sigma^*|_\infty)\|\mathbf{w}-\mathbf{w}\|_{1,q,D}\\
&\quad + C(K_0+K)\|\mathbf{w}-\mathbf{w}^*\|_{1,q,D}.
\end{aligned} \tag{5.15}$$

Furthermore,

$$\begin{aligned}
&\|(\tilde\rho^*-\tilde\rho)\mathbf{U}\cdot\nabla\sigma^*\|_{-1,q,D} \le C(|\mathbf{C}_e^*| + \|\sigma_R^*\|_{2,q,D})\|\rho^*-\rho\|_{0,q,D},\\
&\|\tilde\rho^*(\mathbf{w}^*-\mathbf{w})\cdot\nabla\sigma^*\|_{-1,q,D} \le C\|\sigma^*\|_{1,q,D}\|\mathbf{w}-\mathbf{w}^*\|_{1,q,D}.
\end{aligned} \tag{5.16}$$

Also, replacing $\rho-\rho^*$ in (5.8) by $p_\sigma^* - p_\sigma$, we have

$$\begin{aligned}
\|(\tilde\tau^*-\tilde\tau)\mathrm{div}\mathbf{u}^*\|_{-1,q,D} &\le C|\mathbf{C}^*|\|(p_\sigma^*-p_\sigma)\mathrm{div}\Phi\|_{-1,q,D}\\
&\quad + C\|\mathbf{u}_R^*\|_{2,q,D}\|p_\sigma^*-p_\sigma\|_{0,q,D} + C\|\mathbf{u}^*\|_{1,q,D}\|\tau-\tau^*\|_{1,q,D}\\
&\le C(|\mathbf{C}^*| + \|\mathbf{u}_R^*\|_{2,q,D})\|[\eta-\eta^*,\tau-\tau^*]\|_{0,q,D}\\
&\quad + CK\|\tau-\tau^*\|_{1,q,D}.
\end{aligned} \tag{5.17}$$

Combining (5.6) and (5.15)–(5.17) we have

$$\begin{aligned}
\|H\|_{-1,q,D} &\le C(K+K_0)\|[\mathbf{w}-\mathbf{w}^*,\tau-\tau^*]\|_{1,q,D}\\
&\quad + CK\|\eta-\eta^*\|_{0,q,D}.
\end{aligned} \tag{5.18}$$

Consequently, using (5.11), (5.14), and (5.18), we obtain

$$\begin{aligned}
&\|[\mathbf{u}-\mathbf{u}^*,\sigma-\sigma^*]\|_{1,q,D} + \|p-p^*\|_{0,q,D}\\
&\quad \le a\big(\|[\mathbf{w}-\mathbf{w}^*,\tau-\tau^*]\|_{1,q,D} + \|\eta-\eta^*\|_{0,q,D}\big),
\end{aligned} \tag{5.19}$$

where $a := C(K + K_0)$. If $K + K_0$ is small enough, (5.3) follows. □

Proof of Theorem 1.2. Let $X^0 \in B_K$ be given. Define $X^j = TX^{j-1}$ for $j = 1, 2, \ldots$. Let $Y^j = EX^j$. From (5.3), $\|Y^j - Y^{j-1}\|_{\mathcal{Y}} = \|EX^j - EX^{j-1}\|_{\mathcal{Y}} = \|ETX^{j-1} - ETX^{j-2}\|_{\mathcal{Y}} \le a\|EX^{j-1} - EX^{j-2}\|_{\mathcal{Y}} = a\|Y^{j-1} - Y^{j-2}\|_{\mathcal{Y}} \le a^2 \times \|Y^{j-2} - Y^{j-3}\|_{\mathcal{Y}} \le \cdots \le a^{j-1}\|Y^1 - Y^0\|_{\mathcal{Y}}$. Hence, if $j < k$,

$$\begin{aligned}\|Y^k - Y^j\|_{\mathcal{Y}} &\le \sum_{l=j}^{k-1} \|Y^{l+1} - Y^l\|_{\mathcal{Y}} \\ &\le \left(\sum_{l=j}^{k-1} a^l\right)\|Y^1 - Y^0\|_{\mathcal{Y}} \\ &\le \frac{a^j}{1-a}\|Y^1 - Y^0\|_{\mathcal{Y}} \\ &\to 0 \text{ as } j, k \to \infty.\end{aligned}$$

Therefore the sequence $\{Y^j\}$ is a Cauchy sequence in $\mathcal{Y}$. Hence there is a $Y \in \mathcal{Y}$ such that $\|Y - Y^j\|_{\mathcal{Y}} \to 0$.

Set $X^j = [\mathbf{w}_R^j, \eta^j, \tau_R^j, \tilde{\mathbf{d}}^j]$, $Y^j = [\mathbf{w}^j, \eta^j, \tau^j]$. We have

[a] $[\mathbf{w}^j, \tau^j] \to [\mathbf{w}, \tau] \in \mathbf{H}^{1,q}(D) \times \mathrm{H}^{1,q}(D)$, with the convergence in the topology of $\mathbf{H}^{1,q}(D) \times \mathrm{H}^{1,q}(D)$;

[b] $\eta^j \to \eta \in \mathrm{L}^q(D)$, with the convergence in the topology of $\mathrm{L}^q(D)$.

Since $\{X^j\}$ is a bounded sequence in $\mathcal{X}$, using various compact embeddings we may pick a subsequence $\{X^{j_l}\}$ such that in addition to the above convergences the following holds:

[c] $[\mathbf{w}_R^{j_l}, \tau_R^{j_l}]$ converges *weakly* in the topology of $\mathbf{H}^{2,q}(D) \times \mathrm{H}^{1,q}(D)$ to a function $[\mathbf{w}_R, \tau_R] \in \mathbf{H}^{2,q}(D) \times \mathrm{H}^{1,q}(D)$;

[d] $[\mathbf{w}_R^{j_l}, \tau_R^{j_l}]$ and $[\nabla\mathbf{w}_R^{j_l}, \tau_R^{j_l}]$ converge *uniformly as sequences of continuous functions* to $[\mathbf{w}_R, \tau_R]$ and $[\nabla\mathbf{w}_R, \nabla\tau_R]$, respectively;

[e] η^{j_l} converges *uniformly as a sequence of continuous functions* to η;

[f] the vectors $\tilde{\mathbf{d}}^{j_l}$ converge to a vector, call it $\tilde{\mathbf{d}}$. Furthermore, $\mathbf{w} = \mathbf{w}_R + \mathbf{d}\Phi$ and $\tau = \tau_R + \mathbf{d}_e\phi$. From [d], the sequence of functions $\tau^{j_l} = \tau_R^{j_l} + \mathbf{d}_e^{j_l}\phi$ converges uniformly to τ.

The weak limits $[\mathbf{w}_R, \tau_R]$ and the limits $\tilde{\mathbf{d}}$ are the same for any subsequences. For if two subsequences give rise to two limits, $[\mathbf{w}_R, \tau_R, \tilde{\mathbf{d}}]$ and $[\mathbf{w}_R^*, \tau_R^*, \tilde{\mathbf{d}}^*]$, we have $\mathbf{w} = \mathbf{w}_R + \Phi\mathbf{d} = \mathbf{w}_R^* + \Phi\mathbf{d}^*$, $\tau = \tau_R + \phi\mathbf{d}_e = \tau_R^* + \phi\mathbf{d}_e^*$. Therefore $\mathbf{w}_R - \mathbf{w}_R^* = \Phi(\mathbf{d}^* - \mathbf{d}) \in \mathbf{H}^{2,q}(D)$, $\tau_R - \tau_R^* = \phi(\mathbf{d}_e - \mathbf{d}_e^*) \in \mathrm{H}^{2,q}$, which is impossible unless $\tilde{\mathbf{d}} = \tilde{\mathbf{d}}^*$.

Define $\rho, \kappa_i, p_\sigma, \mathbf{U}, \mathbf{f}, g, h$ by (5.1), and define $\rho^{j_l}, \kappa_i^{j_l}, p_\sigma^{j_l}, \mathbf{U}^{j_l}, \mathbf{f}^{j_l}, g^{j_l}$ by (5.1) with appropriate superscripts j_l attached. From [e] and [f], $\rho^{j_l} \to \rho$, $\kappa_i^{j_l} \to \kappa_i$, and $p_\sigma^{j_l} \to p_\sigma$ uniformly as continuous functions. Also, from [d], and using the theory of initial value problems for ordinary differential equations, the functions $k^{j_l}(x, \bar{y})$ converge uniformly to $k(x, \bar{y})$, and $B^{j_l} \to B$ as operators on $L^q(D)$ or $H^{1,q}(D)$.

We now show that $[\mathbf{f}^{j_l}, g^{j_l}, h^{j_l}] \to [\mathbf{f}, g, h]$ in the topology of $\mathbf{L}^q(D) \times \mathrm{H}^{1,q}(D) \times \mathrm{L}^q(D)$. The terms in $\mathbf{f}^{j_l}$ are uniformly convergent, so the sequence $\mathbf{f}^{j_l}$ converges to $\mathbf{f}$ in $\mathrm{L}^q(D)$. To show the convergence of g^{j_l} in $\mathrm{H}^{1,q}(D)$ we must consider the terms $\kappa_1^{j_l}\mathbf{w}^{j_l} \cdot \nabla p_0$ and $\kappa_2^{j_l}\mathbf{w}^{j_l} \cdot \nabla\sigma_0$. Write

$$\kappa_1\mathbf{w} \cdot \nabla p_0 - \kappa_1^{j_l}\mathbf{w}^{j_l} \cdot \nabla p_0 = (\kappa_1 - \kappa_1^{j_l})\mathbf{w} \cdot \nabla p_0 + \kappa_1^{j_l}(\mathbf{w} - \mathbf{w}^{j_l}) \cdot \nabla p_0 := I_{j_l} + II_{j_l}.$$

Evidently, $\|I_{j_l}\|_{0,q,D} \le C\|\kappa_1 - \kappa_1^{j_l}\|_{\infty,D}\|\mathbf{w}\|_{0,q,D} \to 0$, $\|II_{j_l}\|_{0,q,D} \le C\|\mathbf{w} - \mathbf{w}^{j_l}\|_{0,q,D} \to$

0, so we have convergence in $\mathrm{L}^q(D)$. To show convergence of the derivatives, let D denote any first order derivative. A typical term in DI_{j_l} is $(\kappa_1 - \kappa_1^{j_l})D\mathbf{w}\cdot\nabla p_0$. Using [d] and [e] one sees that the sequence of continuous functions $\kappa_1^{j_l}$ converges uniformly to κ_1. Hence

$$\int_D |(\kappa_1 - \kappa_1^{j_l})D\mathbf{w}\cdot\nabla p_0|^q \mathrm{d}\mathbf{x} \le \max_{\mathbf{x}\in D}|\kappa_1 - \kappa_1^{j_l}|^q \int_D |D\mathbf{w}\cdot\nabla p_0|^q \mathrm{d}\mathbf{x} \to 0.$$

The other terms in ∇g^{j_l} are handled in the same way. All the terms except one in h^{j_l} are treated similarly. To establish the convergence of ψ^{j_l} to ψ in $\mathrm{L}^q(D)$ we note that $\psi^{j_l} = \psi(\mathbf{w}_R^{j_l} + \mathbf{d}^{j_l}\Phi + \mathbf{u}_0, \mathbf{w}_R^{j_l} + \mathbf{d}^{j_l}\Phi + \mathbf{u}_0)$ contains products of the terms $D\mathbf{w}_R^{j_l}$, $\mathbf{d}^{j_l}D\Phi$, and $D\mathbf{u}_0$, where D is an arbitrary first derivative. From [d], $D\mathbf{w}_R^{j_l}$ converges uniformly to $D\mathbf{w}_R$, and from [f], $\mathbf{d}^{j_l}\to\mathbf{d}$. The convergence of ψ^{j_l} follows, provided $|D\Phi|^2 \in \mathrm{L}^q(D)$. Since $D\Phi$ behaves like $r^{\lambda_{1,n}-1}$ at vertex P_n, we must have $q < 1/(1-\lambda_{1,n})$ for each concave vertex P_n. Since $\lambda_{1,n} > \frac{1}{2}$, this gives the restriction

$$q < \min\left\{\frac{1}{1-\lambda_{1,n}} : P_n \text{ is a concave vertex}\right\}, \tag{5.20}$$

which is equivalent to the inequality $q < \frac{1}{2}q_1^*$. With (5.20), h^{j_l} converges to h in $\mathrm{L}^q(D)$.

Let $[\mathbf{u}, p, \sigma]$ be the solution of (5.2) given by Lemma 2.9. Similarly, attach superscripts j_l to the appropriate terms of (5.2), and let $[\mathbf{u}^{j_l}, p^{j_l}, \sigma^{j_l}]$ be the solution of the resulting equation given by Lemma 2.9. Thus, $[\mathbf{u}^{j_l}, p^{j_l}, \sigma^{j_l}] = [\mathbf{w}^{j_l+1}, \eta^{j_l+1}, \tau^{j_l}]$. We now ask whether $[\mathbf{u}^{j_l}, p^{j_l}, \sigma^{j_l}]$ converges to $[\mathbf{u}, p, \sigma]$ in the topology of $\mathcal{Y}$. For this we recall the definition of weak solution given in section 2. Letting $a^{j_l}, \tilde{a}^{j_l}, b^{j_l}, B^{j_l}, \mathcal{E}^{j_l}$ be the bilinear operators and operators corresponding to index j_l, $[\mathbf{u}^{j_l}, p^{j_l}, \sigma^{j_l}]$ satisfies (2.35), (2.32), and (2.33) with indices j_l attached. One has

$$\begin{aligned}
&a^{j_l}(\mathbf{v}^1, \mathbf{v}^2) \to a(\mathbf{v}^1, \mathbf{v}^2), \quad \mathbf{v}^1, \mathbf{v}^2 \in \mathbf{H}_0^1(D),\\
&b^{j_l}(p, \mathbf{v}) \to b(p, \mathbf{v}), \quad p \in \mathrm{L}^q(D),\ \mathbf{v} \in \mathbf{H}_0^1(D),\\
&b^{j_l}(B^{j_l}(\kappa_1^{j_l-1}\mathrm{div}\mathbf{v}^1), \mathbf{v}^2) \to b(B(\kappa_1^{-1}\mathrm{div}\mathbf{v}^1), \mathbf{v}^2), \quad \mathbf{v}^1, \mathbf{v}^2 \in \mathbf{H}_0^1(D),\\
&b^{j_l}(B^{j_l}\mathcal{S}^{j_l}\mathbf{v}^1, \mathbf{v}^2) \to b(B\mathcal{S}\mathbf{v}^1, \mathbf{v}^2), \quad \mathbf{v}^1, \mathbf{v}^2 \in \mathbf{H}_0^1(D),
\end{aligned}$$

where $\mathcal{S}^{j_l} := \bar{\mathbf{U}}^{j_l}\cdot\nabla\mathcal{E}^{j_l}(\tau^{j_l}\mathrm{div})$, and

$$\begin{aligned}
&\tilde{a}^{j_l}(\mathbf{v}^1, \mathbf{v}^2) \to \tilde{a}(\mathbf{v}^1, \mathbf{v}^2), \quad \mathbf{v}^1, \mathbf{v}^2 \in \mathbf{H}_0^1(D),\\
&e^{j_l}(\sigma, \eta) \to e(\sigma, \eta), \quad \sigma, \eta \in \mathrm{H}_0^1(D),\\
&\tilde{b}^{j_l}(\chi, \mathbf{v}) \to \tilde{b}(\chi, \mathbf{v}), \quad \chi \in \mathrm{L}^q(D),\ \mathbf{v} \in \mathbf{H}_0^1(D).
\end{aligned}$$

Considering (2.35), (2.32), and (2.33), one sees that $[\mathbf{u}^{j_l}, \sigma^{j_l}] \to [\mathbf{u}, \sigma]$ in the topology of $\mathbf{H}^{1,q}(D)\times\mathrm{H}^{1,q}(D)$ and $p^{j_l} \to p$ in the topology of $\mathrm{L}^q(D)$. Furthermore, $[\mathbf{u}, p, \sigma]$ is a weak solution of (5.2). Since the coefficients and right-hand side of (5.2) are independent of the subsequence chosen, the pair $[\mathbf{u}, p, \sigma]$ is independent of the subsequence chosen.

We now show that $ETX^j \to [\mathbf{u}, p, \sigma]$ in the topology of $\mathcal{Y}$. Suppose the contrary. Then there is a number $a > 0$ and a subsequence X_{j_l} such that $\|ETX^{j_l} - [\mathbf{u}, p]\|_{\mathcal{Y}} \ge a$. By the above argument there is a subsequence of this subsequence, which we again denote by j_l, such that the convergence assertions [a]–[f] hold, which is a contradiction. Hence $\|ETX^j - [\mathbf{u}, p, \sigma]\|_{\mathcal{Y}} \to 0$.

Since $[\mathbf{u}^j, p^j, \sigma^j] = ETX^j = [\mathbf{w}^{j+1}, \eta^{j+1}, \tau^{j+1}]$, we have $[\mathbf{w}^j, \eta^j, \tau^j] \to [\mathbf{u}, p, \sigma]$ in the topology of $\mathcal{Y}$. From [d] and [f], $[\mathbf{w}_R^j, \tau_R^j] \to [\mathbf{w}_R, \tau_R] = [\mathbf{u}_R, \sigma_R]$ in the topology of $(C^1(\bar{D}))^3$ and $\tilde{\mathbf{d}}^j \to \tilde{\mathbf{d}} = \tilde{\mathbf{C}}$. Hence $[\mathbf{u}, p, \sigma]$ solves (1.5), so $[\mathbf{u} + \mathbf{u}_0,\ p + p_0,\ \sigma + \sigma_0]$ solves (1.1).

Acknowledgments. We thank the anonymous referees for several kind comments.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] J. D. Anderson, Jr., *Fundamentals of Aerodynamics*, 2nd ed., McGraw-Hill, New York, 1991.
[3] G. K. Batchelor, *An Introduction to Fluid Dynamics*, Cambridge University Press, Cambridge, UK, 2000.
[4] H. Beirão da Veiga, *An L^p-theory for the n-dimensional, stationary, compressible Navier-Stokes equations, and the incompressible limit for compressible fluids. The equilibrium solutions*, Comm. Math. Phys., 109 (1987), pp. 229–248.
[5] P. Grisvard, *Elliptic Problems in Non-smooth Domains*, Pitman Advanced Publishing Program, Boston, London, Melbourne, 1985.
[6] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, UK, 1973.
[7] J. Kadlec, *The regularity of the solution of the Poisson problem in a domain whose boundary is similar to that of a convex domain*, Czechoslovak Math. J., 89 (1964), pp. 386–393.
[8] R. B. Kellogg and J. E. Osborn, *A regularity result for the Stokes problem in a convex domain*, J. Funct. Anal., 21 (1976), pp. 397–431.
[9] V. A. Kozlov, V. G. Maz'ya, and J. Rossmann, *Spectral Problems Associated with Corner Singularities of Solutions to Elliptic Equations*, AMS, Providence, RI, 2001.
[10] J. R. Kweon and R. B. Kellogg, *Compressible Navier–Stokes equations in a bounded domain with inflow boundary condition*, SIAM J. Math. Anal., 28 (1997), pp. 94–108.
[11] J. R. Kweon and R. B. Kellogg, *Smooth solution of the compressible Navier-Stokes equations in an unbounded domain with inflow boundary condition*, J. Math. Anal. Appl., 220 (1998), pp. 657–675.
[12] J. R. Kweon and R. B. Kellogg, *Compressible Stokes problem on non-convex polygon*, J. Differential Equations, 176 (2001), pp. 290–314.
[13] J. R. Kweon and R. B. Kellogg, *Regularity of solutions to the Navier-Stokes equations for compressible barotropic flows on a polygon*, Arch. Ration. Mech. Anal., 163 (2002), pp. 35–64.
[14] S. A. Nazarov, A. Novotny, and K. Pileckas, *On steady compressible Navier-Stokes equations in plane damains with corners*, Math. Ann., 304 (1996), pp. 121–150.

SIAM J. MATH. ANAL.
Vol. 35, No. 6, pp. 1486–1511

STABLE PULSE SOLUTIONS FOR THE NONLINEAR SCHRÖDINGER EQUATION WITH HIGHER ORDER DISPERSION MANAGEMENT*

JAMISON T. MOESER†, CHRISTOPHER K. R. T. JONES‡, AND VADIM ZHARNITSKY§

Abstract. The evolution of optical pulses in fiber optic communication systems with strong, higher order dispersion management is modeled by a cubic nonlinear Schrödinger equation with periodically varying linear dispersion at second and third order. Through an averaging procedure, we derive an approximate model for the slow evolution of such pulses and show that this system possesses a stable ground state solution. Furthermore, we characterize the ground state numerically. The results explain the experimental observation of higher order dispersion managed solitons, providing theoretical justification for modern communication systems design.

Key words. higher order dispersion management, homogenization, periodic media, solitary waves, stability, nonlinear Schrödinger equation

AMS subject classifications. 35M99, 35B27, 49J40, 78A48

DOI. 10.1137/S0036141002412586

1. Introduction.

1.1. Conventional dispersion management. The technique of dispersion management (DM), introduced in the early 1980s [16] and refined during the past decade [27], has emerged as the dominant technology for high bandwidth data transmission through optical fibers. In a dispersion managed fiber link, short sections of fiber with opposite linear dispersion characteristics are joined together in a periodically repeated structure, forming a fiber whose linear dispersion is effectively canceled out over each period of DM. In such a system, the characteristic length of local dispersion is much shorter than that of nonlinearity or average dispersion so that on the scale of a typical DM segment the effects of nonlinearity and average dispersion can be made small relative to those of the local dispersion. In this regime, destabilizing effects such as four-wave mixing [2, 5, 22] and Gordon–Haus jitter [14, 36] are minimized.

In the case of DM at second order, the propagation equation for the wave envelope can be written in the dimensionless form

$$iu_z + d_2(z)u_{tt} + \epsilon|u|^2u = 0 \tag{1.1}$$

with the dispersion coefficient $d_2(z)$ decomposed into its varying and average components

$$d_2(z) = \tilde{d_2}(z) + \epsilon\alpha_2,$$

*Received by the editors August 2, 2002; accepted for publication (in revised form) August 29, 2003; published electronically March 11, 2004.
http://www.siam.org/journals/sima/35-6/41258.html

†Department of Applied Mathematics, University of Colorado, Boulder, CO 80309-0526 (moeser@babbage.colorado.edu). This author was supported by NSF grants DMS-0073923 and DMS-9810751.

‡Department of Mathematics, Program in Applied Mathematics, University of North Carolina, Chapel Hill, NC 27599-3250 (ckrtj@amath.unc.edu). This author was supported by NSF grant DMS-0073923.

§Department of Mathematics, University of Illinois at Urbana-Champaign, 1409 W. Green Street, Urbana, IL 61801 (vz@math.uiuc.edu). This author was supported by NSF grant DMS-0219233 and partially supported by NSF grant DMS-0073923.

where $\int_0^1 \tilde{d}_2(z')dz' = 0$. Typically, $d_2(z)$ is piecewise constant and periodic, and the parameter ϵ corresponds to the ratio of the characteristic length scales of local dispersion to that of nonlinearity and average dispersion [1, 10].

The system performance of dispersion managed fiber links is truly remarkable [8, 25]. Not only are stable pulse structures observed for $\alpha_2 > 0$, the focusing (anomalous) regime for the nonlinear Schrödinger (NLS) equation [9], but also for the case when $\alpha_2 \leq 0$ [14, 28]. These DM solitons are characterized by nearly Gaussian central peaks and rapidly decaying secondary peaks which comprise the tails [20, 28]. Also, in contrast to the soliton solutions for the NLS equation, the DM solitons possess a nontrivial quadratic phase component, namely, chirp [10].

The energy of the DM soliton is higher than that of the corresponding NLS soliton with the same full width at half maximum (FWHM) and average dispersion [29, 37]. This makes DM solitons more resistant to the effects of spontaneously emitted amplifier noise, giving dispersion managed systems a higher signal to noise ratio than traditional soliton based systems [34]. Also, this energy enhancement can be exploited to reduce energy variation per channel in wavelength division multiplexing (WDM) systems operating near zero average dispersion [29].

The first analytical results for DM solitons were obtained in the late 1990s, when an averaged equation describing the slow evolution of solutions to (1.1) was derived in [10] through path averaging and in [1] through multiple scales expansion. A rigorous justification for this averaged equation was given in [38], where, moreover, it was shown that for $\alpha_2 > 0$ the Hamiltonian corresponding to the averaged equation possesses a ground state solution in the class of functions $\mathcal{A}_\lambda = \{u : \int_{\mathbb{R}} |u|^2 = \lambda, \int_{\mathbb{R}} |u_t|^2 < \infty\}$. These results indicate the existence of a stable, stationary solution to the averaged equation that propagates nearly periodically for (1.1) on time scales up to $\mathcal{O}(\frac{1}{\epsilon})$. The existence of a standing wave solution for the averaged equation in this regime was also established in [12] by means of a general theorem on bifurcation of solutions from the essential spectrum. Furthermore, the existence of a ground state for the case $\alpha_2 = 0$ was recently demonstrated [13].

1.2. Higher order dispersion management. Waves of the form $\exp(iD(\omega)z - i\omega t)$ traveling through an optical fiber satisfy the dispersion relation

$$D(\omega) = \frac{n(\omega)\omega}{c},$$

where $D(\omega)$ is termed the propagation constant, $n(\omega)$ is the index of refraction, ω is the frequency, and c is the speed of light in a vacuum [3]. Thus, in general, the propagation constant is a complicated function of frequency. Explicit formulae for $D(\omega)$ are generally unknown, and in the derivation of the evolution equation for the electric field's slowly varying amplitude $D(\omega)$ is approximated by its Taylor polynomial in a neighborhood of the carrier frequency:

$$D(\omega) \simeq d_0 + d_1(\omega - \omega_0) + d_2(\omega - \omega_0)^2 + \cdots,$$

where $d_n = \frac{D^{(n)}(\omega_0)}{n!}$. In the derivation of the conventional DM model (1.1), one assumes that pulses are sufficiently narrow in the frequency domain, $|\omega - \omega_0| \ll 1$, so that $D(\omega)$ is accurately approximated by its quadratic Taylor polynomial, and the effects of changes in the values of d_2 in a neighborhood of ω_0 are neglected. However, for the propagation of pulses which are broader in the frequency domain, the inclusion of the cubic term in the Taylor approximation is necessary, with the resulting model taking into account variations in d_2.

This additional term in the Taylor series gives rise to a third order linear dispersive term in the governing NLS-type equation for the electric field's slowly varying envelope [3]. In single channel systems, third order dispersion generally causes an asymmetric broadening of pulses. Moreover, in WDM systems, which utilize many optical channels separated in the frequency domain, third order dispersion can prevent conventional dispersion compensation across neighboring channels.

A natural way to surmount these difficulties is to manage dispersion at both second and third order. By utilizing this technique of higher order dispersion management (HODM), the asymmetric broadening that takes place for ultrashort optical pulses in single channel systems with conventional DM is almost exactly compensated for. Furthermore, in WDM systems, HODM makes it possible to compensate for dispersion over many neighboring frequency channels simultaneously. In fact, advances in fiber manufacturing techniques [18] have made it possible to incorporate this idea into new optical fibers, termed *dispersion slope compensating fibers*, and recent experiments have yielded impressive results [7, 11, 19, 23, 26].

The evolution of optical pulses in a fiber with DM at second and third order is governed by the following dimensionless NLS-type equation [24]:

$$iu_z + d_2(z)u_{tt} + id_3(z)u_{ttt} + \epsilon_{nl}|u|^2u = 0,$$

with

$$d_j(z) = \tilde{d}_j(z) + \epsilon_j\alpha_j$$

and

$$\int_0^1 \tilde{d}_j(z')dz' = 0.$$

Here the α_j, $j = 2, 3$, are order one measures of average dispersion at second and third order, respectively, ϵ_{nl} is a small parameter representing the ratio of characteristic lengths of the local dispersions to the nonlinearity, and the ϵ_j are small parameters representing the ratio of characteristic lengths of the local dispersions to the average dispersions. The dispersion coefficients $d_j(z)$, $j = 2, 3$, are piecewise constant and, due to the manufacture process, periodic with the same period, here normalized to be 1. In the operating regimes we consider, the parameters satisfy $\epsilon_3 \ll \epsilon_{nl} \sim \epsilon_2$, so we set $\epsilon = \epsilon_{nl} = \epsilon_2$, neglect the effects of average third order dispersion by setting $\alpha_3 = 0$, and consider the equation

$$\begin{aligned} iu_z + d_2(z)u_{tt} + id_3(z)u_{ttt} + \epsilon|u|^2u = 0,\\ d_2(z) = \tilde{d}_2(z) + \epsilon\alpha_2,\\ d_3(z) = \tilde{d}_3(z). \end{aligned} \tag{1.2}$$

We develop an averaging theory for (1.2) and show that for the case $\alpha_2 > 0$ the corresponding averaged equation possesses a ground state solution which propagates nearly periodically for the full equation. Furthermore, we solve the Euler–Lagrange equation numerically, revealing the structure of this new DM soliton. We also report that analysis for the case $\alpha_2 = 0$ will appear elsewhere.

2. Averaging. Solutions of (1.2) evolve on two distinct spatial scales, which suggests performing an averaging procedure. We note that the analysis in this section does not require the condition $\alpha_3 = 0$, but it is necessary later when proving the existence of ground states.

2.1. Averaged equation. We first perform the transformation $u(z,t) = \mathcal{L}(z)\{v(z,t)\}$, where $\mathcal{L}\{\cdot\}$ is the unitary semigroup for the linear evolution equation

$$iu_z + \tilde{d}_2(z)u_{tt} + i\tilde{d}_3(z)u_{ttt} = 0. \tag{2.1}$$

The operator is easily computed via Fourier transform:

$$\mathcal{L}(z)\{v(0,t)\} = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \theta(z,k)\hat{v}(0,k)\exp(ikt)dk, \tag{2.2}$$

where

$$\theta(z,k) = \exp\int_0^z -i[k^2\tilde{d}_2(\tau) - k^3\tilde{d}_3(\tau)]d\tau.$$

We observe that $\mathcal{L}(z)$ is an isometry on $H^s(\mathbb{R})$ for all $s \in \mathbb{R}$. Moreover, due to the periodicity of $\tilde{d}_2(z)$ and $\tilde{d}_3(z)$, both $\theta(z,k)$ and $\mathcal{L}(z)$ are periodic in z. For ease of notation, we henceforth suppress the variable dependencies of v.

Using

$$\frac{\partial u}{\partial z} = \frac{\partial\big(\mathcal{L}(z)\{v\}\big)}{\partial z} = \mathcal{L}(z)\left\{\frac{\partial v}{\partial z}\right\} + i\tilde{d}_2(z)\frac{\partial^2\big(\mathcal{L}(z)\{v\}\big)}{\partial t^2} - \tilde{d}_3(z)\frac{\partial^3\big(\mathcal{L}(z)\{v\}\big)}{\partial t^3}$$

we obtain by direct substitution into (1.2) the evolution equation for v:

$$i\frac{\partial v}{\partial z} + \epsilon\left(\alpha_2\frac{\partial^2 v}{\partial t^2} + C(z)\{v\}\right) = 0, \tag{2.3}$$

where

$$C(z)\{v\} = \mathcal{L}(-z)\{|\mathcal{L}(z)\{v\}|^2\mathcal{L}(z)\{v\}\}. \tag{2.4}$$

Formally, the averaged equation is

$$i\frac{\partial v}{\partial z} + \epsilon\left(\alpha_2\frac{\partial^2 v}{\partial t^2} + \langle C\rangle\{v\}\right) = 0, \tag{2.5}$$

where

$$\langle C\rangle\{v\} = \int_0^1 \mathcal{L}(-z')\{|\mathcal{L}(z')\{v\}|^2\mathcal{L}(z')\{v\}\}dz'. \tag{2.6}$$

In Fourier space, (2.5) takes the form

$$i\frac{\partial \hat{v}}{\partial z} + \epsilon(-k^2\alpha_2\hat{v} + \langle\hat{C}\rangle\{v\}) = 0 \tag{2.7}$$

with

(2.8)

$$\langle\hat{C}\rangle\{v\} = \int_0^1\int_{\mathbb{R}^3}\delta(k-k_1+k_2-k_3)\Theta(z',k,k_1,k_2,k_3)\hat{v}_1(z)\bar{\hat{v}_2}(z)\hat{v}_3(z)dk_1dk_2dk_3dz'.$$

Here $\hat{v}_i(z) = \hat{v}(z,k_i)$ and

$$\Theta(z',k,k_1,k_2,k_3) = \exp\int_0^{z'} i\{-\tilde{d}_2(\tau)[k_1^2 - k_2^2 + k_3^2 - k^2] + \tilde{d}_3(\tau)[k_1^3 - k_2^3 + k_3^3 - k^3]\}d\tau.$$

Performing the integration over z' in (2.8) gives

$$\langle \hat{C} \rangle \{v\} = \int_{\mathbb{R}^3} \delta(k - k_1 + k_2 - k_3) \Theta_l(k, k_1, k_2, k_3) \hat{v}_1(z) \bar{\hat{v}}_2(z) \hat{v}_3(z) dk_1 dk_2 dk_3, \tag{2.9}$$

where

$$\Theta_l(k, k_1, k_2, k_3) = \int_0^1 \Theta(z', k, k_1, k_2, k_3) dz' \tag{2.10}$$

is a bounded function on $\mathbb{R}^4$. The averaged equation (2.5) corresponds to the variational equation

$$u_z = J \nabla \langle H \rangle,$$

where $J = -i$ is a skew-symmetric operator, ∇ is the Fréchet derivative, and $\langle H \rangle$ is the Hamiltonian

$$\langle H \rangle (v) = \alpha_2 \int_{\mathbb{R}} |v_t|^2 dt - \frac{1}{2} \int_0^1 \int_{\mathbb{R}} |\mathcal{L}(z')\{v\}|^4 dt dz'. \tag{2.11}$$

We note that $\langle H \rangle (v)$ is a bounded functional on $H^1(\mathbb{R})$, as

$$\begin{aligned} ||\mathcal{L}\{v\}||_{L^4}^4 &\le M \left|\left| \frac{\partial(\mathcal{L}\{v\})}{\partial t} \right|\right|_{L^2} \cdot ||\mathcal{L}\{v\}||_{L^2}^3 \\ &= M ||\mathcal{L}\{v_t\}||_{L^2} \cdot ||\mathcal{L}\{v\}||_{L^2}^3 \\ &= M ||v_t||_{L^2} \cdot ||v||_{L^2}^3, \end{aligned}$$

where we have used the Gagliardo–Nirenberg inequality [4] and the fact that $\mathcal{L}(z)$ is an isometry on any space $H^s(\mathbb{R})$.

We comment briefly on the regularity of the averaged operator $\langle C \rangle$. We will first show that $\langle C \rangle \{\cdot\}$ is bounded on $H^s(\mathbb{R})$ for $s > \frac{1}{2}$. Now

$$\begin{aligned} |\langle \hat{C} \rangle \{v\}| &= \left| \int_{\mathbb{R}^3} \delta(k - k_1 + k_2 - k_3) \Theta_l(k, k_1, k_2, k_3) \hat{v}(k_1) \bar{\hat{v}}(k_2) \hat{v}(k_3) dk_1 dk_2 dk_3 \right| \\ &\le ||\Theta_l||_{L^\infty(\mathbb{R}^4)} \int_{\mathbb{R}^3} \delta(k - k_1 + k_2 - k_3) |\hat{v}(k_1) \bar{\hat{v}}(k_2) \hat{v}(k_3)| dk_1 dk_2 dk_3 \\ &\le ||\Theta_l||_{L^\infty(\mathbb{R}^4)} \int_{\mathbb{R}^2} |\hat{v}(k_1)| |\bar{\hat{v}}(k_2)| |\hat{v}(k - k_1 + k_2)| dk_1 dk_2. \end{aligned}$$

If we denote $\hat{\mathcal{I}}\{v\} = \int_{\mathbb{R}^2} |\hat{v}(k_1)| |\bar{\hat{v}}(k_2)| |\hat{v}(k - k_1 + k_2)| dk_1 dk_2$, then by the above estimate it suffices to show that

$$||\mathcal{I}\{v\}||_{H^s(\mathbb{R})} \le ||v||_{H^s(\mathbb{R})}^3.$$

Now for any u and $w \in H^s(\mathbb{R})$, we have that

$$||uw||_{H^s(\mathbb{R})} \le ||u||_{H^s(\mathbb{R})} ||w||_{H^s(\mathbb{R})}, \tag{2.12}$$

or, equivalently, in Fourier domain,

$$||\hat{u} * \hat{w}||_{L^2_w(\mathbb{R})} \le ||\hat{u}||_{L^2_w(\mathbb{R})} ||\hat{w}||_{L^2_w(\mathbb{R})}, \tag{2.13}$$

where $*$ is the convolution operator and

$$||u||_{L^2_w(\mathbb{R})} = ||(1+|k|^2)^{\frac{s}{2}}\hat{u}||_{L^2(\mathbb{R})}.$$

If we denote

$$\hat{F}(k+k_2) = \int_{\mathbb{R}} |\hat{v}(k_1)||\hat{v}(k-k_1+k_2)|dk_1 = |\hat{v}| * |\hat{v}|,$$

then (2.12) and (2.13) applied to $\hat{u} = \hat{w} = |\hat{v}| \in L^2_w$ yield

$$||(|\hat{v}| * |\hat{v}|)\check{}||_{H^s(\mathbb{R})} \leq ||(|\hat{v}|)\check{}||^2_{H^s(\mathbb{R})} = ||v||^2_{H^s(\mathbb{R})},$$

where $\check{}$ denotes the inverse Fourier transform. Now

$$\hat{\mathcal{I}}\{v\} = \int_{\mathbb{R}} |\hat{v}(k_2)|\hat{F}(k+k_2)dk_2 = |\hat{v}| * \hat{F},$$

so we apply the above argument to $\hat{u} = |\hat{v}|$, $\hat{w} = \hat{F}$ to conclude that

$$||\mathcal{I}\{v\}||_{H^s(\mathbb{R})} \leq ||v||^3_{H^s(\mathbb{R})}.$$

A standard extension of this argument shows that $\langle C\rangle\{\cdot\}$ is locally Lipschitz on $H^s(\mathbb{R})$ for $s > \frac{1}{2}$:

$$||\langle C\rangle\{u-v\}||_{H^s(\mathbb{R})} \leq M||u-v||_{H^s(\mathbb{R})},$$

where M depends on $||u||_{H^s(\mathbb{R})}$ and $||v||_{H^s(\mathbb{R})}$.

2.2. Well posedness. The averaged equation (2.5) is similar in form to the focusing NLS equation, and local well posedness is a straightforward application of semigroup theory.

THEOREM 2.1. *If $v_0 \in H^s(\mathbb{R})$, $s > \frac{1}{2}$, then there exists $z_{max} > 0$ and a unique solution $v(z,t) \in C([0, z_{max}), H^s(\mathbb{R}))$ for (2.5) with initial data v_0, with the property that either $z_{max} = \infty$ or $z_{max} < \infty$ and $\lim_{z \to z_{max}} ||v(z)||_{H^s} = \infty$.*

Proof. The linear part can be solved via Fourier transform, generating a C_0 group of unitary operators $\mathcal{S}(z)$ on $H^s(\mathbb{R})$ for $z \in \mathbb{R}$. Since $\langle C\rangle\{\cdot\}$ is locally Lipschitz from $H^s(\mathbb{R}) \to H^s(\mathbb{R})$, local existence follows [33]. ∎

To prove a global existence theorem, a priori estimates on solutions of (2.5) of the form $||v(z)||_{H^s(\mathbb{R})} < C(z)$ for any $z \in \mathbb{R}^+$ are needed. This is possible for initial data in $H^s(\mathbb{R})$, $s \geq 1$, using conservation of the L^2 norm, conservation of the Hamiltonian, and regularity of the operator $\langle C\rangle$.

THEOREM 2.2. *If $v_0 \in H^s(\mathbb{R})$, $s \geq 1$, then there exists a unique solution $v(z,t) \in C([0,\infty), H^s(\mathbb{R}))$ for (2.5) with initial data v_0.*

Proof. Multiplying (2.7) by $\bar{\hat{v}}$, its conjugate by $\hat{v}$, subtracting, and integrating over $\mathbb{R}$ yield

$$\partial_z \int_{\mathbb{R}} |\hat{v}|^2 dk = -2\mathrm{Im}\int_{\mathbb{R}} \bar{\hat{v}}\langle\hat{C}\rangle\{v\}dk,$$

where

$$\int_{\mathbb{R}} \bar{\hat{v}}\langle\hat{C}\rangle\{v\}dk = \int_{\mathbb{R}} \bar{\hat{v}} \int_{\mathbb{R}^3} \delta(k-k_1+k_2-k_3)\Theta_l(\Delta_2,\Delta_3)\hat{v}_1\bar{\hat{v}}_2\hat{v}_3 dk_1 dk_2 dk_3 dk$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}^3} \delta(k - k_1 + k_2 - k_3)\Theta_l(\Delta_2, \Delta_3)\bar{\hat{v}}\hat{v}_1\bar{\hat{v}}_2\hat{v}_3 dk_1 dk_2 dk_3 dk$$

with

$$\Delta_2 = k_1^2 - k_2^2 + k_3^2 - k^2,$$
$$\Delta_3 = k_1^3 - k_2^3 + k_3^3 - k^3.$$

Since $\overline{\Theta_l(\Delta_2, \Delta_3)} = \Theta_l(-\Delta_2, -\Delta_3)$, making the change of variables $k \to k_3$, $k_1 \to k_2$ we see that

$$\int_{\mathbb{R}} \int_{\mathbb{R}^3} \delta(k - k_1 + k_2 - k_3)\Theta_l(\Delta_2, \Delta_3)\bar{\hat{v}}\hat{v}_1(z)\bar{\hat{v}}_2(z)\hat{v}_3(z) dk_1 dk_2 dk_3 dk$$

$$= \overline{\int_{\mathbb{R}} \int_{\mathbb{R}^3} \delta(k - k_1 + k_2 - k_3)\Theta_l(\Delta_2, \Delta_3)\bar{\hat{v}}\hat{v}_1(z)\bar{\hat{v}}_2(z)\hat{v}_3(z) dk_1 dk_2 dk_3 dk}$$

so that

$$\partial_z \int_{\mathbb{R}} |v|^2 = \partial_z \int_{\mathbb{R}} |\hat{v}|^2 dk = 0$$

and the L^2 norm is conserved.

By conservation of the Hamiltonian,

$$\langle H \rangle(v_0) = \alpha_2 \int_{\mathbb{R}} \left| \frac{\partial v_0}{\partial t} \right|^2 dt - \frac{1}{2} \int_0^1 \int_{\mathbb{R}} |\mathcal{L}(z')\{v_0\}|^4 dt dz'$$

$$= \langle H \rangle(v) = \alpha_2 \int_{\mathbb{R}} \left| \frac{\partial v}{\partial t} \right|^2 dt - \frac{1}{2} \int_0^1 \int_{\mathbb{R}} |\mathcal{L}(z')\{v\}|^4 dt dz'.$$

Thus

$$\begin{aligned} \int_{\mathbb{R}} \left| \frac{\partial v}{\partial t} \right|^2 dt &= \frac{\langle H \rangle(v_0)}{\alpha_2} + \frac{1}{2\alpha_2} \int_0^1 \int_{\mathbb{R}} |\mathcal{L}(z')\{v\}|^4 dt dz' \\ &\leq \frac{\langle H \rangle(v_0)}{\alpha_2} + \frac{M}{2\alpha_2} ||v||^{3/2}_{L^2(\mathbb{R})} \cdot \left|\left| \frac{\partial v}{\partial t} \right|\right|_{L^2(\mathbb{R})} \end{aligned}$$

by the Sobolev inequality. From standard estimates [35],

$$\int_{\mathbb{R}} \left| \frac{\partial v}{\partial t} \right|^2 dt \leq M'.$$

At this point we note that the boundedness of $||v(z)||_{H^1}$ implies the boundedness of $||v(z)||_{L^p}$ for $p \geq 2$.

We complete the proof by demonstrating that $||v(z)||_{H^s(\mathbb{R})} < M(z)$. We start with the integral formulation of (2.5):

$$v(z,t) = \mathcal{S}(z)\{v_0\} - \int_0^z \mathcal{S}(z - z')\{\langle C \rangle\{v(z')\}\} dz' \tag{2.14}$$

so that, for $s \geq 1$,

$$
\begin{aligned}
||v(z)||_{H^s(\mathbb{R})} &\le ||\mathcal{S}(z)\{v_0\}||_{H^s(\mathbb{R})} + \int_0^z ||\mathcal{S}(z-z')\{\langle C\rangle\{v(z')\}\}||_{H^s(\mathbb{R})}dz' \\
&= ||v_0||_{H^s(\mathbb{R})} + \int_0^z ||\langle C\rangle\{v(z')\}||_{H^s(\mathbb{R})}dz' \\
&\le ||v_0||_{H^s(\mathbb{R})} + \int_0^z ||v(z')||^3_{H^s(\mathbb{R})}dz' \\
&\le ||v_0||_{H^s(\mathbb{R})} + \int_0^z ||v(z')||^2_{L^\infty(\mathbb{R})}||v(z')||_{H^s(\mathbb{R})}dz' \\
&\le ||v_0||_{H^s(\mathbb{R})} + M'\int_0^z ||v(z')||_{H^s(\mathbb{R})}dz'.
\end{aligned}
$$

Gronwall's inequality gives that $||v(z)||_{H^s(\mathbb{R})}$ is bounded on $[0, z)$, and this, in combination with the local well posedness result, gives global existence [30]. □

2.3. Averaging theorem. For this section it is most convenient to rescale $z \to \frac{z}{\epsilon}$ in the transformed and averaged equations (2.3) and (2.5) so that

$$
i\frac{\partial v^\epsilon}{\partial z} + \alpha_2\frac{\partial^2 v^\epsilon}{\partial t^2} + C\left(\frac{z}{\epsilon}\right)\{v^\epsilon\} = 0 \tag{2.15}
$$

and

$$
i\frac{\partial v}{\partial z} + \alpha_2\frac{\partial^2 v}{\partial t^2} + \langle C\rangle\{v\} = 0. \tag{2.16}
$$

The validity of the averaging procedure is addressed in the following theorem.

THEOREM 2.3. *Let $v(z,t) \in C^0([0,z^*], H^s(\mathbb{R}))$ be a solution of* (2.16) *on the time interval $[0,z^*]$ for any $z^* > 0$ and $s > \frac{7}{2}$. Then for ϵ sufficiently small there exists $v^\epsilon(z,t)$ a solution of* (2.15) *with initial data $v(0,t)$ such that*
$||v^\epsilon - v||_{L^\infty([0,\frac{z^*}{\epsilon}],H^{s-3}(\mathbb{R}))} < C\epsilon$.

Remark. We note that since $u = \mathcal{L}\{v^\epsilon\}$, the standard averaging result

$$
||u - \mathcal{L}\{v\}||_{L^\infty([0,\frac{z^*}{\epsilon}],H^{s-3}(\mathbb{R}))} < \epsilon
$$

follows immediately by isometry.

Proof. The proof is similar in spirit to classical averaging results in finite dimensions [32] and follows closely the method of [38]. We first split $C(\frac{z}{\epsilon})\{v\}$ into its average and varying components:

$$
C\left(\frac{z}{\epsilon}\right)\{v\} = \langle C\rangle\{v\} + \mathcal{R}\left(\frac{z}{\epsilon}\right)\{v\},
$$

where

$$
\hat{\mathcal{R}}\left(\frac{z}{\epsilon}\right)\{v\} = \int_{\mathbb{R}^3}\delta(k-k_1+k_2-k_3)\mathcal{A}\left(\frac{z}{\epsilon},k,k_1,k_2,k_3\right)\hat{v}(z,k_1)\bar{\hat{v}}(z,k_2)\hat{v}(z,k_3)dk_1dk_2dk_3
$$

and

$$
\mathcal{A}\left(\frac{z}{\epsilon},k,k_1,k_2,k_3\right) = \Theta\left(\frac{z}{\epsilon},k,k_1,k_2,k_3\right) - \Theta_l(k,k_1,k_2,k_3).
$$

The function $\mathcal{A}$ is bounded in its spatial variables, uniformly in z and ϵ. Now consider

$$\mathcal{B}_\epsilon(z,k,k_1,k_2,k_3) = \int_0^z \mathcal{A}\left(\frac{\tau}{\epsilon},k,k_1,k_2,k_3\right)d\tau = \epsilon\int_0^{\frac{z}{\epsilon}} \mathcal{A}(\tau',k,k_1,k_2,k_3)d\tau',$$

where $\tau' = \frac{\tau}{\epsilon}$. Since the integrand is 1-periodic with zero mean, we may write

$$\epsilon\int_0^{\frac{z}{\epsilon}} \mathcal{A}(\tau',k,k_1,k_2,k_3)d\tau' = \epsilon\int_0^{z''} \mathcal{A}(\tau',k,k_1,k_2,k_3)d\tau',$$

where $z'' = \frac{z}{\epsilon} - [\frac{z}{\epsilon}] \in [0,1)$, with $[\cdot]$ denoting the greatest integer function. Thus

$$||\mathcal{B}_\epsilon||_{L^\infty(\mathbb{R}^5)} \leq \epsilon\int_0^{z''} ||\mathcal{A}||_{L^\infty(\mathbb{R}^5)} \leq \epsilon z''||\mathcal{A}||_{L^\infty(\mathbb{R}^5)} \leq M\epsilon,$$

where M is independent of z.

We define the local average $\tilde{v} = v + v_1$, where

$$\hat{v}_1(z,k) = i\int_{\mathbb{R}^3} \delta(k-k_1+k_2-k_3)\mathcal{B}_\epsilon(z,k,k_1,k_2,k_3)\hat{v}(z,k_1)\bar{\hat{v}}(z,k_2)\hat{v}(z,k_3)dk_1dk_2dk_3.$$

By direct estimation

$$\begin{aligned}|\hat{v}_1(z,k)| &= \left|i\int_{\mathbb{R}^3} \delta(k-k_1+k_2-k_3)\mathcal{B}_\epsilon(z,k,k_1,k_2,k_3)\hat{v}(z,k_1)\bar{\hat{v}}(z,k_2)\hat{v}(z,k_3)dk_1dk_2dk_3\right| \\ &\leq ||\mathcal{B}_\epsilon||_{L^\infty(\mathbb{R}^5)}\left|\int_{\mathbb{R}^3} \delta(k-k_1+k_2-k_3)|\hat{v}(z,k_1)||\bar{\hat{v}}(z,k_2)||\hat{v}(z,k_3)|dk_1dk_2dk_3\right|.\end{aligned}$$

Repeating the arguments for regularity of $\langle C\rangle$ in section 2.1, we have that for every $\sigma > \frac{1}{2}$,

$$\|v_1(z)\|_{H^\sigma} \leq M\epsilon\|v(z)\|_{H^\sigma}^3.$$

Moreover, by the energy estimate in the well posedness theorem, Theorem 2.2, the bound is uniform in z for $s \geq 1$:

$$\sup_{0\leq z\leq z^*} ||v_1(z)||_{H^\sigma} \leq M\epsilon \sup_{0\leq z\leq z^*} \|v(z)\|_{H^\sigma}^3 \leq M'\epsilon.$$

We directly compute

$$i\frac{\partial\hat{v}_1}{\partial z} = \hat{\mathcal{R}}' - \hat{\mathcal{R}},$$

where

$$\hat{\mathcal{R}}' = i\int_{\mathbb{R}^3} \delta(k-k_1+k_2-k_3)\mathcal{B}_\epsilon(z,k,k_1,k_2,k_3)\partial_z\{\hat{v}(z,k_1)\bar{\hat{v}}(z,k_2)\hat{v}(z,k_3)\}dk_1dk_2dk_3.$$

Using equation (2.5), we estimate

$$\|\mathcal{R}'\|_{H^{s-3}} \leq M\epsilon.$$

The local average $\tilde{v}$ satisfies

$$i\frac{\partial\tilde{v}}{\partial z} + \alpha_2\frac{\partial^2\tilde{v}}{\partial t^2} + C\left(\frac{z}{\epsilon}\right)\{\tilde{v}\} = \mathcal{R}'',$$

where

$$\mathcal{R}'' = C\left(\frac{z}{\epsilon}\right)\{\tilde{v}\} - C\left(\frac{z}{\epsilon}\right)\{v\} + \mathcal{R}' + \alpha_2 \frac{\partial^2 v_1}{\partial t^2}.$$

By continuity of $C(\frac{z}{\epsilon})\{\cdot\}$,

$$\left\| C\left(\frac{z}{\epsilon}\right)\{\tilde{v}\} - C\left(\frac{z}{\epsilon}\right)\{v\} \right\|_{H^s} \leq M\epsilon$$

for all $s > \frac{1}{2}$, and so

$$\|\mathcal{R}''\|_{H^{s-3}} \leq M\epsilon.$$

Finally, we consider

$$f_\epsilon = v^\epsilon - \tilde{v},$$

which satisfies

$$i\frac{\partial f_\epsilon}{\partial z} + \alpha_2 \frac{\partial^2 f_\epsilon}{\partial t^2} + C\left(\frac{z}{\epsilon}\right)\{\tilde{v} + f_\epsilon\} - C\left(\frac{z}{\epsilon}\right)\{\tilde{v}\} = -\mathcal{R}''. \tag{2.17}$$

Again by continuity of $C(\frac{z}{\epsilon})\{\cdot\}$,

$$\left\| C\left(\frac{z}{\epsilon}\right)\{\tilde{v} + f_\epsilon\} - C\left(\frac{z}{\epsilon}\right)\{\tilde{v}\} \right\|_{H^{s-3}} \leq M\|f_\epsilon\|_{H^{s-3}}.$$

Writing (2.17) in Fourier space, multiplying the equation by $(1+|k|^2)^{s-3}\hat{f}_\epsilon$, its conjugate by $(1+|k|^2)^{s-3}f_\epsilon$, subtracting, and integrating over k, one obtains

$$\frac{\partial}{\partial z}\|f_\epsilon\|^2_{H^{s-3}} \leq M\epsilon\|f_\epsilon\|^2_{H^{s-3}} + M\|f_\epsilon\|^4_{H^{s-3}}.$$

Since $\|f_\epsilon\|^2_{H^{s-3}} < M$, we can write the estimate

$$\frac{\partial}{\partial z}\|f_\epsilon\|^2_{H^{s-3}} \leq M^2\epsilon + M^2\|f_\epsilon\|^2_{H^{s-3}}.$$

Now using the fact that $\|f_\epsilon(0)\|^2_{H^{s-3}} = 0$ and applying Gronwall's inequality we have

$$\|f_\epsilon(z)\|_{H^{s-3}} \leq e^{M^2\epsilon z}M^2\epsilon \leq e^K M^2\epsilon,$$

where K is a time-independent constant for $z \sim \mathcal{O}(\frac{1}{\epsilon})$, so

$$\sup_{0\leq z\leq \frac{z^*}{\epsilon}} \|f_\epsilon\|_{H^{s-3}} \leq M'\epsilon.$$

Overall, we have

$$\begin{aligned}\sup_{0\leq z\leq \frac{z^*}{\epsilon}} \|v^\epsilon - v\|_{H^{s-3}} &\leq \sup_{0\leq z\leq \frac{z^*}{\epsilon}} \|v^\epsilon - \tilde{v}\|_{H^{s-3}} + \sup_{0\leq z\leq \frac{z^*}{\epsilon}} \|\tilde{v} - v\|_{H^{s-3}} \\ &= \sup_{0\leq z\leq \frac{z^*}{\epsilon}} \|f_\epsilon\|_{H^{s-3}} + \sup_{0\leq z\leq \frac{z^*}{\epsilon}} \|v_1(z)\|_{H^{s-3}} \leq M\epsilon. \qquad \square\end{aligned}$$

3. Existence of a ground state. Here we show that for the cases $\alpha_2 > 0$ and $\alpha_3 = 0$ the averaged Hamiltonian possesses a minimizer in the class of admissible functions $\mathcal{A}_\lambda = \{v : \int_{\mathbb{R}} |v|^2 = \lambda, \int_{\mathbb{R}} |v_t|^2 < \infty\}$. We adapt an argument first established for the NLS equation [6] and later adapted for the case of second order DM [38]. We first present properties of the Hamiltonian that are essential to the minimization argument.

3.1. Properties of $\langle H \rangle$.

3.1.1. $\inf_{\mathcal{A}_\lambda} \langle H \rangle\{v\} < 0$.

Proof. We first assume that the 1-periodic dispersion maps $d_i(z)$, $i = 2, 3$, are piecewise constant and of the following form, which is standard in optical communications:

$$\tilde{d}_i(z) = \begin{cases} \tilde{D}_i & \text{if } z \in [0, \theta) \text{ or } z \in [1-\theta, 1), \\ -\tilde{D}_i & \text{if } z \in [\theta, 1-\theta) \ . \end{cases}$$

For these dispersion profiles we define the map strength parameters s_i by $s_i = \theta \tilde{D}_i$. The Hamiltonian can be written as

$$\langle H \rangle\{v\} = \alpha_2 \int_{\mathbb{R}} |v_t|^2 dt$$

$$- \ \frac{1}{2} \int_0^1 \int_{\mathbb{R}} \int_{\mathbb{R}^4} e^{i(k_1 - k_2 + k_3 - k_4)t} e^{-i\Delta_2 \int_0^z d_2(z') + i\Delta_3 \int_0^z d_3(z') dz'}$$

$$\times \ \hat{v}(k_1)\bar{\hat{v}}(k_2)\hat{v}(k_3)\bar{\hat{v}}(k_4) dk_1 dk_2 dk_3 dk_4 dt dz,$$

where $\Delta_2 = k_1^2 - k_2^2 + k_3^2 - k_4^2$ and $\Delta_3 = k_1^3 - k_2^3 + k_3^3 - k_4^3$. Performing the integration in z yields

$$\langle H \rangle\{v\} = \alpha_2 \int_{\mathbb{R}} |v_t|^2 dt$$

$$- \ \frac{1}{2} \int_{\mathbb{R}} \int_{\mathbb{R}^4} e^{i(k_1 - k_2 + k_3 - k_4)t} \Big(\frac{\theta \sin(s_2\Delta_2 + s_3\Delta_3)}{s_2\Delta_2 + s_3\Delta_3} \Big)$$

$$\times \ \hat{v}(k_1)\bar{\hat{v}}(k_2)\hat{v}(k_3)\bar{\hat{v}}(k_4) dk_1 dk_2 dk_3 dk_4 dt.$$

Let v be an arbitrary element of $\mathcal{A}_\lambda$, and consider the rescaled function

$$v_\gamma(t) = \gamma^{\frac{1}{2}} v\big(\gamma t\big),$$

which is also an element of $\mathcal{A}_\lambda$. A scaling property of the Fourier transform gives that

$$\hat{v}_\gamma(k) = \gamma^{\frac{-1}{2}} \hat{v} \left(\frac{k}{\gamma} \right),$$

and the chain rule yields

$$\frac{\partial v_\gamma(t)}{\partial t} = \gamma^{\frac{3}{2}} \frac{\partial v(t')}{\partial t'},$$

where $t' = \gamma t$. Substituting v_γ into the Hamiltonian, we have

$$F(\gamma) = \langle H \rangle \{v_\gamma\} = \alpha_2 \int_{\mathbb{R}} \left| \gamma^{\frac{3}{2}} \frac{\partial v(t')}{\partial t'} \right|^2 dt$$

$$- \frac{1}{2\gamma^2} \int_{\mathbb{R}} \int_{\mathbb{R}^4} e^{i(k_1 - k_2 + k_3 - k_4)t} \left(\frac{\theta \sin(s_2 \Delta_2 + s_3 \Delta_3)}{s_2 \Delta_2 + s_3 \Delta_3} \right)$$

$$\times \hat{v}\left(\frac{k_1}{\gamma}\right) \bar{\hat{v}}\left(\frac{k_2}{\gamma}\right) \hat{v}\left(\frac{k_3}{\gamma}\right) \bar{\hat{v}}\left(\frac{k_4}{\gamma}\right) dk_1 dk_2 dk_3 dk_4 dt$$

$$= \gamma^2 \alpha_2 \int_{\mathbb{R}} \left| \frac{\partial v(t')}{\partial t'} \right|^2 dt'$$

$$- \frac{1}{2\gamma^3} \int_{\mathbb{R}} \int_{\mathbb{R}^4} e^{i(\frac{k_1}{\gamma} - \frac{k_2}{\gamma} + \frac{k_3}{\gamma} - \frac{k_4}{\gamma})t'} \left(\frac{\theta \sin(s_2 \gamma^2 \tilde{\Delta}_2 + s_3 \gamma^3 \tilde{\Delta}_3)}{s_2 \gamma^2 \tilde{\Delta}_2 + s_3 \gamma^3 \tilde{\Delta}_3} \right)$$

$$\times \hat{v}\left(\frac{k_1}{\gamma}\right) \bar{\hat{v}}\left(\frac{k_2}{\gamma}\right) \hat{v}\left(\frac{k_3}{\gamma}\right) \bar{\hat{v}}\left(\frac{k_4}{\gamma}\right) dk_1 dk_2 dk_3 dk_4 dt',$$

where

$$\tilde{\Delta}_2 = \left(\frac{k_1}{\gamma}\right)^2 - \left(\frac{k_2}{\gamma}\right)^2 + \left(\frac{k_3}{\gamma}\right)^2 - \left(\frac{k_4}{\gamma}\right)^2$$

and

$$\tilde{\Delta}_3 = \left(\frac{k_1}{\gamma}\right)^3 - \left(\frac{k_2}{\gamma}\right)^3 + \left(\frac{k_3}{\gamma}\right)^3 - \left(\frac{k_4}{\gamma}\right)^3.$$

Making the change of variable $k_j' = \frac{k_j}{\gamma}$, we have

$$F(\gamma) = C\gamma^2$$

$$- \frac{\gamma}{2} \int_{\mathbb{R}} \int_{\mathbb{R}^4} e^{i(k_1' - k_2' + k_3' - k_4')t'} \left(\frac{\theta \sin(s_2 \gamma^2 \Delta_2' + s_3 \gamma^3 \Delta_3')}{s_2 \gamma^2 \Delta_2' + s_3 \gamma^3 \Delta_3'} \right)$$

$$\times \hat{v}(k_1') \bar{\hat{v}}(k_2') \hat{v}(k_3') \bar{\hat{v}}(k_4') dk_1' dk_2' dk_3' dk_4' dt'$$

$$= C\gamma^2 - \gamma G(v; \gamma, s_2, s_3, \theta),$$

where $\Delta_2' = (k_1')^2 - (k_2')^2 + (k_3')^2 - (k_4')^2$, $\Delta_3' = (k_1')^3 - (k_2')^3 + (k_3')^3 - (k_4')^3$, and G is a functional of v also depending on γ, s_2, s_3, and θ. At this stage, we can see that as $s_j \to 0$ and $\theta \to 1$ we recover the exact scaling of the integrable NLS Hamiltonian

evaluated at v_γ. To show that this Hamiltonian can be made negative, we first note that by continuity of the kernel

$$K(\gamma,\theta,s_j,\Delta_j') = \Big(\frac{\theta\sin(s_2\gamma^2\Delta_2' + s_3\gamma^3\Delta_3')}{s_2\gamma^2\Delta_2' + s_3\gamma^3\Delta_3'}\Big)$$

in γ, $F(0)=0$. Moreover, differentiating the functional F in γ yields

$$F'(\gamma) = 2C\gamma - \gamma G'(v;\gamma,s_2,s_3,\theta) - G(v;\gamma,s_2,s_3,\theta).$$

To compute G', we differentiate under the integral sign and apply the chain rule. This gives $G'(v;0,s_2,s_3,\theta)=0$ so that

$$F'(0) = -G(v;0,s_2,s_3,\theta) = -\frac{1}{2}\int_{\mathbb{R}}\int_{\mathbb{R}^4} e^{i(k_1'-k_2'+k_3'-k_4')t'}$$

$$\times\ \hat v(k_1')\bar{\hat v}(k_2')\hat v(k_3')\bar{\hat v}(k_4')dk_1'dk_2'dk_3'dk_4'dt'$$

$$= -\frac{1}{2}\int_{\mathbb{R}} |v(t')|^4 dt' < 0.$$

Thus for γ small enough the Hamiltonian is negative. □

3.1.2. $\langle H\rangle\{v\}$ is subadditive . If $I_\lambda = \inf_{v\in\mathcal{A}_\lambda}\langle H\rangle\{v\}$, then $I_{\lambda_1+\lambda_2} < I_{\lambda_1} + I_{\lambda_2}$.

Claim. For $\theta>1$, $I_{\theta\lambda} < \theta I_\lambda$

Proof of claim.

$$I_{\theta\lambda} = \inf_{v\in\mathcal{A}_{\theta\lambda}}\langle H\rangle\{v\}$$
$$= \inf_{w\in\mathcal{A}_\lambda}\langle H\rangle\{\sqrt{\theta}w\}$$

since

$$||w||_{L^2}^2 = \lambda \Rightarrow ||\sqrt{\theta}w||_{L^2}^2 = \theta\lambda.$$

But

$$\langle H\rangle(\sqrt{\theta}w) = \alpha_2\int_{\mathbb{R}}|(\sqrt{\theta}w)_t|^2dt - \frac{1}{2}\int_0^1\int_{\mathbb{R}}|\mathcal{L}(z)\{\sqrt{\theta}w\}|^4dtdz$$

$$= \theta\alpha_2\int_{\mathbb{R}}|w_t|^2dt - \frac{\theta^2}{2}\int_0^1\int_{\mathbb{R}}|\mathcal{L}(z)\{w\}|^4dtdz$$

$$< \theta\big(\alpha_2\int_{\mathbb{R}}|w_t|^2dt - \frac{1}{2}\int_0^1\int_{\mathbb{R}}\int_{\mathbb{R}}|w|^4dtdz\big)$$

for $\theta>1$. So

$$I_{\theta\lambda} = \inf_{w\in\mathcal{A}_\lambda}\langle H\rangle\{\sqrt{\theta}w\} < \theta\inf_{w\in\mathcal{A}_\lambda}\langle H\rangle\{w\} = \theta I_\lambda.$$ □

Proof of subadditivity. If we set $\lambda_1 = \alpha\lambda_2$ with $\alpha<1$, we have

$$I_{\lambda_1+\lambda_2} = I_{\alpha\lambda_2+\lambda_2} < (\alpha+1)I_{\lambda_2} = \alpha I_{(\alpha^{-1}\lambda_1)} + I_{\lambda_2} < I_{\lambda_1} + I_{\lambda_2}.$$

3.1.3. Localization of minimizing sequences. In the minimization proof, we used the fact that for a minimizing sequence $v_k(t) \in H^1(\mathbb{R})$ there exists a subsequence $v_{k_m}(t)$ which remains localized. That is, for any $\epsilon > 0$ there exists an $R > 0$ such that

$$\int_{-R}^{+R} |w_m(t)|^2 dt > \lambda - \epsilon,$$

where $w_m(t) = v_{k_m}(t - t_m)$ and $\lambda = \int_{\mathbb{R}} |w_m(t)|^2 dt$. To prove this result, we apply a version of Lions's concentration-compactness lemma [17, 38].

LEMMA 3.1. *If $u_m \in H^1(\mathbb{R})$ is a bounded sequence with $||u_m||_{L^2} = \lambda$, then there exists a subsequence u_{m_k} for which one of the following properties hold:*

1. *(localization) There exists a sequence t_k such that for any $\epsilon > 0$ there exists $R > 0$ and*

$$\int_{t_k-R}^{t_k+R} |u_{m_k}|^2 dx \geq \lambda - \epsilon.$$

2. *(vanishing) For any $R > 0$*

$$\lim_{k\to\infty} \sup_{y\in\mathbb{R}} \int_{y-R}^{y+R} |u_{m_k}|^2 dx \to 0.$$

3. *(splitting) There exists $0 < \gamma < \lambda$ such that for any $\epsilon > 0$ there exist k_0 and two sequences v_k, w_k with compact support so that for $k \geq k_0$*

$$||v_k||_{H^1} + ||w_k||_{H^1} \leq 4 \sup_{k\in\mathbf{N}} ||u_{m_k}||_{H^1}, \tag{3.1}$$

$$||u_{m_k} - (v_k + w_k)||_{L^2} \leq 2\epsilon, \tag{3.2}$$

$$|||v_k||_{L^2} - \gamma| \leq \epsilon \;\; |||v_k||_{L^2} - (\lambda - \gamma)| \leq \epsilon, \tag{3.3}$$

$$\left|\left|\frac{\partial v_k}{\partial x}\right|\right|_{L^2} + \left|\left|\frac{\partial w_k}{\partial x}\right|\right|_{L^2} \leq \left|\left|\frac{\partial u_{m_k}}{\partial x}\right|\right|_{L^2} + \epsilon \tag{3.4}$$

and $\operatorname{dist}(\operatorname{supp}(v_k), \operatorname{supp}(w_k)) > 2\epsilon^{-1}$.

Thus, for the minimization problem there exists a localized subsequence of the minimizing sequence if vanishing and splitting can be ruled out.

We first rule out vanishing. Let v_k be a minimizing sequence for $\langle H\rangle\{v\}$, and assume that a subsequence v_{m_k} vanishes. Since v_k is a minimizing sequence, for some k we have

$$\alpha_2 \int_{\mathbb{R}} \left|\frac{\partial v_k}{\partial t}\right|^2 dt - \frac{1}{2}\int_0^1 \int_{\mathbb{R}} |\mathcal{L}(z')\{v_k\}|^4 dt dz' < 0$$

so that

$$\int_0^1 \int_{\mathbb{R}} |\mathcal{L}(z')\{v_k\}|^4 dt dz' > 0.$$

Thus for some z^* we have

$$\int_{\mathbb{R}} |\mathcal{L}(z^*)\{v_k\}|^4 dt > 0.$$

Applying a lemma of Cazenave [6] for arbitrary $H^1(\mathbb{R})$ functions

$$\int_{\mathbb{R}} |u|^4 dt \leq C||u||^2_{H^1} \sup_{y\in\mathbb{R}} \int_{y-1}^{y+1} |u|^2 dt$$

gives that

$$\sup_{y\in\mathbb{R}} \int_{y-1}^{y+1} |\mathcal{L}(z^*)\{v_k\}|^2 dt > 0. \tag{3.5}$$

Now we relate $\sup_{y\in\mathbb{R}} \int_{y-1}^{y+1} |\mathcal{L}(z^*)\{v_k\}|^2 dt$ to $\sup_{y\in\mathbb{R}} \int_{y-1}^{y+1} |v_k|^2 dt$ with the following localization lemma, which is similar to the lemma of [38].

LEMMA 3.2. *Consider the following linear dispersive equation:*

$$iu_z + \tilde{d}_2(z)u_{tt} + i\tilde{d}_3(z)u_{ttt} = 0 \tag{3.6}$$

with $u \in H^1(\mathbb{R})$, $||u||_{L^2(\mathbb{R})} = 1$, and $\tilde{d}_i(z)$ piecewise constant. Let $u_n(t,z)$ be a sequence of solutions of (3.6)*, and define*

$$\epsilon_n(z) = \sup_{y\in\mathbb{R}} \int_{y-1}^{y+1} |u_n(t,z)|^2.$$

If $u_n(t,0)$ is vanishing initial data ($\lim_{n\to\infty} \epsilon_n(0) = 0$) with the constraint $||u_n||_{L^2(\mathbb{R})} = 1$, then the sequence of the solutions $u_n(t,z)$ is also vanishing ($\lim_{n\to\infty} \epsilon_n(z) = 0$).

Proof. Let $\chi_m(t)$ be a smooth approximation to the characteristic function on the interval $[-m,m]$, with the property that $|\partial_t \chi_m| < \frac{C}{m}$, $\chi_m(t) = 1$ if $|t| \leq 1$ and $\chi(t) = 0$ if $|t| \geq m$. Multiplying (3.6) by $\bar{u}\chi_m(t)$, its conjugate by $u\chi_m(t)$, subtracting, and integrating over t yield

$$\frac{d}{dz}\int_{\mathbb{R}} \chi_m |u|^2 dt = -2\tilde{d}_2(z)\mathrm{Im}\int_{\mathbb{R}} \chi_m \bar{u}u_{tt}dt - 2\tilde{d}_3(z)Re\int_{\mathbb{R}} \chi_m \bar{u}u_{ttt}dt.$$

Now

$$\mathrm{Im}\int_{\mathbb{R}} \chi_m \bar{u}u_{tt}dt = -\mathrm{Im}\int_{\mathbb{R}} (\chi_m \bar{u}_t + \chi'_m \bar{u})u_t dt = -\mathrm{Im}\int_{\mathbb{R}} \chi'_m \bar{u}u_t dt$$

and

$$\mathrm{Re}\int_{\mathbb{R}} \chi_m \bar{u}u_{ttt}dt = -\mathrm{Re}\int_{\mathbb{R}} (\chi_m \bar{u}_t + \chi'_m \bar{u})u_{tt}dt = -\mathrm{Re}\int_{\mathbb{R}} \chi_m \bar{u}_t u_{tt}dt - \mathrm{Re}\int_{\mathbb{R}} \chi'_m \bar{u}u_{tt}dt$$

$$= \frac{1}{2}\int_{\mathbb{R}} \frac{d|u_t|^2}{dt}\chi_m + \mathrm{Re}\int_{\mathbb{R}} (\chi''_m \bar{u} + \chi'_m \bar{u}_t)u_t dt = \frac{3}{2}\int_{\mathbb{R}} |u_t|^2 \chi'_m dt + \mathrm{Re}\int_{\mathbb{R}} \chi''_m \bar{u}u_t dt.$$

Overall,

$$\frac{d}{dz}\int_{\mathbb{R}} \chi_m |u|^2 dt = 2\tilde{d}_2(z)\mathrm{Im}\int_{\mathbb{R}} \chi'_m \bar{u}u_t dt - 2\tilde{d}_3(z)\left(\frac{3}{2}\int_{\mathbb{R}} |u_t|^2\chi'_m dt + \mathrm{Re}\int_{\mathbb{R}} \chi''_m \bar{u}u_t dt\right),$$

and integrating from 0 to z gives

$$\int_{\mathbb{R}} \chi_m |u(z)|^2 dt = \int_{\mathbb{R}} \chi_m |u(0)|^2 dt + \int_0^z \left(2\tilde{d}_2(z')\mathrm{Im}\int_{\mathbb{R}} \chi'_m \bar{u}u_t dt + 2\tilde{d}_3(z')\left(\frac{3}{2}\int_{\mathbb{R}} |u_t|^2\chi'_m dt + \mathrm{Re}\int_{\mathbb{R}} \chi''_m \bar{u}u_t dt\right)\right) dz'$$

$$\leq \int_{\mathbb{R}} \chi_m |u(0)|^2 dt + C_m(||u||_{H^1}, ||\chi_m'||_{L^\infty}, ||\chi_m''||_{L^\infty}, ||\tilde{d}_j||_{L^\infty}),$$

where $C_m \to 0$ as $m \to \infty$.

Let $u_n(t,z)$ denote a sequence of solutions of (3.6) with vanishing initial data, i.e.,

$$\epsilon_n(0) = \sup_{y\in\mathbb{R}} \int_{y-1}^{y+1} |u_n(t,0)|^2 \to 0.$$

If $\epsilon_n(z) < \epsilon_n(0)$, then we are done, so let $\epsilon_n(z) > \epsilon_n(0)$. Choosing $\chi_{m_n}(* - t_n)$ such that it is centered with respect to $u_n(z,t)$, we have

$$\int_{\mathbb{R}} \chi_{m_n} |u_n(t,z)|^2 dt \geq \epsilon_n(z)$$

and also

$$\int_{\mathbb{R}} \chi_{m_n} |u_n(t,0)|^2 dt \leq 2m_n \epsilon_n(0),$$

and taking the limit $m_n \to \infty$ with $m_n \sim \sqrt{\frac{1}{\epsilon_n(0)}}$ gives the result. □

Returning to (3.5) and applying the contrapositive of the localization lemma, we have

$$\sup_{y\in\mathbb{R}} \int_{y-1}^{y+1} |v_k|^2 dt > 0,$$

contradicting the assumption that a subsequence v_{m_k} vanishes.

To rule out splitting, it is enough to show that

$$\langle H\rangle\{v_{m_k}\} > \langle H\rangle\{w_k\} + \langle H\rangle\{u_k\} + \alpha(\epsilon),$$

where $\alpha(\epsilon)$ is independent of k and goes to 0 as $\epsilon \to 0$, as this causes $\langle H\rangle\{v_{m_k}\}$ to violate subadditivity. We directly evaluate $\langle H\rangle\{v_{m_k}\}$:

$$\langle H\rangle\{v_{m_k}\} = \alpha_2 \int_{\mathbb{R}} \left| \frac{\partial(u_k + w_k + h_k)}{\partial t} \right|^2 dt$$

$$- \frac{1}{2} \int_0^1 \int_{\mathbb{R}} |\mathcal{L}\{u_k + w_k + h_k\}|^4 dt dz,$$

where

$$||h_k||_{L^2}^2 < \epsilon,$$

and we have suppressed the notation $\mathcal{L}(z)$.
Expanding the terms, this can be rewritten as

$$\langle H\rangle\{v_{m_k}\} = \langle H\rangle\{u_k\} + \langle H\rangle\{w_k\}$$

$$+ 2\alpha_2 \mathrm{Re} \int_{\mathbb{R}} (\partial_t \overline{u}_k \partial_t w_k + \partial_t \overline{u}_k \partial_t h_k + \partial_t \overline{w}_k \partial_t h_k + |\partial_t h_k|^2) dt$$

$$-\mathrm{Re}\int_0^1\int_{\mathbb{R}}\big(|\mathcal{L}\{u_k+w_k\}|^2|\mathcal{L}\{h_k\}|^2+\frac{1}{2}|\mathcal{L}\{h_k\}|^4$$
$$+\ 2|\mathcal{L}\{u_k+w_k\}|^2(\mathcal{L}\{u_k+w_k\})(\mathcal{L}\{h_k\})+(\mathcal{L}\{u_k+w_k\})^2(\overline{\mathcal{L}\{h_k\}})^2$$
$$+\ 2\mathcal{L}\{u_k+w_k\}|\mathcal{L}\{h_k\}|^2\overline{\mathcal{L}h_k}\big)\ dtdz$$

$$+\ \frac{1}{2}\int_0^1\int_{\mathbb{R}}\big(2|\mathcal{L}\{u_k\}|^2|\mathcal{L}\{w_k\}|^2+2|\mathcal{L}\{u_k\}|^2\mathcal{L}\{u_k\}\mathcal{L}\{\overline{w}_k\}$$
$$+\ (\mathcal{L}\{u_k\})^2(\mathcal{L}\{\overline{w}_k\})^2+2|\mathcal{L}\{w_k\}|^2\mathcal{L}\{u_k\}\mathcal{L}\{\overline{w}_k\}\big)\ dtdz.$$

We proceed exactly as in [38]. The terms

$$2\alpha_2\mathrm{Re}\int_{\mathbb{R}}(\partial_t\overline{u}_k\partial_t w_k+\partial_t\overline{u}_k\partial_t h_k+\partial_t\overline{w}_k\partial_t h_k+|\partial_t h_k|^2)dt$$

can be estimated from below by $-C_1\epsilon$, with C_1 depending only on λ and α_2. The terms

$$\mathrm{Re}\int_0^1\int_{\mathbb{R}}\big(2|\mathcal{L}\{u_k+w_k\}|^2|\mathcal{L}\{h_k\}|^2+\frac{1}{2}|\mathcal{L}\{h_k\}|^4$$
$$+\ 2|\mathcal{L}\{u_k+w_k\}|^2(\mathcal{L}\{u_k+w_k\})(\mathcal{L}\{h_k\})+(\mathcal{L}\{u_k+w_k\})^2(\overline{\mathcal{L}\{h_k\}})^2$$
$$+\ 2\mathcal{L}\{u_k+w_k\}|\mathcal{L}\{h_k\}|^2\overline{\mathcal{L}h_k}\big)\ dtdz$$

are all estimated by Holder's inequality and the Sobolev inequality

$$\int_0^1\int_{\mathbb{R}}|\mathcal{L}(z)\{v\}|^4dtdz\le M||v(t)||^3_{L^2(\mathbb{R})}||v_t(t)||_{L^2(\mathbb{R})},$$

yielding a lower bound of the form $-C_2(\epsilon)$. The remaining terms

$$\int_0^1\int_{\mathbb{R}}\big(2|\mathcal{L}\{u_k\}|^2|\mathcal{L}\{w_k\}|^2+2|\mathcal{L}\{u_k\}|^2\mathcal{L}\{u_k\}\mathcal{L}\{\overline{w}_k\}$$
$$+\ (\mathcal{L}\{u_k\})^2(\mathcal{L}\{\overline{w}_k\})^2+2|\mathcal{L}\{w_k\}|^2\mathcal{L}\{u_k\}\mathcal{L}\{\overline{w}_k\}\big)dtdz$$

are estimated using the boundedness of $H^1(\mathbb{R})$ solutions of linear Schrödinger equations in $L^\infty(\mathbb{R})$ and the following lemma, which is a straightforward consequence of the localization lemma, Lemma 3.2:

In the notation of the concentration-compactness lemma the following estimates hold:

$$\int_{|t-t_k|\le t_c}|\mathcal{L}\{w_k\}|^2dt\le C\epsilon,$$
$$\int_{|t-t_k|\ge t_c}|\mathcal{L}\{u_k\}|^2dt\le C\epsilon,$$

where $t_c=\frac{t_1+t_2}{2}$.

Overall, we have

$$\langle H\rangle\{v_{m_k}\}>\langle H\rangle\{w_k\}+\langle H\rangle\{u_k\}+\alpha(\epsilon),$$

where $\alpha(\epsilon)$ is independent of k and goes to 0 as $\epsilon\to 0$. Thus splitting causes $\langle H\rangle$ to violate subadditivity, a contradiction.

3.2. Minimization theorem.

THEOREM 3.3. *Let $\alpha_2 > 0$ and $\alpha_3 = 0$. Then there exists a solution to the following constrained minimization problem:*

Minimize

$$\langle H\rangle\{v\} = \alpha_2 \int_{\mathbb{R}} |v_t|^2 dt - \frac{1}{2}\int_0^1 \int_{\mathbb{R}} |\mathcal{L}(z')\{v\}|^4 dt dz'$$

over the set of admissible functions

$$A_\lambda = \left\{ v \in H^1(\mathbb{R}), \int_{\mathbb{R}} |v|^2 = \lambda \right\}.$$

Moreover, every minimizing sequence has a subsequence which converges strongly in $H^1(\mathbb{R})$.

Remark. We note that the constraint $\int_{\mathbb{R}} |v|^2 = \lambda$ is quite natural, as the L^2 norm of the initial data is preserved by solutions of the Euler–Lagrange equation (2.5). Posing the problem in this way is also critical to the proof of the stability of the ground state which is given in a later section.

Proof. We follow the arguments of [6, 38]. The idea is to first show strong convergence in $L^2(\mathbb{R})$ by using Lions's concentration-compactness principle. This involves using structural properties of the Hamiltonian to rule out possible loss of compactness. Strong convergence in $L^2(\mathbb{R})$, along with an appropriate Sobolev inequality, implies convergence of the quartic term in the Hamiltonian. These results, in combination with lower semicontinuity of the $H^1(\mathbb{R})$ norm, give the existence of a minimizer. We show a posteriori that all minimizing sequences have a subsequence which converges strongly in $H^1(\mathbb{R})$.

We first argue that $I_\lambda > -\infty$. To prove the lower bound, we use the Sobolev inequality [4]

$$||\mathcal{L}v||_{L^4}^4 \le C||\mathcal{L}v_t||_{L^2}||\mathcal{L}v||_{L^2}^3 = C||v_t||_{L^2}||v||_{L^2}^3 = C\lambda^{3/2}||v_t||_{L^2}.$$

Integrating the inequality over z' gives

$$\int_0^1 \int_{-\infty}^{+\infty} |\mathcal{L}(z)v|^4 dt dz \le C\lambda^{3/2}||v_t||_{L^2}.$$

Thus

$$\langle H\rangle(v) \ge ||v_t||_{L^2}^2 - C\lambda^{3/2}||v_t||_{L^2} = \left(||v_t||_{L^2}^2 - \frac{C\lambda^{3/2}}{2}\right)^2 - \frac{C^2\lambda^3}{4} > -\infty$$

for all $v \in H^1(\mathbb{R})$. Taking the infimum over $v \in A_\lambda$ gives the desired result.

Let v_k be a minimizing sequence for $\langle H\rangle(v)$. By the previous inequality, $||v_k||_{H^1}$ must be bounded. By Alaoglu's theorem, there exists a weakly converging subsequence in $H^1(\mathbb{R})$, v_{k_m}. We will prove strong convergence of v_{k_m} to a minimizer in $H^1(\mathbb{R})$ and first establish strong convergence in $L^2(\mathbb{R})$.

From previous analysis, we conclude that the minimizing sequence remains localized as $m \to \infty$. That is, for any $\epsilon > 0$ there exists an $R > 0$ such that

$$\int_{-R}^{+R} |w_m(t)|^2 dt > \lambda - \epsilon, \tag{3.7}$$

where $w_m(t) = v_{k_m}(t - t_m)$. Now $w_m \rightharpoonup w^*$ for some $w^* \in H^1(\mathbb{R})$. For any $R > 0$, the embedding $H^1(\mathbb{R}) \hookrightarrow L^2([-R, R])$ is compact, and we have

$$\int_{-R}^{R} |w^*|^2 dt = \lim_{m\to\infty} \int_{-R}^{+R} |w_m|^2 dt.$$

Together with (3.7), this implies

$$\int_{-\infty}^{+\infty} |w^*|^2 dt > \lambda - \epsilon \text{ for any } \epsilon > 0,$$

and therefore

$$\int_{-\infty}^{+\infty} |w^*|^2 dt = \lambda.$$

This norm convergence, along with weak convergence in $L^2(\mathbb{R})$, gives strong convergence in $L^2(\mathbb{R})$.

Since w_m converges weakly to w^* and the Sobolev norm $||*||_{H^1(\mathbb{R})}$ is weakly lower semicontinuous, we have

$$||w^*||_{H^1(\mathbb{R})} \leq \liminf_{m\to\infty} ||w_m||_{H^1(\mathbb{R})},$$

which together with $w_m \to w^* \in L^2(\mathbb{R})$ implies that

$$||\partial_t w_m||_{L^2(\mathbb{R})} \leq \liminf_{m\to\infty} ||\partial_t w_m||_{L^2(\mathbb{R})}. \tag{3.8}$$

Now for any u^*, $u_m \in H^1(\mathbb{R})$, the Sobolev inequality gives

$$\int_{-\infty}^{+\infty} |u_m - u^*|^4 dt \leq C \int_{-\infty}^{+\infty} |\partial_t u_m - \partial_t u^*|^2 dt \left(\int_{-\infty}^{+\infty} |u_m - u^*|^2 dt \right)^{3/2}$$
$$\leq C \left(\int_{-\infty}^{+\infty} |u_m - u^*|^2 dt \right)^{3/2}.$$

It follows that if $u_m \to u^*$ in $L^2(\mathbb{R})$,

$$\int_{-\infty}^{+\infty} |u_m - u^*|^4 dt \to 0.$$

Applying the same argument to $\mathcal{L}(z)w_m$ and $\mathcal{L}(z)w^*$, we establish that

$$\mathcal{L}(z)w_m \to \mathcal{L}(z)w^* \text{in } L^4(\mathbb{R}),$$

and so

$$||\mathcal{L}(z)w^*||_{L^4(\mathbb{R})} = \lim_{m\to\infty} ||\mathcal{L}(z)w_m||_{L^4(\mathbb{R}).} \tag{3.9}$$

Combining (3.8) and (3.9),

$$\langle H \rangle(w^*) \leq \liminf_{m\to\infty} \langle H \rangle(w_m),$$

which can only happen if

$$\langle H\rangle(w^*) = \lim_{m\to\infty} \langle H\rangle(w_m), \tag{3.10}$$

so the weak limit w^* is a minimizer. Furthermore, by (3.10)

$$||\partial_t w^*||_{L^2(\mathbb{R})} = \lim_{m\to\infty} ||\partial_t w_m||_{L^2(\mathbb{R})}.$$

Together with weak convergence, this implies strong convergence of $\partial_t w_m$ in $L^2(\mathbb{R})$, so $w_m \to w^*$ strongly in $H^1(\mathbb{R})$. □

4. Properties of the ground state.

4.1. Regularity. The minimizer for the constrained minimization problem is also a weak solution to the Euler–Lagrange equation

$$-\omega v + \alpha_2 v_{tt} + \langle C\rangle\{v\} = 0. \tag{4.1}$$

If we rewrite (4.1) in the form

$$v_{tt} = \frac{1}{\alpha_2}(\omega v - \langle C\rangle\{v\}) = f(v),$$

where $f(v) \in H^1(\mathbb{R})$ by continuity of $\langle C\rangle\{v\}$, we may use standard elliptic regularity theory [15] to conclude that $v \in H^3(\mathbb{R})$. Again, by continuity of $\langle C\rangle\{v\}$,

$$\omega v - \alpha_2 v_{tt} = \langle C\rangle\{v\} \in H^3(\mathbb{R}),$$

forcing $v \in H^5(\mathbb{R})$. We repeat this procedure indefinitely, obtaining $v \in H^s(\mathbb{R})$, for every $s \geq 1$, or $v \in C^\infty(\mathbb{R})$. We note, however, that the $H^s(\mathbb{R})$ norm of v may depend on α_2.

4.2. Stability. It is clear that the minimizer is not unique, as any translation $v(\cdot + \tau_0)$, $\tau_0 \in \mathbb{R}$, or rotation $e^{i\theta}v$, $\theta \in \mathbb{R}$, of the minimizer is also a solution of the constrained minimization problem. Also, it is not known that translations and rotations give *all* possible minimizers. From now on we consider the class of ground state solutions $\mathcal{S}_\lambda = \{v_g \in A_\lambda, \langle H\rangle(v_g) = I_\lambda\}$. Using the strong convergence of minimizing sequences and conservation laws for (2.5), one can show that the minimizer is stable in the following orbital sense.

THEOREM 4.1. *Let $\mathcal{S}_\lambda$ be the set of ground states $\mathcal{S}_\lambda = \{v_g \in A_\lambda, \langle H\rangle(v_g) = I_\lambda\}$. For any $\epsilon > 0$, there exists a $\delta > 0$ such that if $\inf_{\mathcal{S}_\lambda} ||v - v_g||_{H^1} \leq \delta$, then the solutions of (2.5) corresponding to initial data v and v_g, denoted $v(z)$ and $v_g(z)$, satisfy $\sup_z \inf_{\mathcal{S}_\lambda} ||v(z) - v_g(z)||_{H^1} \leq \epsilon$.*

Proof. We argue by contradiction. Let $v_k(0)$ be a sequence of initial conditions such that $\inf_{\mathcal{S}_\lambda} ||v_k(0) - v_g||_{H^1} \to 0$, and assume that $v_k(z)$ and $v_g(z)$ satisfy $\sup_z \inf_{\mathcal{S}_\lambda} ||v_k(z) - v_g(z)||_{H^1} \geq \epsilon$ for some $\epsilon > 0$. For definiteness, let z_n be the first time that $\inf_{\mathcal{S}_\lambda} ||v_k(z) - v_g(z)||_{H^1} = \epsilon$. By conservation of the L^2 norm and of the Hamiltonian, we have

$$\int_{\mathbb{R}} |v_k(z_n)|^2 dt = \int_{\mathbb{R}} |v_k(0)|^2 dt,$$
$$\langle H\rangle\{v_k(z_n)\} = \langle H\rangle\{v_k(0)\}.$$

By the assumption on $v_k(0)$ and continuity of $\langle H\rangle$, we have

$$\int_{\mathbb{R}} |v_k(z_n)|^2 dt = \int_{\mathbb{R}} |v_k(0)|^2 dt \to \lambda,$$
$$\langle H\rangle\{v_k(z_n)\} = \langle H\rangle\{v_k(0)\} \to \langle H\rangle\{v_g\}.$$

By choosing, for example, $w_k = \frac{\lambda^{\frac{1}{2}} v_k(z_n)}{(\int_{\mathbb{R}} |v_k(z_n)|^2 dt)^{\frac{1}{2}}}$, let w_k be a sequence of H^1 functions such that

$$||w_k - v_k(z_n)||_{H^1} \to 0$$

and $\int_{\mathbb{R}} |w_k(z)|^2 dt = \lambda$. By continuity of $\langle H\rangle$, w_k is a minimizing sequence and must have a subsequence w_{m_k} which converges to a ground state. But

$$||v_k(z_n) - v_g(z)||_{H^1} \le ||v_k(z_n) - w_{m_k}||_{H^1} + ||w_{m_k} - v_g(z)||_{H^1},$$

and taking the infimum over $\mathcal{S}_\lambda$ gives

$$\epsilon = \inf_{\mathcal{S}_\lambda} ||v_k(z_n) - v_g(z)||_{H^1} \le ||v_k(z_n) - w_{m_k}||_{H^1} + \inf_{\mathcal{S}_\lambda} ||w_{m_k} - v_g(z)||_{H^1} \to 0,$$

a contradiction. Thus the class of ground states must be orbitally stable. ∎

5. Numerical studies.

5.1. Solution of the eigenvalue problem. In Fourier space, the Euler–Lagrange equation (4.1) becomes

$$-\omega\hat{v} - \alpha_2 k^2 \hat{v} + \langle \hat{C}\rangle\{v\} = 0. \tag{5.1}$$

We propose the following explicit iteration scheme:

$$-\omega_n \hat{v}_{n+1} - \alpha_2 k^2 \hat{v}_{n+1} + \langle \hat{C}\rangle\{v_n\} = 0$$

so that

$$\hat{v}_{n+1} = \frac{\langle \hat{C}\rangle\{v_n\}}{\omega_n + \alpha_2 k^2}.$$

If we multiply (5.1) by $\bar{\hat{v}}(k)$ and integrate over k, we can derive a formula for ω_{n+1}:

$$\omega_{n+1} = \frac{\int_{\mathbb{R}} \bar{\hat{v}}_{n+1} \langle \hat{C}\rangle\{v_{n+1}\} dk - \alpha_2 \int_{\mathbb{R}} k^2 |\hat{v}_{n+1}|^2 dk}{\int_{\mathbb{R}} |\hat{v}_{n+1}|^2 dk}.$$

This idea also suggests a definition for a relaxation factor to hasten convergence [21, 31]:

$$s_{n+1} = \frac{\omega_{n+1} \int_{\mathbb{R}} |\hat{v}_{n+1}|^2 dk}{\int_{\mathbb{R}} \bar{\hat{v}}_{n+1} \langle \hat{C}\rangle\{v_{n+1}\} dk - \alpha_2 \int_{\mathbb{R}} k^2 |\hat{v}_{n+1}|^2 dk}.$$

The factor s_n can be used to compensate the nonlinearity, and as the scheme converges, $s_n \to 1$. We also note that convergence depends strongly on the initial guess,

which is typically taken to be Gaussian. Incorporating the relaxation factor, the overall scheme becomes

$$\hat{v}_{n+1} = (s_n)^p \frac{\langle \hat{C} \rangle \{v_n\}}{\omega_n + \alpha_2 k^2},$$

$$\omega_{n+1} = \frac{\int_{\mathbb{R}} \bar{\hat{v}}_{n+1} \langle \hat{C} \rangle \{v_{n+1}\} dk - \alpha_2 \int_{\mathbb{R}} k^2 |\hat{v}_{n+1}|^2 dk}{\int_{\mathbb{R}} |\hat{v}_{n+1}|^2 dk},$$

$$s_{n+1} = \frac{\omega_{n+1} \int_{\mathbb{R}} |\hat{v}_{n+1}|^2 dk}{\int_{\mathbb{R}} \bar{\hat{v}}_{n+1} \langle \hat{C} \rangle \{v_{n+1}\} dk - \alpha_2 \int_{\mathbb{R}} k^2 |\hat{v}_{n+1}|^2 dk},$$

where we use $p = 1.5$.

To confine our search for minimizers to a fixed level set of $L^2(\mathbb{R})$, we normalize the result of each iteration so that its energy is that of the initial guess,

$$||v_n||^2_{L^2}(\mathbb{R}) = ||v_0||^2_{L^2}(\mathbb{R}) = \lambda.$$

In practice the algorithm is as follows:

1. Choose an initial profile $\hat{v}_0$, typically Gaussian.
2. Compute ω_0 with $\hat{v}_0$.
3. Compute s_0 with $\hat{v}_0$ and ω_0.
4. Compute $\hat{v}_1$ using the scheme.
5. Rescale $\hat{v}_1$ so that $||\hat{v}_1||^2_{L^2}(\mathbb{R}) = ||\hat{v}_0||^2_{L^2}(\mathbb{R})$.
6. Compute ω_1 and s_1.
7. Repeat 3, 4, and 5 until desired accuracy is reached.

5.2. Solution of evolution equations. Both the full evolution equation (1.2) and averaged equation (2.5) can easily be solved with a version of the well-known Fourier split-step scheme, which applies to a wide class of NLS-type equations. Given an evolution equation of the form

$$iu_z + \mathcal{L}\{u\} + \mathcal{N}\{u\} = 0,$$

where $\mathcal{L}$ is a self-adjoint operator on a Hilbert space and $\mathcal{N}$ is a continuous nonlinear operator, the solution may be written formally as

$$u(z,t) = u(0,t) e^{i \int_0^z (\mathcal{L}(s)\{u\} + \mathcal{N}(s)\{u\}) ds}. \tag{5.2}$$

We consider the evolution for a small propagation step Δz so that we may approximate (5.2) using a formal Taylor expansion:

$$\begin{aligned} u(\Delta z, t) &= u(0,t) e^{i \int_0^{\Delta z} (\mathcal{L}(s)\{u\} + \mathcal{N}(s)\{u\}) ds} \\ &\approx u(0,t) e^{i \int_0^{\Delta z/2} \mathcal{L}(s)\{u\}} e^{i \int_0^{\Delta z} \mathcal{N}(s)\{u\} ds} e^{i \int_0^{\Delta z/2} \mathcal{L}(s)\{u\} ds}, \end{aligned}$$

with a local error on the order of $(\Delta z)^3$. Performing the approximation for $\mathcal{O}(\frac{1}{\Delta z})$ time steps gives a global error on the order of $(\Delta z)^2$. Formally, $e^{i \int_0^z \mathcal{L}(s)\{u\} ds}$ is the semigroup for the linear evolution

$$iu_z + \mathcal{L}\{u\} = 0$$

and can be computed explicitly in Fourier domain. Also,

$$e^{i\int_0^z \mathcal{N}(s)\{u\}ds}$$

is the solution operator for the evolution equation

$$iu_z + \mathcal{N}\{u\} = 0$$

and can be computed either by a standard ODE method such as fourth order Runge–Kutta or, in special cases, by using conservation laws for the equation.

The overall scheme becomes the following:

1. Choose an initial profile $u(0,t)$ and compute its Fourier transform with the FFT.
2. In Fourier space, evolve the linear dispersive operator for $\Delta z/2$.
3. Evolve the nonlinear operator for Δz.
4. Evolve the linear dispersive operator for $\Delta z/2$.
5. Repeat 3 and 4 until the final propagation step.
6. Evolve the linear dispersive operator for $\Delta z/2$ and compute the inverse Fourier transform with the IFFT.

5.3. Existence of nearly periodic solutions. Combining the averaging and minimization results, we see that there exist stationary solutions for (2.5) that evolve nearly periodically for (1.2). The ground state $v_g(t)$ for the variational problem corresponds to a standing wave solution for (2.5), $v(z,t) = \exp(i\omega z)v_g(t)$. The averaging theorem gives that

$$\left|\left|u - \mathcal{L}\left(\frac{z}{\epsilon}\right)\{v(z,t)\}\right|\right|_{L^\infty([0,\frac{z^*}{\epsilon}],H^{s-3}(\mathbb{R}))} \leq \epsilon$$

so that $u(z,t)$ is nearly periodic on the scale of validity for the averaging theorem. Moreover, by well posedness of the averaged equation, the same result is true for initial data chosen close to the class of ground states $\inf_{\mathcal{S}_\lambda} ||v - v_g||_{H^1} \leq \epsilon$.

We define a *higher order dispersion managed soliton* to be an element from the class of ground states $\mathcal{S}_\lambda$ and demonstrate the existence of such solutions numerically. Figure 5.1 shows the shape of the ground state solution for the parameters

$$\tilde{d}_2(z') = \tilde{d}_3(z') = \begin{cases} 5.0 & \text{if } z' \in [0,.25) \text{ or } z' \in [.75,1.0), \\ -5.0 & \text{if } z' \in [.25,.75) \end{cases} \tag{5.3}$$

and $\alpha_2 = 1.0$, $\epsilon = 0.1$. The solution is computed on the time domain $[-30,30]$ with 2048 Fourier modes. The logarithm of the amplitude $|v(t)|$ is plotted versus time on the interval $[-20,20]$. One observes a nearly Gaussian central peak, along with many secondary peaks which decay rapidly. This is similar to the structure of the ground states observed for DM at second order [1, 28].

From the averaging theorem, one would expect the ground state to evolve nearly periodic for $z \sim \mathcal{O}(10)$. Figure 5.2 depicts the evolution of the maximum amplitude of the ground state for the corresponding full equation. The individual oscillations are due to the linear compensation of dispersion, and we observe that the evolution of the amplitude is, in fact, nearly periodic on z scales *much longer* than those predicted by the averaging theorem.

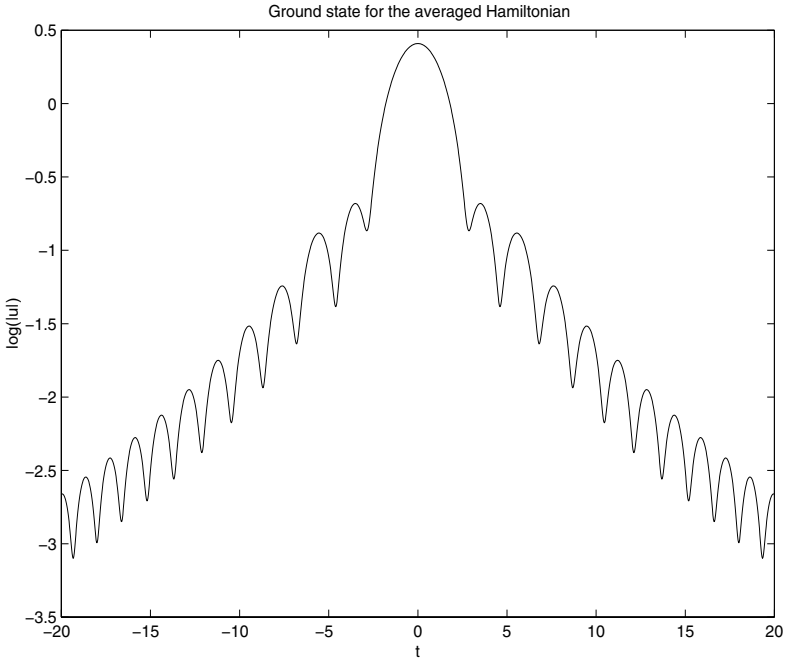


FIG. 5.1. *Higher order dispersion managed soliton.*

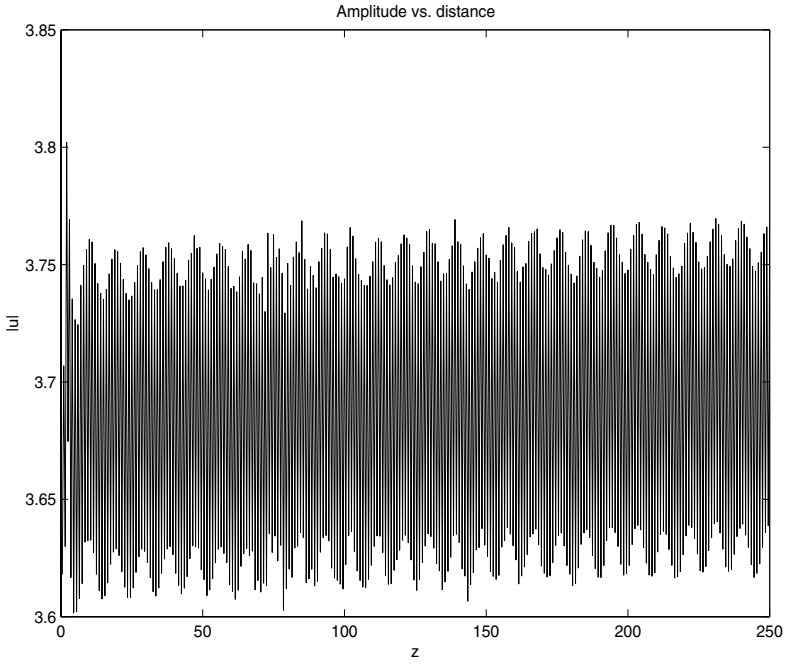


FIG. 5.2. *Nearly periodic evolution of the ground state.*

Acknowledgments. We are grateful to all of the reviewers, who provided many insightful suggestions and comments. We are particularly indebted to the reviewer who provided an alternate proof of Lemma 3.2, which was used in this text. We also would like to thank Ildar Gabitov for bringing this problem to our attention. Part of this work was done when V. Zharnitsky was visiting Program in Applied and Computational Mathematics at Princeton University. He would like to thank Ingrid Daubechies for her hospitality and for providing a stimulating research environment.

REFERENCES

[1] M. J. Ablowitz and G. Biondini, *Multiscale pulse dynamics in communication systems with strong dispersion management*, Opt. Lett., 23 (1998), pp. 1668–1670.

[2] M. J. Ablowitz and T. Hirooka, *Resonant nonlinear interactions in strongly dispersion-managed transmission systems*, Opt. Lett., 25 (2000), pp. 1750–1752.

[3] G. P. Agrawal, *Fiber-Optic Communication Systems*, 2nd ed., John Wiley and Sons, New York, 1997.

[4] H. Brezis, *Analyse fonctionnelle. Théorie et applications*, Masson, Paris, 1983.

[5] S. K. Burtsev and I. Gabitov, *Four-wave mixing in fiber links with dispersion management*, in Proceedings of the Second International Symposium on Physics and Applications of Optical Solitons in Fibers, Kyoto, Japan, 1997, pp. 261–265.

[6] T. Cazenave, *An Introduction to Nonlinear Schrödinger Equations*, UFRJ, Rio de Janeiro, Brazil, 1993.

[7] L. du Mouza, E. Seve, H. Mardoyn, S. Wabnitz, P. Sillard, and P. Nouchi, *High-order dispersion-managed solitons for dense wavelength-division multiplexed transmissions*, Opt. Lett., 26 (2001), pp. 1128–1130.

[8] F. Favre, D. Le Guen, M. L. Moulinard, M. Henry, and T. Georges, 320*Gbit/s soliton WDM transmission over* 1300 *km with* 100 *km dispersion compensated spans of standard fibre*, Elec. Lett., 33 (1997), pp. 2135–2136.

[9] F. Favre, D. Le Guen, and T. Georges, *Experimental evidence of pseudo-periodical soliton propagation in dispersion managed link*, Elec. Lett., 34 (1998), pp. 1868–1869.

[10] I. Gabitov, E. G. Shapiro, and S. K. Turitsyn, *Asymptotic breathing pulse in optical transmission systems with dispersion compensation*, Phys. Rev. E, 55 (1995), pp. 3624–3633.

[11] E. A. Golovchenko, V. J. Mazurczyk, D. G. Duff, and S. M. Abbott, *Four-wave mixing penalties in long-haul WDM transmission links*, IEEE Photon. Technol. Lett., 11 (1999), pp. 821–823.

[12] M. Kunze, *Bifurcation from the essential spectrum without sign condition on the nonlinearity*, Proc. Roy. Soc. Edinburgh, Sect. A, 131 (2001), pp. 927–943.

[13] M. Kunze, *On a variational problem with lack of compactness related to the Strichartz inequality*, Calc. Var. Partial Differential Equations, to appear.

[14] T. I. Lakoba, J. Yang, D. J. Kaup, and B. A. Malomed, *Conditions for stationary pulse propagation in the strong dispersion management regime*, Opt. Comm., 149 (1998), pp. 366–375.

[15] E. H. Lieb and M. Loss, *Analysis*, AMS, Providence, RI, 1997.

[16] C. Lin, H. Kogelnik, and L. G. Cohen, *Optical pulse equalization and low dispersion transmission in single-mode fibers in the* 1.3-1.7*mm spectral region*, Opt. Lett., 5 (1980), pp. 476–478.

[17] P. L. Lions, *The concentration-compactness principle in the calculus of variations. The locally compact case, part* 1, Ann. Inst. H. Poincaré Anal. Non Linéaire, 1 (1984), pp. 109–145.

[18] Y. Liu, *Challenging the Limits of Chromatic Dispersion*, http://lw.pennnet.com/Articles/Article_Display.cfm?Section=ARCHI&Subsection=Display&ARTICLE_ID=98944&KEYWORD=liu&p=13 (May 2001).

[19] F. Liu, J. Bennike, S. Dey, C. Rasmussen, H. Mikkelsen, P. Mamyshev, D. Gapontse, and V. Ivshin, 1.6 *Tbit/s (*40x42.7 *Gbit/s) transmission over* 3600 *km UltraWave*^{tm} *fiber with all-Raman amplified* 100 *km terrestrial spans using ETDM transmitter and receiver*, OFC 2002 Post Deadline Digest session FC7 1-3, Optical Fiber Communications, Anaheim, CA, 2002.

[20] P. M. Lushnikov, *Dispersion-managed soliton in optical fibers with zero average dispersion*, Opt. Lett., 25 (2000), pp. 1144–1146.

[21] P. M. Lushnikov, *Dispersion-managed soliton in a strong dispersion map limit*, Opt. Lett., 26, (2001), pp. 1535–1537.

[22] P. V. Mamyshev and N. A. Mamysheva, *Pulse-overlapped dispersion managed data transmission and intrachannel four-wave mixing*, Opt. Lett., 24 (1999), pp. 1454–1456.
[23] M. Manna and E. A. Golovchenko, *FWM resonances in dispersion slope-matched and non-zero-dispersion fiber maps: Impact on system performance*, Optical Fiber Communication Conference and Exhibition. Technical Digest, Anaheim, CA, 2001.
[24] J. Moeser, I. Gabitov, and C. K. R. T. Jones, *Pulse stabilization by high order dispersion management*, Opt. Lett., 27 (2002), pp. 2206–2208.
[25] L. F. Mollenauer, P. V. Mamyshev, J. Gripp, M. J. Neubelt, N. Mamysheva, L. Gruner-Nielsen, and T. Veng, *Demonstration of massive wavelength-division multiplexing over transoceanic distances by use of dispersion-managed solitons*, Opt. Lett., 25 (2000), pp. 704–706.
[26] M. Murakami, H. Maeda, and T. Imai, *Long-haul* 16x10 *Gb/s WDM transmission experiment using higher order fiber dispersion management technique*, IEEE Photon. Technol. Lett., 11 (1999), pp. 898–900.
[27] M. Nakazawa, H. Kubota, K. Suzuki, and E. Yamada, *Recent progress in soliton transmission technology*, Chaos, 10, (2000), pp. 486–514.
[28] J. H. B. Nijhof, N. J. Doran, W. Forysiak, and F. M. Knox, *Stable soliton–like propagation in dispersion managed systems with net anomalous, zero, and normal dispersion*, Elec. Lett., 33 (1997), pp. 1726–1728.
[29] J. H. B. Nijhof, N. J. Doran, W. Forysiak, and A. Berntson, *Energy enhancement of dispersion managed solitons and WDM*, Elec. Lett., 34 (1998), pp. 481–483.
[30] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[31] V. I. Petviashvili and O. A. Pokhotelov, *Solitary Waves in Plasmas and in the Atmosphere*, Gordon and Breach, Philadelphia, 1992.
[32] J. A. Sanders and F. Verhulst, *Averaging Methods in Nonlinear Dynamical Systems*, Springer-Verlag, New York, 1985.
[33] I. Segal, *Non-linear semi-groups*, Ann. of Math. (2), 78 (1963), pp. 339–364.
[34] N. J. Smith, N. J. Doran, F. M. Knox, and W. Forysiak, *Energy-scaling characteristics of solitons in strongly dispersion-managed fibers*, Opt. Lett., 21 (1996), pp. 1981–1983.
[35] W. A. Strauss, *Nonlinear Wave Equations*, AMS, Providence, RI, 1989.
[36] M. Suzuki, I. Morita, N. Edagawa, S. Yamamoto, H. Taga, and S. Akiba, *Reduction of Gordon-Haus timing jitter by periodic dispersion compensation in soliton transmission*, Elec. Lett., 31 (1995), pp. 2027–2030.
[37] S. K. Turitsyn, I. Gabitov, E. W. Laedke, V. K. Mezentsev, S. L. Musher, E. G. Shapiro, T. Schäfer, and K. H. Spatschek, *Variational approach to optical pulse propagation in dispersion compensated transmission systems*, Opt. Comm., 151 (1998), pp. 117–135.
[38] V. Zharnitsky, E. Grenier, C. K. R. T. Jones, and S. K. Turitsyn, *Stabilizing effects of dispersion management*, Phys. D, 152/153 (2001), pp. 794–817.

SIAM J. MATH. ANAL.
Vol. 35, No. 6, pp. 1512–1533

DIFFUSIVE SCALING LIMITS OF MUTUALLY INTERACTING PARTICLE SYSTEMS*

SHUI FENG†, ILIE GRIGORESCU‡, AND JEREMY QUASTEL§

Abstract. We prove the diffusive scaling limits of some interacting particle systems in random dynamical environments. The limits are identified as nonlinear parabolic systems, with coefficients given by equilibrium variational problems. Three related models are studied that correspond to different environments. All the models are of nongradient type, and one is nonreversible. The proofs involve techniques of entropy production estimates, the nongradient method and asymmetric tools, in particular a proof of the strong sector condition.

Key words. hydrodynamic limit, nongradient system

AMS subject classifications. 60F10, 60J75, 60K35, 82C26

DOI. 10.1137/S0036141002409520

1. Introduction. In this article we study the diffusive scaling limits of three models of random walks with simple exclusion on a multidimensional lattice subject to rapidly fluctuating jump rates determined by another system of similar walks. In each of the three models there are two types of particles which we denote by η and ξ. Each particle, independently of the others, waits a random, exponentially distributed length of time and then attempts to jump to a neighboring site. The interaction enters in two ways. If a particle attempts to jump to a site already occupied by a particle of the same type, the jump is suppressed. This hard core interaction between particles of the same type is called simple exclusion. The second interaction is through a speed change. The expected waiting time for a given particle depends on the local configuration of particles of the *other* type. The three models differ in the exact form of this *speed change*. It is more convenient to think of this in terms of the inverse of the expected waiting time, or the rate of jumping. In Model 1, the rate for an η particle to jump from a site x to a neighboring site y is $\gamma_1 + \xi_x + \xi_y$, where ξ_x and ξ_y are the numbers of ξ particles at x and y, while the ξ particles all jump at rate γ_2. Here γ_1 and γ_2 are two positive numbers. In other words, the ξ particles perform the symmetric simple exclusion process and the η particles perform a "simple exclusion in a symmetric simple exclusion environment." In Model 2, the rate for an η particle to jump from x to the nearest neighbor y is $\gamma_1 + \frac{1}{2}(\xi_x + \xi_y)$, and the rate for a ξ particle to jump is $\gamma_2 + (1 - \frac{1}{2}(\eta_x + \eta_y))$. Hence the two processes dynamically drive each other through the interdependence of the jump rates. Models 1 and 2 are in some sense warmups for Model 3, in which an η particle jumps from x to nearest neighbor y at rate $\gamma_1 + \xi_x$ and a ξ particle does the same at rate $\gamma_2 + 1 - \eta_x$.

Such models can be thought of as microscopic pursuit and evasion predator-prey models. In Models 2 and 3, for example, the η particles represent prey and the ξ

*Received by the editors June 12, 2002; accepted for publication May 23, 2003; published electronically March 11, 2004. This research was supported by the Natural Science and Engineering Research Council of Canada.

http://www.siam.org/journals/sima/35-6/40952.html

†Department of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada L8S 4K1 (shuifeng@mcmaster.ca).

‡Department of Mathematics, University of Miami, 1365 Memorial Drive, Ungar Building, Coral Gables, FL 33146-4250 (igrigore@math.miami.edu).

§Departments of Mathematics and Statistics, University of Toronto, Toronto, ON, Canada M5S 3G3 (quastel@math.toronto.edu).

particles represent predators. The predators jump fast until they find a prey and then slow down, while the prey jump slowly until they see a predator, at which time they speed up to run away. Very little work has been done on pursuit and evasion systems, as opposed to birth and death predator prey systems, where birth and death rates are functions of the other species. Both types of predator-prey systems are usually modelled by continuum equations—systems of partial differential equations—though particle systems are more realistic. There is quite a bit of work in the probability literature on systems of birth and death processes with random walks [1] and on speed change models with a scalar conservation law [11] but, as far as we know, no previous results on speed change (i.e., pursuit and evasion) systems. Parabolic systems have much more interesting behavior than scalar parabolic equations [9] and are more relevant to biology.

With respect to modelling in biology using parabolic systems, a question arises as to whether one should use divergence or nondivergence form. Our models shed some light on this issue: Microscopically they are in (discrete) nondivergence form, but macroscopically the bulk equations take divergence form.

Another motivation for these models comes from the theory of homogenization. Let us recall two well-known examples.

1. To each bond $x, x+e$ of the multidimensional integer lattice $\mathbf{Z}^d$ is associated an independent random variable $a(x, x+e) \geq \delta > 0$. We let $x(t)$ be a continuous time random walk on $\mathbf{Z}^d$ with generator $Lf(x) = \sum_{|e|=1} a(x, x+e)(f(x+e) - f(x))$, and we ask for the asymptotic behavior of $x_\epsilon(t) = \epsilon x(\epsilon^{-2}t)$. This is the reversible case, in which the rate of jumping from x to nearest neighbor $x+e$ is the same as the rate of jumping back, that is, $a(x, x+e) = a(x+e, x)$. The uniform measure is an invariant and reversible (unnormalized) measure, and by standard methods of homogenization one finds that the limiting process is a Brownian motion with covariance $\ell \cdot \bar{a}\ell = \inf_f \sum_e E[a(0,e)(e \cdot \ell + \tau_e f - f)^2]$, the infimum ranging over stationary processes.

2. To each site x of $\mathbf{Z}^d$ is associated an independent random variable $a(x) \geq \delta > 0$, and $y(t)$ is a continuous time random walk with generator $Lf(x) = \sum_{|e|=1} a(x)(f(x+e) - f(x))$. This is the nonreversible case, in which the rate of jumping from x, $a(x)$ depends on x alone. Here the (unnormalized) invariant measure gives mass $a^{-1}(x)/E[a^{-1}]$ to site x. The rescaled process $y_\epsilon(t) = \epsilon y(\epsilon^{-2}t)$ again converges to Brownian motion. The variance of $y_\epsilon(t)$ can be computed explicitly, and an application of the ergodic theorem tells us that the asymptotic variance in this case is $E[a^{-1}(0)]^{-1}$. Now in each of the two models, suppose that we replaced the static random field by one varying in time. Similar questions can be answered in the reversible case of example 1, but in the nonreversible case of example 2 little is known. The problem is that in the second case there is no invariant measure.

Interacting systems with two types of particles provide examples of dynamic random environments which can be analyzed. In particular, the speed change in our Model 3 is of the type of the second example.

The main results of the article are scaling limits for the diffusively rescaled density fields in the three models. The limits are coupled parabolic systems, with diffusion matrices which can be obtained from certain variational problems. We use the nongradient method (see [10], [13]) and its adaptation to the mean zero nonreversible setting [15]. The main idea is to consider the models as bounded perturbations of symmetric simple exclusion, and for this we have to assume $\gamma_1, \gamma_2 > 0$. The work [15] is unpublished. The only other case we know treating the nonreversible, nongradient

case using the strong-sector estimate to bound the asymmetry in terms of the symmetry is in [6], where a very interesting model related to vortex flow (see [7]) is studied. The method used is the relative entropy method, which requires certain a priori regularity for solutions of the hydrodynamic equation. However, to prove this one needs first some regularity of the diffusion coefficient as a function of the density, and this has not been obtained for the model in [6] at the present time. Hence the proof is not complete. For parabolic systems as considered in the present article, even some regularity of the coefficients would not help, as the needed regularity results for solutions are not available. Hence one is forced to use the method of [10], [13], [15]. Because of the lack of references we have provided a sketch of the argument, referring to the existing literature whenever possible.

2. The models. In each of the three models there are two types of particles which we call η particles and ξ particles. The particles perform symmetric nearest neighbor random walks on the multidimensional integer lattice $\mathbf{Z}^d$, with exclusion within their type. In other words, each particle waits an exponential amount of time, then attempts a jump to a neighboring site chosen with equal probabilities. The jump is only executed if the target site is free of a particle of the same type. If we start with at most one particle of each type at each site, it will stay so forever, so the state space of all three models is $X = (\{0,1\} \times \{0,1\})^{\mathbf{Z}^d}$. Configurations will be denoted (η, ξ), and for each $x \in \mathbf{Z}^d$, $\eta_x \in \{0,1\}$, and $\xi_x \in \{0,1\}$ denote the presence or absence of a particle at that site.

The interaction is through the expected length of the holding time, which will depend on the local environment. Let us introduce some notation. The operations $\eta \mapsto \eta^{x,y}$ and $\xi \mapsto \xi^{x,y}$ exchange the occupation numbers at the two sites x and y. More precisely, they are defined as $\eta_x^{x,y} = \eta_y$, $\eta_y^{x,y} = \eta_x$, and $\eta_z^{x,y} = \eta_z$ otherwise, and analogously for ξ. It is convenient to use the η and ξ lattice gradients acting on functions on X, which are given by

$$\nabla_{x,y}^{\eta} f(\eta,\xi) = f(\eta^{x,y},\xi) - f(\eta,\xi), \qquad \nabla_{x,y}^{\xi} f(\eta,\xi) = f(\eta,\xi^{x,y}) - f(\eta,\xi). \tag{2.1}$$

We can now describe the three models (from easiest to hardest).

- *Model* 1. The ξ particles attempt jumps to each neighbor at rate γ_2. An η particle at x attempts to jump to nearest neighbor y at rate $\gamma_1 + \frac{\xi_x+\xi_y}{2}$. The infinitesimal generator is

$$L^{(1)} f = \sum_{x\sim y} \left(\gamma_1 + \frac{\xi_x+\xi_y}{2}\right) \nabla_{x,y}^{\eta} f + \gamma_2 \nabla_{x,y}^{\xi} f. \tag{2.2}$$

  The sum is over ordered nearest neighbor pairs $x \sim y$.
- *Model* 2. A ξ particle at x attempts to jump to nearest neighbor y at rate $\gamma_2 + 1 - \frac{\eta_x+\eta_y}{2}$. An η particle at x attempts to jump to nearest neighbor y at rate $\gamma_1 + \frac{\xi_x+\xi_y}{2}$. The infinitesimal generator is

$$L^{(2)} f = \sum_{x\sim y} \left(\gamma_1 + \frac{\xi_x+\xi_y}{2}\right) \nabla_{x,y}^{\eta} f + \left(\gamma_2 + 1 - \frac{\eta_x+\eta_y}{2}\right) \nabla_{x,y}^{\xi} f. \tag{2.3}$$

- *Model* 3. A ξ particle at x attempts to jump to each nearest neighbor site at rate $\gamma_2 + 1 - \eta_x$. An η particle at x attempts to jump to each nearest neighbor site at rate $\gamma_1 + \xi_x$. The infinitesimal generator is

$$L^{(3)} f = \sum_{x\sim y} (\gamma_1 + \xi_x\eta_x(1-\eta_y)) \nabla_{x,y}^{\eta} f + (\gamma_2 + (1-\eta_x)\xi_x(1-\xi_y)) \nabla_{x,y}^{\xi} f. \tag{2.4}$$

We shall use the generic notation L for the infinitesimal generator of the three models, unless we need to differentiate between them (especially in section 5).

Models 1 and 2 are reversible with respect to the family of product (Bernoulli) measures $\pi_{u,v} = (m_u \times m_v)^{\otimes \mathbf{Z}_N^d}$, $u, v \in [0,1]$, where $m_u(1) = u$ and $m_u(0) = 1-u$. The corresponding Dirichlet forms

$$\mathcal{D}_{u,v}(f) = -E^{\pi_{u,v}}[fLf]$$

are given by

$$\frac{1}{2}\sum_{x\sim y} E^{\pi_{u,v}}\left[\left(\gamma_1 + \frac{\xi_x+\xi_y}{2}\right)(\nabla^\eta_{x,y}f)^2 + \gamma_2(\nabla^\xi_{x,y}f)^2\right] \qquad \text{(Model 1)}$$

and

$$\frac{1}{2}\sum_{x\sim y} E^{\pi_{u,v}}\left[\left(\gamma_1 + \frac{\xi_x+\xi_y}{2}\right)(\nabla^\eta_{x,y}f)^2 + \left(\gamma_2 + 1 - \frac{\eta_x+\eta_y}{2}\right)(\nabla^\xi_{x,y}f)^2\right] \text{ (Models 2,3).}$$

We learned about Model 3 from Donatis Surgailis, who also indicated the following key fact, which is easy to check.

PROPOSITION 2.1 (Surgailis). *The product measures $\pi_{u,v}$, $u, v \in [0,1]$ are invariant for $L^{(3)}$ in Model 3.*

However, $L^{(3)}$ is *not* reversible with respect to the $\pi_{u,v}$. The generator of Model 2 is nothing but the symmetric part of the generator in Model 3.

One could, of course, consider much more general speed change models, where the holding time of a particle is a general function of the local configuration. The basic problem then becomes one of finding the set of invariant measures, which is extremely hard in general.

On the other hand, one can start with a family of invariant measures and construct appropriate Dirichlet forms. This produces dynamics for which the measures are guaranteed to be reversible and invariant. However, dynamics for which we can determine a nice family of measures which are invariant but *not* reversible are rare, a fact underlying the importance of Model 3.

For each of the three models one can check that for $\gamma_1, \gamma_2 > 0$ the two particle densities are the only conserved quantities. A consequence is that on a box of side length ϵ^{-1} with periodic or reflecting boundary conditions, once we fix the number of η and the number of ξ particles, then the continuous time Markov chain $(\eta(\cdot), \xi(\cdot))$ is ergodic and the distribution converges to the uniform distribution on configurations with those numbers of particles.

We also have the obvious lower bound

$$\mathcal{D}(f) \geq \gamma \mathcal{D}^{(0)}(f), \quad \gamma \leq \gamma_1 \wedge \gamma_2,$$

for each of the three Dirichlet forms in terms of the Dirichlet form $\mathcal{D}^{(0)}$ of two independent copies of the symmetric simple exclusion process,

$$\mathcal{D}^{(0)}_{u,v}(f) = \frac{1}{2}\sum_{x\sim y} E^{\pi_{u,v}}[(\nabla^\eta_{x,y}f)^2 + (\nabla^\xi_{x,y}f)^2].$$

We can also rewrite the Dirichlet form as $\mathcal{D}^{(2)}(f) = \sum_{x\sim y} \mathcal{D}_{x,y}(f)$, where

$$D_{x,y}(f) = \frac{1}{2}E\left[\left(\gamma_1 + \frac{\xi_x+\xi_y}{2}\right)(\nabla^\eta_{x,y}f)^2 + \left(\gamma_2 + 1 - \frac{\eta_x+\eta_y}{2}\right)(\nabla^\xi_{x,y}f)^2\right]. \tag{2.5}$$

Since it is well known that on a box of side length ϵ^{-1} the spectral gap of symmetric simple exclusions is bounded below by some constant multiple of ϵ^2, we immediately obtain for our three Dirichlet forms the Poincaré inequalities $Var(f) \leq C\epsilon^{-2}\mathcal{D}(f)$ on boxes of side length ϵ^{-1}, uniformly in the density. For fixed $\gamma_1, \gamma_2 > 0$ we will prove diffusive scaling limits for the joint empirical densities of particles. The limits are coupled systems of parabolic partial differential equations. The diffusion matrices for limits of such systems cannot in general be expected to be elementary functions of the densities. However, we can obtain variational formulae for the diffusion matrices, and these can be used to show some structure of the equations.

This is made precise by the *hydrodynamic scaling limit*. To avoid technicalities we work on the torus $\mathbf{T}^d$ instead of $\mathbf{R}^d$, though it is known how to deal with infinite systems [3]. We are given functions $u_0(\mathbf{x})$ and $v_0(\mathbf{x})$ of $\mathbf{x} \in \mathbf{T}^d$ taking values in $[0,1]$. The small scaling parameter $\epsilon > 0$ represents the separation between macroscopic and microscopic pictures. To keep ourselves on the torus we assume that ϵ^{-1} is an integer. Macroscopic space and time variables $\mathbf{x} \in \mathbf{T}^d$ and $\mathbf{t} \geq 0$ are related to microscopic variables $x \in \mathbf{Z}^d/\epsilon^{-1}\mathbf{Z}^d$ and $t \geq 0$ by

$$x = \lfloor \epsilon^{-1}\mathbf{x} \rfloor, \qquad t = \epsilon^{-2}\mathbf{t}.$$

We assume that the initial distribution μ_0^ϵ of the process running on $\mathbf{Z}^d/\epsilon^{-1}\mathbf{Z}^d$ is such that the following law of large numbers holds: *As $\epsilon \to 0$, in μ_0^ϵ-probability, the empirical density fields $(\eta_{\lfloor \epsilon^{-1}\mathbf{x} \rfloor}, \xi_{\lfloor \epsilon^{-1}\mathbf{x} \rfloor})$ converge weakly to $(u_0(\mathbf{x}), v_0(\mathbf{x}))$*, where $u_0(\mathbf{x})$ and $v_0(\mathbf{x})$ are some nice functions on the torus. Consider $\hat{P}_\epsilon$, the distributions of

$$\mathbf{t} \longrightarrow (\eta_{\lfloor \epsilon^{-1}\mathbf{x} \rfloor}(\epsilon^{-2}\mathbf{t})d\mathbf{x}, \xi_{\lfloor \epsilon^{-1}\mathbf{x} \rfloor}(\epsilon^{-2}\mathbf{t})d\mathbf{x}), \tag{2.6}$$

seen as measures on $D([0,\infty); M(\mathbf{T}^d) \times M(\mathbf{T}^d))$, the Skorohod space of left-limit and right-continuous maps from $[0,\infty)$ into $M(\mathbf{T}^d) \times M(\mathbf{T}^d)$, the space of pairs of probability measures with the topology of weak convergence, indexed by the scaling parameter $\epsilon > 0$.

We shall denote by $L_l^{(i)}$, for $i = 0, 1, 2, 3$, the restrictions of the infinitesimal generators of the processes confined to a box Λ_l of size $l \in Z_+$ centered at the origin. For fixed numbers of particles m and n, we denote by $P^{n,m,l}$ the product Bernoulli measure π_ϱ conditional on the hyperplane $\sum_{x \in \Lambda_l} \zeta_x = (m,n) = \lfloor (2l+1)^d \varrho \rfloor$, where $\zeta_x = (\xi_x, \eta_x)$, $\varrho = (u,v) \in [0,1] \times [0,1]$.

Let $\mathcal{F}$ be the class of local functions f on the state space $\{0,1\}^{Z^d} \times \{0,1\}^{Z^d}$ satisfying the bound

$$E^{n,m,l}[fh] \leq C \sum_{|x-y|=1,\, |x|,|y| \leq l'} \mathcal{D}_{x,y}^{(0)}(h), \tag{2.7}$$

with a constant $C > 0$, uniformly over boxes of size $l \in Z_+$ for functions with finite support h (*local functions*). The integer $l' \leq l$ stands for the largest integer such that the box $\Lambda_{l'} + \mathrm{supp}(f)$ be included in Λ_l. In particular, mean-zero local functions like the gradients $\nabla\zeta$, the currents W_{0,e_i} and the fluctuations Lg for g local satisfy the property.

We shall see in (3.25) that, for any $\varrho = (u,v) \in [0,1] \times [0,1]$ and for $i = 0, 1, 2, 3$, we can define the equivalent seminorms

$$\langle f\,,\, f \rangle_{-1,\varrho}^{(i)} = \lim_{(n(2l+1)^{-d}\,,\, m(2l+1)^{-d}) \to \varrho} (2l)^{-d} E^{n,m,l}\left[\sum_{x \leq l'} \tau_x f\,,\, (-L_l^{(i)})^{-1}(\tau_x f) \right]. \tag{2.8}$$

If $\mathcal{N}$ is the null space corresponding to $i = 0$, we denote the completion of the quotient space $\mathcal{F}/\mathcal{N}$ by $\mathcal{H}^{(i)}_{-1,\varrho}$, a Hilbert space for the symmetric cases $i = 0$ and $i = 2$. The null space is the same for all i due to the fact that $L^{(3)}_{sym} = L^{(2)}$ and equivalence of the norms warranted by the *strong-sector condition* described in Lemma 2.5.

We need the compressibility matrix

$$\chi(\varrho) = \chi(u,v) = \begin{pmatrix} u(1-u)I_d & 0 \\ 0 & v(1-v)I_d \end{pmatrix}, \tag{2.9}$$

where I_d is the d-dimensional identity matrix.

THEOREM 2.2 (Model 1). *Assume $\gamma_1, \gamma_2 > 0$. Then $\hat{P}_\epsilon \Rightarrow \delta_{u,v}$, the Dirac mass on the trajectory $(u(\mathbf{t},\mathbf{x}), v(\mathbf{t},\mathbf{x}))d\mathbf{x}$, where (u,v) is the unique weak solution of*

$$\frac{\partial}{\partial t}\begin{pmatrix} u \\ v \end{pmatrix} = \nabla \begin{pmatrix} e(u,v) & 0 \\ 0 & \gamma_2 I_d \end{pmatrix} \nabla \begin{pmatrix} u \\ v \end{pmatrix}, \qquad \mathbf{x} \in \mathbf{T}^d,\ t \geq 0, \tag{2.10}$$

with $(u(0,\mathbf{x}), v(0,\mathbf{x})) = (u_0(\mathbf{x}), v_0(\mathbf{x}))$, satisfying $\int_0^T \int_{\mathbf{T}^d} [|\nabla u|^2 + |\nabla v|^2] d\mathbf{x} d\mathbf{t} < \infty$. The matrix $e(u,v)$ is continuous in u and v and is given by the variational formula for any $\mathbf{r} = (r_1, \ldots, r_d) \in \mathbf{R}^d$:

$$\begin{aligned} &\mathbf{r} e(u,v) \mathbf{r}' \\ &\quad = \frac{1}{2u(1-u)} \inf_{g \in \mathcal{F}} E^{\pi_{u,v}} \left[\sum_{i=1}^d \left(\gamma_1 + \frac{\xi_0 + \xi_{e_i}}{2} \right) (r_i(\eta_{e_i} - \eta_0) \right. \\ &\qquad \left. - \nabla^\eta_{0,e_i} \Omega_g)^2 + \gamma_2 (\nabla^\xi_{0,e_i} \Omega_g)^2 \right]. \end{aligned}$$

Here $\Omega_g = \sum_{x \in \mathbf{Z}^d} \tau_x g$ with τ_x the shift operator.

THEOREM 2.3 (Model 2). *Assume $\gamma_1, \gamma_2 > 0$. Then $\hat{P}_\epsilon$ are tight, and any limit point is supported on the set of weak solutions of*

$$\frac{\partial}{\partial t}\begin{pmatrix} u \\ v \end{pmatrix} = \nabla D^{(2)}(u,v) \nabla \begin{pmatrix} u \\ v \end{pmatrix},$$

where $\mathbf{x} \in \mathbf{T}^d$, $t \geq 0$, with $(u(0,\mathbf{x}), v(0,\mathbf{x})) = (u_0(\mathbf{x}), v_0(\mathbf{x}))$ satisfying $\int_0^T \int_{\mathbf{T}^d} [|\nabla u|^2 + |\nabla v|^2] d\mathbf{x} d\mathbf{t} < \infty$. The diffusion matrix $D^{(2)}(u,v)$ is continuous in u and v and is given by

$$\begin{aligned} D^{(2)}(u,v) &= \begin{pmatrix} (\gamma_1 + v) I_d & 0 \\ 0 & (\gamma_2 + 1 - u) I_d \end{pmatrix} \\ &\quad + \frac{1}{4} \left[B - (u(1-u)(\gamma_1 + v) + v(1-v)(\gamma_2 + 1 - u)) I_d \right] \chi^{-1}(u,v) \begin{pmatrix} I_d & I_d \\ I_d & I_d \end{pmatrix}, \end{aligned} \tag{2.11}$$

where for any $\mathbf{r} = (r_1, \ldots, r_d)$,

$$\begin{aligned} \mathbf{r} B \mathbf{r}' = \frac{1}{2} \inf_{g \in \mathcal{F}} E^{\pi_{u,v}} \left[\sum_{i=1}^d \left(\gamma_1 + \frac{\xi_0 + \xi_{e_i}}{2} \right) (r_i(\eta_{e_i} - \eta_0) - \nabla^\eta_{0,e_i} \Omega_g)^2 \right. \\ \left. + \left(\gamma_2 + 1 - \frac{\eta_0 + \eta_{e_i}}{2} \right) (r_i(\xi_{e_i} - \xi_0) - \nabla^\xi_{0,e_i} \Omega_g)^2 \right]. \end{aligned}$$

Before stating the third hydrodynamic limit, we need to recall that on hyperplanes $\sum \zeta = (n, m)$ for fixed nonnegative integers m and n, the generators $L_l^{(i)}$ are invertible.

THEOREM 2.4 (Model 3). *Assume $\gamma_1, \gamma_2 > 0$. Then $\hat{P}_\epsilon$ are tight, and any limit point is supported on the set of weak solutions of*

$$\frac{\partial}{\partial t}\begin{pmatrix} u \\ v \end{pmatrix} = \nabla D^{(3)}(u,v)\nabla \begin{pmatrix} u \\ v \end{pmatrix}, \qquad \mathbf{x} \in \mathbf{T}^d,\ t \geq 0, \tag{2.12}$$

with $(u(0,\mathbf{x}), v(0,\mathbf{x})) = (u_0(\mathbf{x}), v_0(\mathbf{x}))$ satisfying $\int_0^T \int_{\mathbf{T}^d} [|\nabla u|^2 + |\nabla v|^2] d\mathbf{x} d\mathbf{t} < \infty$, where $D^{(3)}(u,v)$ is a $2d \times 2d$ matrix valued function continuous in u and v given by

$$\begin{aligned} &\Big(D^{(3)}(u,v)\Big)^{-1}\chi(\varrho) \\ &\quad = \lim_{(n(2l+1)^{-d},\, m(2l+1)^{-d}) \to (u,v)} (2l)^{-d} E^{n,m,l}\left[\sum_{x \leq l'} \tau_x \nabla\zeta\ (-L_l^{(3)})^{-1}(\tau_x \nabla \zeta)\right]. \end{aligned} \tag{2.13}$$

Furthermore, there exist $2d \times 2d$ matrices Q and V, with V symmetric, such that

$$D^{(3)}(u,v)Q = D^{(2)}(u,v)V \tag{2.14}$$

and $Q_{sym} < V$ in the sense of quadratic forms.

A comment related to the asymmetric diffusion coefficient $D^{(3)}(u,v)$ is included at the end of section 5.

Remark 1 (on uniqueness). Uniqueness of the hydrodynamic equations for Models 2 and 3 is a hard problem and we have not pursued it here.

Remark 2 (on the degenerate case). If $\gamma_1 = \gamma_2 = 0$, then Models 2 and 3 are no longer ergodic. For example, any configuration in which every site where there is a ξ particle is also occupied by an η particle and there are η but no ξ particles in all nearest neighbor[ing] sites is an absorbing state for Model 3. We can construct such configurations which have macroscopic profiles, and since every state in our systems has bounded specific entropy, it follows that the diffusion coefficients simply vanish. It is an interesting question whether the scaling limit could hold after removing some bad configurations from the space, but we do not know how to answer this. On the other hand, if only one of γ_1 and γ_2 vanish the situation is not so bad. One can check, for example, in Model 2 that the spectral gap on a box of side length ϵ^{-1} is correct, say, if $\gamma_1 = 0$ but $\gamma_2 > 0$, but with a factor $Cv\epsilon^2$, where v is the density of ξ particles, and with a factor $C(1-u)\epsilon^2$ if $\gamma_1 > 0$ but $\gamma_2 = 0$. Analogous results hold for Model 1. In a similar way, one can check that the diffusion matrices of Models 2 and 3 dominate

$$\begin{pmatrix} C(\gamma_2)vI_d & 0 \\ 0 & \gamma_2 I_d \end{pmatrix} \tag{2.15}$$

if $\gamma_1 = 0$ and

$$\begin{pmatrix} \gamma_1 I_d & 0 \\ 0 & C(\gamma_1)(1-u)I_d \end{pmatrix} \tag{2.16}$$

if $\gamma_2 = 0$ for some $C(d) > 0$ for $d > 0$. For the rest of the article we concentrate exclusively on the case

$$\gamma_1, \gamma_2 \geq \gamma > 0.$$

Remark 3 (on the birth-death model). Let $a(\eta,\xi), b(\eta,\xi)$ be positive local functions, and

$$L_{reaction} f(\eta,\xi) = \sum_x a(\tau_x\eta,\tau_x\xi)(f(\eta^x,\xi) - b(\eta,\xi)) + d(\tau_x\eta,\tau_x\xi)(f(\eta,\xi^x) - f(\eta,\xi))$$

with $\eta_x^x = 1 - \eta_x$ and $\eta_y^x = \eta_y$ otherwise, and analogously for ξ. Let $L_\epsilon^{(i)} = \epsilon^{-2}L^{(i)} + L_{reaction}$, $i = 1,2,3$. The hydrodynamic limit is a nonlinear reaction-diffusion equation of the form

$$\frac{\partial}{\partial t}\begin{pmatrix} u \\ v \end{pmatrix} = \nabla D^{(i)}(u,v)\nabla \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} F(u,v) \\ G(u,v) \end{pmatrix},$$

where $F(u,v) = E^{\pi_{u,v}}[a(\eta,\xi)(1-2\eta_0)]$, $G(u,v) = E^{\pi_{u,v}}[b(\eta,\xi)(1-2\xi_0)]$. See [5] for details.

The method of proof for Models 1 and 2 which are reversible, nongradient systems is by now rather standard (in the sense that they have been worked out for the Ginzburg–Landau model [13] and the symmetric simple exclusion process [14]). These methods are all based on entropy and its rate of change. Fix $\varrho = (u,v) \in (0,1)\times(0,1)$ and let π_ϱ be a reference probability measure on the state space. If $\mu = f\pi_\varrho$ is any other probability measure on the state space we define its entropy as

$$H(f) = E^{\pi_\varrho}[f\log f].$$

If $\mu_t = f_t\pi_\varrho$ denotes the marginal distribution of our process with Dirichlet form $\mathcal{D}(f)$, then we have the general inequality

$$\frac{dH(f_t)}{dt} \le -\frac{1}{4}\mathcal{D}(f_t).$$

Changing to the macroscopic time scale $t = \epsilon^{-2}\mathbf{t}$ corresponds to multiplying the generator, or Dirichlet form, by a factor ϵ^{-2}. Hence the initial entropy bound

$$H(f_0) \le K(\log 4)\epsilon^{-d} \tag{2.17}$$

with K a constant independent of ϵ produces the bound

$$\int_0^\infty \mathcal{D}(f_\mathbf{t})d\mathbf{t} \le K(\log 4)\epsilon^{2-d}.$$

The $\log 4$ is just the maximum entropy per site in a model with 4 possible values at each site (the constant K will take care of any arbitrary pair $\varrho_0 = (u_0,v_0)$, but we can assume $u = v = 1/2$ for this purpose). Now if $\gamma_1,\gamma_2 \ge \gamma > 0$ for each of the models, the Dirichlet form of the process dominates $\gamma\mathcal{D}^{(0)}$, the Dirichlet form of the symmetric simple exclusion. Hence we have the *entropy production bound*

$$\int_0^\infty \mathcal{D}^{(0)}(f_\mathbf{t})d\mathbf{t} \le \gamma^{-1}(\log 4)\epsilon^{2-d}. \tag{2.18}$$

From this bound follow the key estimates for nongradient reversible systems. These will be described in section 3 with references to the original proofs.

Model 3 is of nongradient, nonreversible, mean-zero type. For such models, a method was developed in Xu's thesis [15], specifically applied to the mean-zero asymmetric simple exclusion process. We did not have access to [15] but relied on notes of Varadhan's lectures on this topic at the Fields Institute [14]. Our proof follows their ideas very closely. The main ingredient, which allows the extension of the standard reversible machinery in these types of nonreversible systems, is the following *strong-sector condition.*

LEMMA 2.5. *There exists a constant $C > 0$ such that for any $\pi = \pi_\varrho$, $\varrho \in [0,1] \times [0,1]$, $f, g \in \mathcal{F}$, and any of our three models,*

$$\left| \int f L g d\pi \right| \leq C \sqrt{\mathcal{D}(f)} \sqrt{\mathcal{D}(g)}. \tag{2.19}$$

Proof. For Models 1 and 2 the result is immediate from the reversibility. We prove it for Model 3 with $\gamma_1 = \gamma_2 = 0$. It then extends immediately to nonnegative γ_1, γ_2. We rewrite the generator as $L = \sum_{x \sim y} L_{x,y}$, where

$$L_{x,y} g(\eta, \xi) = \frac{1}{2}[(\xi_x + \xi_y)(g(\eta, \xi^{x,y}) - g(\eta, \xi)) + (\eta_x \xi_x + \eta_y \xi_y)(g(\eta^{x,y}, \xi) - g(\eta, \xi^{x,y}))].$$

Recall the Dirichlet form $\mathcal{D}(f) = \sum_{x \sim y} \mathcal{D}_{x,y}(f)$ from (2.5). We write $E[f L_{x,y} g] = A + B$, where

$$A = \frac{1}{2} E\left[(\xi_x + \xi_y)(g(\eta, \xi^{x,y}) - g(\eta, \xi)) f(\eta, \xi)\right], \tag{2.20}$$

$$B = \frac{1}{2} E\left[(\eta_x \xi_x + \eta_y \xi_y)(g(\eta^{x,y}, \xi) - g(\eta, \xi^{x,y})) f(\eta, \xi)\right]. \tag{2.21}$$

Applying the exchange operator $\xi \mapsto \xi^{x,y}$ to A and resumming we obtain

$$A = -\frac{1}{4} E\left[(\xi_x + \xi_y)(g(\eta, \xi^{x,y}) - g(\eta, \xi))(f(\eta, \xi^{x,y}) - f(\eta, \xi))\right].$$

Applying $\eta \mapsto \eta^{x,y}$ and $\xi \mapsto \xi^{x,y}$ simultaneously in B we obtain

$$B = -\frac{1}{4} E\left[(\eta_x \xi_x + \eta_y \xi_y)(g(\eta, \xi^{x,y}) - g(\eta^{x,y}, \xi))(f(\eta^{x,y}, \xi^{x,y}) - f(\eta, \xi))\right]. \tag{2.22}$$

We write $B = B_1 + B_2 + B_3 + B_4$, where

$$B_1 = \frac{1}{4} E\left[(\eta_x \xi_x + \eta_y \xi_y)(g(\eta^{x,y}, \xi) - g(\eta, \xi))(f(\eta^{x,y}, \xi^{x,y}) - f(\eta, \xi^{x,y}))\right],$$

$$B_2 = \frac{1}{4} E\left[(1 - \eta_x \eta_y)(\eta_x \xi_x + \eta_y \xi_y)(g(\eta^{x,y}, \xi) - g(\eta, \xi))(f(\eta, \xi^{x,y}) - f(\eta, \xi))\right],$$

$$B_3 = \frac{1}{4} E\left[(1 - \eta_x \eta_y)(\eta_x \xi_x + \eta_y \xi_y)(g(\eta, \xi) - g(\eta, \xi^{x,y}))(f(\eta^{x,y}, \xi^{x,y}) - f(\eta, \xi^{x,y}))\right],$$

$$B_4 = \frac{1}{4} E\left[(\eta_x \xi_x + \eta_y \xi_y)(g(\eta, \xi) - g(\eta, \xi^{x,y}))(f(\eta, \xi^{x,y}) - f(\eta, \xi))\right].$$

Notice that in B_2 and B_3 we have slipped in the term $1 - \eta_x \eta_y$, which vanishes when the lattice gradients vanish but otherwise is 1. Now we have $(\eta_x \xi_x + \eta_y \xi_y) \leq (\xi_x + \xi_y)$, and therefore by Schwarz's inequality

$$|B_1| \leq \sqrt{\mathcal{D}_{x,y}(f)} \sqrt{\mathcal{D}_{x,y}(g)}.$$

For B_2 and B_3 note that

$$(1-\eta_x\eta_y)(\eta_x\xi_x+\eta_y\xi_y)\le(\xi_x+\xi_y)\wedge((1-\eta_x)+(1-\eta_y)).$$

Again by Schwarz's inequality

$$|B_2+B_3|\le 4\sqrt{\mathcal{D}_{x,y}(f)}\sqrt{\mathcal{D}_{x,y}(g)}.$$

From $(\eta_x\xi_x+\eta_{x+e}\xi_{x+e})=-(1-\eta_x)\xi_x-(1-\eta_{x+e})\xi_{x+e}+(\xi_x+\xi_{x+e})$,

$$B_4+A=-\frac{1}{4}E\left[((1-\eta_x)\xi_x+(1-\eta_y)\xi_y)(g(\eta,\xi)-g(\eta,\xi^{x,y}))(f(\eta,\xi^{x,y})-f(\eta,\xi))\right].$$

Since $((1-\eta_x)\xi_x+(1-\eta_y)\xi_y)\le(1-\eta_x)+(1-\eta_y)$, Schwarz's inequality gives

$$|B_4+A|\le\sqrt{\mathcal{D}_{x,y}(f)}\sqrt{\mathcal{D}_{x,y}(g)}.$$

This proves that $E[fL_{x,y}g]\le 6\sqrt{\mathcal{D}_{x,y}(f)}\sqrt{\mathcal{D}_{x,y}(g)}$. Summing over nearest neighbor pairs x and y, an application of Schwarz's inequality completes the proof. □

3. Nongradient systems. Let $\zeta=(\eta,\xi)$ be the vector valued occupancy number. For each $\epsilon>0$ and initial distribution μ_0^ϵ our three models define Markov processes $\zeta(t)$ with state space $X_\epsilon=(\{0,1\}\times\{0,1\})^{\mathbf{Z}^d/\epsilon\mathbf{Z}^d}$. We denote by P_ϵ the corresponding measure on $D([0,\infty);X_\epsilon)$, the space of right-continuous paths with left-limits, equipped with the topology of convergence at continuity points. We are primarily interested in the comportment of $\zeta_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}(\epsilon^{-2}\mathbf{t})d\mathbf{x}$. Let $M(\mathbf{T}^d)$ be the set of nonnegative measures on $\mathbf{T}^d$ with total mass bounded above by 1 and let $\hat{P}_\epsilon$ denote the corresponding probability measure on $D([0,T];M(\mathbf{T}^d)\times M(\mathbf{T}^d))$.

In any such model we have

$$d\zeta_x(t)=\sum_{i=1}^d\Big(W_{x-e_i,x}(t)-W_{x,x+e_i}(t)\Big)dt+dM_x(t), \tag{3.1}$$

where $W_{x,x+e}=W_{x,x+e}(t)$, the (vector) rate of particle jumps from x to $x+e$, is a local function of the form $W_{x,x+e}=\tau_xW_{0,e_i}=W_x^i$, and the M_x are martingales. We use e_i for the vector of unit length in the positive i direction on the lattice. The precise form of W_{0,e_i} will be given later. Let ϕ be a smooth function on the torus taking values in $\mathbf{R}^2$. We have

$$\begin{aligned}\int_{\mathbf{T}^d}\Big(\zeta_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}(\epsilon^{-2}\mathbf{t})-\zeta_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}(0)\Big)\phi(\mathbf{x})d\mathbf{x}\\ =\int_0^{\mathbf{t}}\int_{\mathbf{T}^d}\nabla_\epsilon\phi(\mathbf{x})\epsilon^{-1}W_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}(\epsilon^{-2}\mathbf{s})d\mathbf{x}d\mathbf{s}+M_\phi(\mathbf{t}),\end{aligned} \tag{3.2}$$

where $(\nabla_\epsilon\phi)(\mathbf{x})=\epsilon^{-1}[\phi(\mathbf{x}+\epsilon e_i)-\phi(\mathbf{x})]=\nabla\phi(\mathbf{x})+O(\epsilon)$ and M_ϕ is a martingale with variance

$$E[(M_\phi(\mathbf{t}))^2]=\epsilon^d\int_0^{\mathbf{t}}\int_{\mathbf{T}^d}|\nabla_\epsilon\phi|^2(\mathbf{x})\sigma^2_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}(\epsilon^{-2}\mathbf{s})d\mathbf{x}d\mathbf{s}, \tag{3.3}$$

where σ_x is a (bounded) local function specific to the model. Hence the martingale term is of order $\epsilon^{d/2}$ and is negligible in the limit. The problem is therefore to show that as $\epsilon\to 0$,

$$\epsilon^{-1}W_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}(\epsilon^{-2}\mathbf{t})\rightharpoonup D(\varrho)\nabla\varrho, \tag{3.4}$$

where $\varrho = (u, v)$ is the weak limit of $\zeta_{\lfloor \epsilon^{-1}\mathbf{x} \rfloor}(\epsilon^{-2}\mathbf{t})$, and $D = D(\varrho)$ is the diffusion matrix specific to the model. The symbol $\rightharpoonup$ is used to denote weak convergence. In other words, (3.4) means that for any smooth $\phi(\mathbf{x}, \mathbf{t})$,

$$\int_0^{\mathbf{t}} \int_{\mathbf{T}^d} \phi(\mathbf{x}, \mathbf{s}) \epsilon^{-1} W_{\lfloor \epsilon^{-1}\mathbf{x} \rfloor}(\epsilon^{-2}\mathbf{s}) d\mathbf{x} d\mathbf{s} \to \int_0^{\mathbf{t}} \int_{\mathbf{T}^d} \phi(\mathbf{x}, \mathbf{s}) D(\varrho(\mathbf{x}, \mathbf{s})) \nabla \varrho(\mathbf{x}, \mathbf{s}) d\mathbf{x} d\mathbf{s}$$

in probability.

At this point it helps to know what $W_{0,e}$, the current, is in each specific model. In Model 1 it is

$$W_{0,e}^{(1)} = \left(\left(\gamma_1 + \frac{\xi_0 + \xi_e}{2} \right)(\eta_0 - \eta_e), \gamma_2(\xi_0 - \xi_e) \right), \tag{3.5}$$

in Model 2 it is

$$W_{0,e}^{(2)} = \left(\left(\gamma_1 + \frac{\xi_0 + \xi_e}{2} \right)(\eta_0 - \eta_e), \left(\gamma_2 + 1 - \frac{\eta_0 + \eta_e}{2} \right)(\xi_0 - \xi_e) \right), \tag{3.6}$$

and in Model 3 the current is the pair $W_{0,e}^{(3)} = (W_{0,e}^{(3),\eta}, W_{0,e}^{(3),\xi})$, where

$$\begin{aligned} W_{0,e}^{(3),\eta} &= (\gamma_1 + \xi_0)\eta_0(1 - \eta_e) - (\gamma_1 + \xi_e)\eta_e(1 - \eta_0), \\ W_{0,e}^{(3),\xi} &= (\gamma_2 + 1 - \eta_0)\xi_0(1 - \xi_e) - (\gamma_2 + 1 - \eta_e)\xi_e(1 - \xi_0). \end{aligned} \tag{3.7}$$

We shall denote the current generically as $W_{0,e}$ unless we need to differentiate between the three models. Remember that the first coordinate is the current of the η particles and the second coordinate is the current of the ξ particles. For some of the terms above a special simplification occurs; for example, even in Model 3 there are some terms in the current of the form $\eta_0\xi_0 - \eta_e\xi_e$. Since it is a difference of a shift $\tau_e h$ of a function h with itself, called a *gradient*, a summation by parts reduces the key term on the right-hand side of (3.2) to

$$\int_0^t \int_{\mathbf{T}^d} \Delta\phi(\mathbf{x}) h_{\lfloor \epsilon^{-1}\mathbf{x} \rfloor}(\epsilon^{-2}\mathbf{s}) d\mathbf{x} d\mathbf{s}.$$

The difficult ϵ^{-1} is absorbed into the gradient on the test function through an integration by parts, and the much easier problem is now to show that

$$h_{\lfloor \epsilon^{-1}\mathbf{x} \rfloor}(\epsilon^{-2}\mathbf{t}) \rightharpoonup \bar{h}(\varrho), \tag{3.8}$$

where $\bar{h}(\varrho) = E^{\pi_\varrho}[h]$. A system whose currents are of this form is called a *gradient system* (see [14] for a discussion of the question).

Notice that all three systems we are studying are of *nongradient* type. So we have to prove (3.4). One way to do it might be to generate a microscopic variable which we knew converged to $D(\varrho)\nabla\varrho$ and then show that the difference between it and the field $\epsilon^{-1} W_{\lfloor \epsilon^{-1}\mathbf{x} \rfloor}(\epsilon^{-2}\mathbf{t})$ converges weakly to zero.

The simplest candidate is the following. Let ℓ be a positive integer and let $\bar{\zeta}_x^\ell$ denote the average value of ζ on a box Λ_ℓ of side length ℓ around site x, and let

$$\Xi_x^\ell = \left(\sum_{j=1}^d a_{ij}(\bar{\zeta}_x^\ell)(\bar{\zeta}_{x+e_j}^\ell - \bar{\zeta}_x^\ell) \right)_{1 \le i \le 2d},$$

where $D(\varrho) = (a_{ij}(\varrho))_{1\le i,j\le d}$ is the diffusion coefficient in the model. Then $\epsilon^{-1}\Xi^{\ell}_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}(\epsilon^{-2}\mathbf{t})$ is our natural candidate. On the other hand, for a given $\delta > 0$ it is an easy computation using Itô's formula to show that if L is the generator of the process and $g(\varrho,\zeta)$ is any function continuous in the density ϱ and depending only locally on ζ, then the field $\epsilon^{-1}Lg(\bar{\zeta}^{\delta\epsilon^{-1}}_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}, \tau_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}\zeta)(\epsilon^{-2}\mathbf{t})$ converges weakly to 0. Here $\bar{\zeta}^{\delta\epsilon^{-1}}_x$ is just the empirical density on a box of side length $\delta\epsilon^{-1}$ around x (the intermediate scale between the micro- and macroscopic levels). Hence we can replace our simple candidate by a linear combination of *gradient-type* terms plus a negligible part

$$\Xi^{\ell,g}_x = \left(\sum_{j=1}^{d} a_{ij}(\bar{\zeta}^{\ell}_x)(\bar{\zeta}^{\ell}_{x+e_j} - \bar{\zeta}^{\ell}_x) + \tau_x Lg(\bar{\zeta}^{\ell},\zeta)\right)_{1\le i\le 2d} \tag{3.9}$$

with coefficients a_{ij} dependent on $\varrho = (u,v)$ which determine the diffusion matrix $D(\varrho) = (a_{ij}(\varrho))_{1\le i,j\le d}$ uniquely. The problem can now be reduced to the following three lemmas.

LEMMA 3.1. *There exists a sequence g_n of local functions such that*

$$\epsilon^{-1}\left[W_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}(\epsilon^{-2}\mathbf{t}) - \Xi^{\ell,g_n}_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}(\epsilon^{-2}\mathbf{t})\right] \rightharpoonup 0 \tag{3.10}$$

in P_ϵ probability, as $\epsilon\to 0$ followed by $\ell\to\infty$ and $n\to\infty$.

LEMMA 3.2. *The sequence of probability measures $\hat{P}_\epsilon$, as defined in* (2.6)*, is relatively compact, and every limit point $\hat{P}$ is concentrated on absolutely continuous paths with marginal densities $\varrho(\mathbf{t},\mathbf{x})$ satisfying*

$$E^{\hat{P}}\left[\int_0^T\int |\nabla\varrho(\mathbf{t},\mathbf{x})|^2 d\mathbf{x}d\mathbf{t}\right] < \infty. \tag{3.11}$$

We recall the definition of the Hilbert space $\mathcal{H}^{(0)}_{-1,\varrho}$ from (2.8).

LEMMA 3.3. *Let $\tilde{P}_{\epsilon,\ell}$ denote the joint distribution of the fields*

$$(\zeta_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}(\epsilon^{-2}\mathbf{t}), \Xi^{\ell}_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}(\epsilon^{-2}\mathbf{t}))$$

as elements of $\mathcal{H}^{(0)}_{-1,\varrho}$. The sequence is tight, and any limit measure is concentrated on fields of the form $(\varrho, D(\varrho)\nabla\varrho)$.

Suppose we have a functional $F_{\epsilon,K}$ depending on ϵ and some additional parameters which we denote by K and we want to show that $\lim_K \lim_{\epsilon\to 0} E^{P_\epsilon}[F_{\epsilon,K}] = 0$. We now recall the standard machinery which reduces such problems to eigenvalue estimates. Recall that Q_ϵ denotes the equilibrium process, with initial distribution $\pi_{1/2,1/2}$, and that we have the entropy bound $H(P_\epsilon/Q_\epsilon) \le (\log 4)\epsilon^{-d}$ (see (2.17)).

LEMMA 3.4. *Suppose that P_ϵ and Q_ϵ are probability measures with*

$$H(P_\epsilon/Q_\epsilon) = \int \log\frac{dP_\epsilon}{dQ_\epsilon} dP_\epsilon \le C\epsilon^{-d}.$$

If for any $\lambda > 0$,

$$\lim_K \limsup_{\epsilon\to 0} \epsilon^d \log E^{Q_\epsilon}\left[\exp\{\lambda\epsilon^{-d}F_{\epsilon,K}\}\right] \le 0, \tag{3.12}$$

then

$$\lim_K \limsup_{\epsilon \to 0} E^{P_\epsilon}[F_{\epsilon,K}] = 0. \tag{3.13}$$

Proof. This follows from the entropy inequality

$$E^{P_\epsilon}[F] \le \log E^{Q_\epsilon}[\exp F] + H(P_\epsilon/Q_\epsilon). \qquad \square \tag{3.14}$$

LEMMA 3.5. *Let Q be a Markov process $\zeta(s)$, $s \ge 0$, with generator L which is in equilibrium with invariant measure μ. Let $\mathcal{D}$ denote the corresponding Dirichlet form $\mathcal{D}(f) = -E_\mu[fLf]$. Let $V(s,\zeta)$ be bounded. Then*

$$E^Q\left[\exp\left\{\int_0^t V(s,\zeta(s))ds\right\}\right] \le \exp\left\{\int_0^t \lambda(V(s))ds\right\}, \tag{3.15}$$

where $\lambda(V)$ is the principal eigenvalue of $S+V$, $S = (L+L^*)/2$, given by the Raliegh–Ritz formula

$$\lambda(V) = \sup_{f\ge 0,\ \int f d\mu = 1} \left\{\int V f d\mu - \mathcal{D}(\sqrt{f})\right\}. \tag{3.16}$$

Proof. By the Feynman–Kac formula, $u(t,\zeta) = E_\zeta[\exp\{\int_0^t V(t-s,\zeta(s))ds\}]$ solves the equation $\partial_t u = [A+V]u$ with $u(0,\zeta) = 1$. Hence

$$\frac{d}{dt}\int u^2 d\mu = 2\left\{\int Vu^2 - \mathcal{D}(u)\right\} \le 2\lambda(V)\int u^2 d\mu. \tag{3.17}$$

Therefore

$$\begin{aligned} &E^Q\left[\exp\left\{\int_0^t V(t-s,\zeta(s))ds\right\}\right] \\ &\quad = \int u(t)d\mu \le \sqrt{\int u^2(t)d\mu} \le \exp\int_0^t \lambda(V(t-s,\zeta(s)))ds. \qquad \square \end{aligned} \tag{3.18}$$

In our applications $t = \epsilon^{-2}\mathbf{t}$, and hence after rescaling the variational formula becomes

$$\sup_{f\ge 0,\ \int f d\mu = 1} \left\{\epsilon^{-d}\int V f d\mu - \epsilon^{-2}\mathcal{D}(\sqrt{f})\right\}$$

so that we can restrict the variational problem to f with $\mathcal{D}_\epsilon(\sqrt{f}) \le C\epsilon^{2-d}$, which is the same as (2.17).

Since all of our Dirichlet forms have a lower bound in terms of the Dirichlet form $\mathcal{D}^{(0)}$ of symmetric simple exclusions, we can use $\mathcal{D}^{(0)}$ instead of the real Dirichlet form $\mathcal{D}$ in the variational problem to get an upper bound. Thus the key lemmas are reduced to eigenvalue problems for the generator of the symmetric simple exclusion process.

Next we state the standard one and two block estimates in our context (see Chapter 5 of [4] for a proof).

LEMMA 3.6. *Suppose f_ϵ is a sequence of densities of the particle system on $\mathbf{Z}^d/\epsilon^{-1}\mathbf{Z}^d$ with respect to invariant measures $\pi = \pi_{u,v}$ for some fixed $0 < u < 1$, $0 < v < 1$ and satisfying*

$$\mathcal{D}_\epsilon^0(\sqrt{f_\epsilon}) \le C\epsilon^{2-d}.$$

Let g be a local function and $\bar{g}(\varrho) = E^{\pi_\varrho}[g]$. Then

$$\limsup_{\ell\to\infty} \limsup_{\epsilon\to 0} E^{f_\epsilon \pi}\left[Av_{x\in\mathbf{Z}^d/\epsilon^{-1}\mathbf{Z}^d}\left|Av_{|y-x|\le\ell} g(\tau_x\zeta) - \bar{g}(\bar{\zeta}_x^\ell)\right|\right] = 0.$$

Let F be a continuous function on $[0,1]\times[0,1]$. Then

$$\lim_{\substack{\delta\to 0\\ \ell\to\infty}} \lim_{\epsilon\to 0} E^{f_\epsilon\pi}\left[Av_{x\in\mathbf{Z}^d/\epsilon^{-1}\mathbf{Z}^d}\left|Av_{|y-x|\le\delta\epsilon^{-1}} F(\bar{\zeta}_y^{\delta\epsilon^{-1}}) - Av_{|y-x|\le\ell} F(\bar{\zeta}_y^\ell)\right|\right] = 0.$$

Here Av denotes the average and $\bar{\zeta}_x^\ell = Av_{|y-x|\le\ell}\zeta_y$, the average over y in a box of size ℓ.

Now we return to the key replacement, which is Lemma 3.1, which in microscopic variables takes the form

$$\epsilon^{1+d/2}\int_0^{\epsilon^{-2}\mathbf{t}} \sum_{x\in\mathbf{Z}^d/\epsilon^{-1}\mathbf{Z}^d} \Omega_x(s)ds,$$

where $\Omega_x(s) = \phi(\epsilon\mathbf{x},\epsilon^2\mathbf{s})[W_x - \Xi_x]$. Where for gradient systems the key replacement (3.8) is a local law of large numbers, which is proved in the one/two block estimates, for nongradient systems the key replacement is a local central limit theorem.

Let us make this more rigorous. For any vector local function $\mathbf{g}$ define

$$\Omega_x^{\ell,\mathbf{g}} = \phi(\epsilon\mathbf{x},\epsilon^2\mathbf{s})\tau_x\left[\frac{1}{(2\ell'+1)^d}\sum_{|y|\le\ell'} W_y - D(\bar{\zeta}_0^\ell)(\bar{\zeta}_e^{\ell'} - \bar{\zeta}_0^{\ell'}) - \frac{1}{(2\ell'+1)^d}\sum_{|y|\le\ell'}\tau_y L\mathbf{g}\right],$$

where $\ell' = \ell - |\operatorname{supp}(g)|$ so that $\Omega^{\ell,\mathbf{g}}$ depends only on variables in a box of side length $2\ell+1$ about $0\in\mathbf{Z}^d$. Let $L_\ell^{(0)}$ denote the generator of the process where the η and ξ particles independently perform symmetric random walks with simple exclusion on a box of side length ℓ with reflecting boundary conditions. Let $E^{n,m,\ell}$ denote expectation with respect to the canonical measure $u_{n,m}^\ell$, the uniform distribution on configurations on this box with n particles of type η and m of type ξ. Since the system is ergodic when restricted to such a set of configurations and $\Omega^{\ell,\mathbf{g}}$ has mean 0, we can define a nonnegative definite matrix

$$\sigma_{n,m,\ell}^2(\mathbf{g}) = E^{\ell,n,m}[\Omega^{\ell,\mathbf{g}}(-L_\ell^{(0)})^{-1}\Omega^{\ell,\mathbf{g}}]. \tag{3.19}$$

Let L be the generator of a Markov process X_t, $t\ge 0$, on a state space $\mathcal{S}$, reversible with respect to a probability measure μ and with Dirchlet form $\mathcal{D}(f) = -E_\mu[fLf]$. Given a function V on $\mathcal{S}$, let $\lambda(\epsilon V)$ be the principal eigenvalue of $L+\epsilon V$, as in (3.16). Let $m = E_\mu[V]$ and

$$\begin{aligned}\sigma^2(V) &= E_\mu[V(-L)^{-1}V]\\ &= \lim_{T\to\infty}\frac{1}{2}E_\mu\left[\left(\frac{1}{\sqrt{T}}\int_0^T V(X_t)dt\right)^2\right]\\ &= \sup_f\left\{2E_\mu[Vf] - \mathcal{D}(f)\right\}.\end{aligned}$$

We now make this more precise. Let γ be the spectral gap of L,

$$\gamma = \inf_f \{\mathcal{D}(f)/Var(f)\}.$$

The Rayleigh–Schrodinger perturbation series is

$$\lambda(\epsilon V) = \epsilon m + \epsilon^2 \sigma^2(V) + \cdots.$$

We are interested in the following result, which can be found in [13] and also in [4].

LEMMA 3.7. *Assume that L has a spectral gap $\gamma > 0$. Let V be bounded with $E_\mu[V] = 0$. Then*

$$0 \le \lambda(\epsilon V) \le \frac{\epsilon^2}{1 - 2\epsilon\gamma^{-1}\|V\|_\infty}\sigma^2(V).$$

Returning to the setup of our problem, recall the definition (3.19) of $\sigma^2_{n,m,\ell}(g)$.

LEMMA 3.8. *To prove Lemmas* 3.1 *and* 3.3*, it suffices to prove that*

$$\inf_{\mathbf{g}} \limsup_{\ell\to\infty} \sup_{0\le n,m\le(2\ell+1)^d} \ell^d \sigma^2_{n,m,\ell}(g) = 0. \tag{3.20}$$

Proof. If ϕ is a smooth test function, then it is clear that

$$\int_0^{\mathbf{T}} \int_{\mathbf{T}^d} \phi(\mathbf{x}) W_{\lfloor \epsilon^{-1}\mathbf{x}\rfloor}(\epsilon^{-2}\mathbf{t}) d\mathbf{x} d\mathbf{t} \quad \text{and} \quad \int_0^{\mathbf{T}} \int_{\mathbf{T}^d} \phi(\mathbf{x}) W^\ell_{\lfloor \epsilon^{-1}\mathbf{x}\rfloor}(\epsilon^{-2}\mathbf{t}) d\mathbf{x} d\mathbf{t}$$

will have the same limit, where

$$W^\ell_0 = \frac{1}{(2\ell'+1)^d} \sum_{|y|\le\ell'} W_y.$$

Let

$$V^\ell_x = \epsilon^{-1} D(\bar\zeta^\ell_x)(\bar\zeta^{\ell'}_{x+e} - \bar\zeta^{\ell'}_x).$$

By an elementary resummation $\sum_x \phi(\epsilon x)[V^{\epsilon^{-1}\delta}_x - V^\ell_x] = \epsilon^{-1}\sum_x B_x \nabla\zeta_x$, where $\nabla\zeta_x$ is the vector whose ith entry is $\zeta_{x+e_i} - \zeta_x$ and

$$B_x = Av_{|y-x|\le\epsilon^{-1}\delta} D(\bar\zeta^{\epsilon^{-1}\delta}_y)\varphi(\epsilon^{-1}y) - Av_{|y-x|\le\ell'} D(\bar\zeta^\ell_y)\varphi(\epsilon^{-1}y).$$

There is a very simple integration by parts formula which says that for any function $f(\eta,\xi)$, $E^{\ell,n,m}[(\eta_{x+e}-\eta_x)f(\eta,\xi)] = -\frac12 E^{\ell,n,m}[(\eta_{x+e}-\eta_x)(f(\eta^{x,x+e},\xi) - f(\eta,\xi))]$, and analogously for ξ. Since B_x is invariant under the transformations $\eta \mapsto \eta^{x,x+e_i}$ and $\xi \mapsto \xi^{x,x+e_i}$ for any $c > 0$ there exists a $C < \infty$ so that

$$|E[B_x \nabla\zeta_x f]| \le CE[|B_x|^2 f] + \epsilon^{-2}\frac{c}{2}\sum_{i=1}^d \mathcal{D}^{(0)}_{x,x+e_i}(\sqrt{f}). \tag{3.21}$$

Hence for any $c > 0$, f, and bounded ϕ_x

$$\begin{aligned}\epsilon^{d-1} E\left[\sum_x \phi_x (V^{\epsilon^{-1}\delta}_x - V^\ell_x) f\right] - c\epsilon^{d-2}\mathcal{D}^{(0)}(\sqrt{f}) \\ \le C\epsilon^d \sum_x E[|B_x|^2 f] - \frac{c}{2}\epsilon^{d-2}\mathcal{D}^{(0)}(\sqrt{f}).\end{aligned}$$

From the continuity of $D(\varrho)$, this vanishes uniformly over densities f, in the limit $\epsilon \to 0$, followed by $\delta \to 0$, by the two-block estimate.

Remark 4. In order to use the two-block estimate from above, one needs the continuity of the diffusive coefficient $D(u,v)$. We refer the reader to Theorem 5.8 from [4].

By applying Lemmas 3.4 and 3.5, to prove Lemma 3.1 it suffices to verify that for any $\delta > 0$ and bounded ϕ_x

$$\inf_{\mathbf{g}} \limsup_{\ell\to\infty} \limsup_{\epsilon\to 0} \sup_{f} \left\{ \epsilon^{d-1} E^{\pi_\varrho}\left[\sum_x \Omega_x^{\ell,\mathbf{g}} f \right] - \delta \epsilon^{d-2} \mathcal{D}^{(0)}(\sqrt{f}) \right\} \le 0. \tag{3.22}$$

The state space here is $(\{0,1\}\times\{0,1\})^{\mathbf{Z}^d/\epsilon^{-1}\mathbf{Z}^d}$ and the expectations are with respect to a product measure with some fixed $0 < u < 1$, $0 < v < 1$. If we denote $\mathcal{D}_\ell^{(0)} = \sum_{|x|\le\ell, |x+e|\le\ell, |e|=1} \mathcal{D}_{x,x+e}^{(0)}(\sqrt{f})$, where the expectation is with respect to the product measure $\pi_{u,v}^\ell$ on configurations on a box of side length $2\ell+1$, then we have $\sum_x \mathcal{D}_\ell^{(0)}(\sqrt{\tau_{-x} f}) \le L^d \mathcal{D}^{(0)}(\sqrt{f})$, where $L = 2\ell+1$, and therefore

$$\begin{aligned} &\epsilon^{d-1} E\left[\sum_x \Omega_x^{\ell,\mathbf{g}} f \right] - \delta \epsilon^{d-2} \mathcal{D}^{(0)}(\sqrt{f}) \\ &\quad \le \frac{\epsilon^{d-2}\delta}{L^d} \sum_x \sup_f \left\{ \frac{L^d \epsilon}{\delta} E[\Omega_x^{\ell,\mathbf{g}} f] - \mathcal{D}_\ell^{(0)}(\sqrt{f}) \right\}. \end{aligned} \tag{3.23}$$

The expectation is with respect to $\pi_{u,v}^\ell$, but we could instead use the canonical measure $u_{n,m}^\ell$. Since the product measure is just a linear combination of the latter, if we prove it uniformly over n and m we have the result. Now by the previous lemma and the fact that the spectral gap of the exclusion process on a box of side length L is of order L^{-2},

$$\frac{\epsilon^{d-2}\delta}{L^d} \sum_x \sup_f \left\{ \frac{L^d\epsilon}{\delta} E[\Omega_x^{\ell,\mathbf{g}} f] - \mathcal{D}_\ell^{(0)}(\sqrt{f}) \right\} \le C\delta^{-1} L^d \sigma_{n,m,\ell}^2(g).$$

Letting $\ell \to \infty$ we obtain the desired result. □

The previous lemma reduces the proof of the hydrodynamic limit to the evaluation of the asymptotics of certain central limit theorem variances. We now describe how to make these computations. Note that $\Omega^{\ell,\mathbf{g}}$ is an average of shifts of local functions f of three types: 1. the current W; 2. the microscopic density gradients $\nabla\zeta$; 3. incoherent fluctuations Lg. All three have the property that their expectation is zero with respect to any canonical measure on any box containing their support. They also satisfy the following integration by parts formulae with respect to any such measure: For any local h, and nearest neighbors x and y, $E[W_{x,y}h] = -\frac{1}{2}E[W_{x,y}(h(\eta^{x,y},\xi^{x,y}) - h(\eta,\xi))]$, $E[(\zeta_y - \zeta_x)h] = -\frac{1}{2}E[(\zeta_y-\zeta_x)(h(\eta^{x,y},\xi^{x,y}) - h(\eta,\xi))]$, and $E[L_{x,y}gh] = -\frac{1}{2}\mathcal{D}_{x,y}(g,h)$. In particular, each of the three functions f satisfies a bound

$$E[fh] \le C \sum_{|x-y|=1, |x|,|y|\le R} \mathcal{D}_{x,y}^{(0)}(h) \tag{3.24}$$

for some $C, R < \infty$, uniformly over boxes containing $|x|, |y| \le R$ and over the canonical measures on that box. The class of local functions f satisfying a bound of type (3.24)

was denoted by $\mathcal{F}$ in (2.7). Note that this corresponds to local functions for which the asymptotic variance

$$\langle f, f\rangle^{(0)}_{-1,\varrho} = \lim_{\ell\to\infty,\ (\frac{n}{(2\ell+1)^d}, \frac{m}{(2\ell+1)^d})\to\varrho} \frac{1}{(2\ell)^d} E^{n,m,\ell}\left[\sum_{x\le\ell'} \tau_x f, (-L^{(0)}_\ell)^{-1}\left(\sum_{|x|\le\ell'} \tau_x f\right)\right] \tag{3.25}$$

is finite. For any g, h in $\mathcal{F}$ we can define $\langle g, h\rangle^{(0)}_{-1,\varrho}$ by polarization, giving a semi-inner product on $\mathcal{F}$, and $||g||^{(0)}_{-1,\varrho} = (\langle g^2\rangle^{(0)}_\varrho)^{\frac{1}{2}}$ becomes a seminorm. Let $\mathcal{N} = \{g \in \mathcal{F} : ||g||^{(0)}_{-1,\varrho} = 0\}$. The completion of the quotient space $\mathcal{F}/\mathcal{N}$, denoted by $\mathcal{H}^{(0)}_{-1,\varrho}$, is thus a Hilbert space. The first part of the following result first appeared in [10]. A complete proof can be found in [4], so we will not prove it again here. The second part was first proved in a different context (mean-zero asymmetric simple exclusion) by [15]. A nice review is [14, Theorem A, Varadhan's Lecture 5, page 2, at Fields].

THEOREM 3.9. *For each $\varrho = (u, v) \in (0, 1) \times (0, 1)$,*

(1) *the closure of $L^{(0)}\mathcal{F}$ in $\mathcal{H}^{(0)}_{-1,\varrho}$ is a linear subspace of codimension $2d$ and the orthogonal subspace is provided by the span of $\nabla\zeta$;*

(2) *the closure of $L^{(i)}\mathcal{F}$, $i = 1, 2, 3$, in $\mathcal{H}^{(0)}_{-1,\varrho}$ is a linear subspace of codimension $2d$ and a complementary subspace is provided by the span of $\nabla\zeta$.*

Proof. We prove only (2). From (1), it suffices to prove the triviality of the kernel $\mathcal{K}$ of the orthogonal projection from $\overline{L^{(0)}\mathcal{F}}$ to $L\mathcal{F}$. Let $g \in \mathcal{K}$ and $\delta > 0$. Since $g \in \overline{L^{(0)}\mathcal{F}}$ there is an $f \in \mathcal{F}$ with $\| g - L^{(0)} f \|^{(0)}_{-1,\varrho} \le \delta$. From the equivalence of the Dirichlet forms $\mathcal{D}^{(i)}$, $i = 0, 1, 2, 3$, we have $\| L^{(0)} f \|^{(0)}_{-1,\varrho} \le (\gamma^{-1}\langle L^{(0)} f, Lf\rangle^{(0)}_{-1,\varrho})^{1/2}$. Since $g \in \mathcal{K}$, $\langle L^{(0)} f, Lf\rangle^{(0)}_{-1,\varrho} = \langle L^{(0)} f - g, Lf\rangle^{(0)}_{-1,\varrho} \le \delta$. By Schwarz's inequality, $\langle L^{(0)} f - g, Lf\rangle^{(0)}_{-1,\varrho} \le \delta \| Lf \|^{(0)}_{-1,\varrho}$. Hence $\| L^{(0)} f \|^{(0)}_{-1,\varrho} \le \gamma^{-1}\delta \| Lf \|^{(0)}_{-1,\varrho}$. By the strong-sector condition Lemma 2.5, $\| Lf \|^{(0)}_{-1,\varrho} \le C \| L^{(0)} f \|^{(0)}_{-1,\varrho}$. Letting $\delta \downarrow 0$, we have $\| g \|^{(0)}_{-1,\varrho} = 0$. □

4. Tightness. Hence the diffusion coefficient can be identified by the formula $W_{0,e_i} - D^{(i)}(\varrho)\nabla\zeta \in \overline{L^{(i)}\mathcal{F}}$ in $\mathcal{H}^{(0)}_{-1,\varrho}$. In the final section we derive more explicit expressions for D. It remains only to prove compactness of the density fields, Lemma 3.2. We start with a general lemma. For a pure jump function $x(\cdot)$ with a finite number of jumps, the polygonalization $\hat{x}(\cdot)$ is obtained by linearly interpolating between values at successive jumps.

LEMMA 4.1. *Let $\{(Q_\epsilon, P_\epsilon)\}_{\epsilon>0}$ be probability measures on $D([0, T]; \mathbf{R})$ which are supported on pure jump functions such that for some $C_1, C_2 < \infty$, $H(Q_\epsilon/P_\epsilon) \le C_1\epsilon^{-d}$. If, for any $0 \le s < t \le T$ and any $\lambda > 0$,*

$$E^{P_\epsilon}\left[\exp\{\lambda\epsilon^{-d}(x(t) - x(s))\}\right] \le \exp\{C_2\epsilon^{-d}\lambda^2(t - s)\}, \tag{4.1}$$

then there exists $C_3 < \infty$ so that, for any $0 < \delta \le T$,

$$\limsup_{\epsilon\to 0} E^{Q_\epsilon}\left[\sup_{|t-s|<\delta,\ 0\le s,t\le T} |\hat{x}(t) - \hat{x}(s)|\right] \le C_3\sqrt{\delta}\log\delta^{-1}.$$

Proof. The Garsia–Rodemich–Rumsey inequality [12] states that if $x(t)$ is a continuous function and $\psi(x)$ a strictly increasing function such that $\psi(0) = 0$,

$\lim_{x\to\infty}\psi(x)=\infty$, if

$$B=\int_0^T\int_0^T\psi\left(|x(t)-x(s)|/\sqrt{|t-s|}\right)dsdt,$$

then

$$\sup_{|t-s|<\delta\ 0\le s,t\le T}|x(t)-x(s)|\le 4\int_0^\delta\psi^{-1}\left(4Bu^{-2}\right)u^{-1/2}du.$$

Choosing $\psi(x)=\exp\{\epsilon^{-d}x\}-1$ one obtains after some computation that

$$4\int_0^\delta\psi^{-1}\left(4Bu^{-2}\right)u^{-1/2}du\le\epsilon^d C_4(\delta)(1+\log(4B+\delta^2)\vee 0),$$

where $C_4(\delta)=32\sqrt{\delta}\log\delta^{-1}$. Applying this to the polygonalization of $x(t)$,

$$E^{P_\epsilon}\left[\exp\left\{\lambda\epsilon^{-d}\sup_{|t-s|<\delta,0\le s<t\le T}|\hat{x}(t)-\hat{x}(s)|\right\}\right]$$
$$\le E^{P_\epsilon}\left[\exp\{\lambda C_4(\delta)(1+\log(4B+\delta^2))\}\right].$$

By choosing $\lambda=1/C_4(\delta)$, the right-hand side is bounded by $C_5(T)\epsilon^{-d}$ for some $C_5(T)<\infty$ for each $T>0$, from (4.1). It only remains to apply the entropy inequality (3.14). □

LEMMA 4.2. *Let P_ϵ^{eq} be the process starting from equilibrium on $\mathbf{Z}^d/\epsilon^{-1}\mathbf{Z}^d$ and let $V_x=\tau_xV$, where V is any local function satisfying a bound of the form* (3.24). *Then there exists a constant $C<\infty$ such that for any smooth test function $\phi:[0,\mathbf{T}]\times\mathbf{T}^d\to\mathbf{R}$,*

$$E^{P_\epsilon^{eq}}\left[\exp\left\{\epsilon^{-d}\int_{\mathbf{s}}^{\mathbf{t}}\int_{\mathbf{T}^d}\phi(\mathbf{u},\mathbf{x})V_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}(\epsilon^{-2}\mathbf{u})d\mathbf{x}d\mathbf{u}\right\}\right]$$
$$\le\exp\{C\epsilon^{-d}\|\phi(\mathbf{u})\|^2_{L^2([\mathbf{s},\mathbf{t}]\times\mathbf{T}^d)}\}.$$

Proof. By stationarity and Lemma 3.5 $\exp\{2(\mathbf{t}-\mathbf{s})\Lambda_\epsilon\}$ is an upper bound for the left-hand side, where

$$\Lambda_\epsilon=\sup_{E^{\pi_\varrho}[f]=1,\ f\ge 0}\left\{\epsilon^{-(d+1)}\int_{\mathbf{T}^d}\phi(\mathbf{x})E^{\pi_\varrho}[V_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}f]d\mathbf{x}-\epsilon^{-2}\mathcal{D}(\sqrt{f})\right\}.$$

By (3.24) and $a-b=(\sqrt{a}-\sqrt{b})(\sqrt{a}+\sqrt{b})$ we obtain the result. □

THEOREM 4.3. *$\hat{P}_\epsilon$ is tight.*

Proof. By (3.1),

$$\hat{P}_\epsilon\left(\sup_{0\le s<t\le T,\ |t-s|<\delta}\left|\int_{\mathbf{T}^d}[\zeta_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}(\epsilon^{-2}\mathbf{t})-\zeta_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}(\epsilon^{-2}\mathbf{s})]\phi(\mathbf{x})d\mathbf{x}\right|\ge 4\varepsilon\right)$$
$$\le\hat{P}_\epsilon\left(\sup_{0\le s<t\le T,|t-s|<\delta}\left|\int_{\mathbf{s}}^{\mathbf{t}}\int_{\mathbf{T}^d}\nabla_e\phi(\mathbf{x})\epsilon^{-1}W_{\lfloor\epsilon^{-1}\mathbf{x}\rfloor}(\epsilon^{-2}\mathbf{u})d\mathbf{x}d\mathbf{u}\right|\ge 2\varepsilon\right)$$
$$+\hat{P}_\epsilon\left(\sup_{0\le s<t\le T}|M_\phi(\mathbf{t})-M_\phi(\mathbf{s})|\ge 2\varepsilon\right).$$

The third term is of order ϵ^d by Doob's inequality. By the previous lemmas applied to the second term we obtain (3.12). By Lemma 3.4 this suffices. □

5. Diffusion coefficient. In this section we derive formulae for the various diffusion coefficients. For any two vector functions $\mathbf{g}, \mathbf{h}$, we write

$$\langle \mathbf{g}, \mathbf{h} \rangle_{-1,\varrho} = (\langle g_i, h_j \rangle_{-1,\varrho})_{i,j}. \tag{5.1}$$

The current is given by $W = (W^\eta, W^\xi)$, where

$$W^\eta = (W^\eta_{0,e_1}, \dots, W^\eta_{0,e_d}), \qquad W^\xi = (W^\xi_{0,e_1}, \dots, W^\xi_{0,e_d}).$$

The diffusion coefficient in each of the three models $i = 1, 2, 3$ is defined by the equation $W^{(i)} - D^{(i)}\nabla\zeta \in \otimes\overline{L^{(i)}\mathcal{F}}$ as elements of $\mathcal{H}^{(0)}_{-1,\varrho}$. We can also define spaces $\mathcal{H}^{(i)}_{-1,\varrho}$, $i = 1, 2, 3$, by using the analogue of (3.25) as in (2.8) with

$$\langle f, f \rangle^{(i)}_{-1,\varrho} = \lim_{\ell\to\infty} \frac{1}{(2\ell)^d} E^{n,m,\ell}\left[\sum_{x\le \ell'} \tau_x f, (-L^{(i)}_\ell)^{-1}\left(\sum_{|x|\le \ell'} \tau_x f \right)\right].$$

Since the corresponding Dirichlet forms are equivalent, $\mathcal{H}^{(1)}_{-1,\varrho}$ and $\mathcal{H}^{(2)}_{-1,\varrho}$ are equivalent to $\mathcal{H}^{(0)}_{-1,\varrho}$. Hence we can solve $W^{(i)} - D^{(i)}\nabla\zeta \in \otimes\overline{L^{(i)}\mathcal{F}}$ in $\mathcal{H}^{(i)}_{-1,\varrho}$ for $i = 1, 2$. Model 1 is more straightforward, so we describe the details in the case of Model 2 and leave Model 1 to the reader.

Model 2. In Model 2, the two components of the current (3.6) read

$$W^{(2),\eta}_{0,e} = \left(\gamma_1 + \frac{\xi_0 + \xi_e}{2}\right)(\eta_e - \eta_0), \quad W^{(2),\xi}_{0,e} = \left(\gamma_2 + 1 - \frac{\eta_0 + \eta_e}{2}\right)(\xi_e - \xi_0).$$

For any local g and any $1 \le i \le d$, we can compute explicitly

$$\langle W^{(2),\eta}_{0,e_i}, L^{(2)} g\rangle^{(2)}_{-1,\varrho} = \frac{1}{2} E^{\pi_\varrho}\left[\left(\gamma_1 + \frac{\xi_{e_i} + \xi_0}{2}\right)(\eta_0 - \eta_{e_i})\nabla^\eta_{0,e_i} \sum_x \tau_x g\right], \tag{5.2}$$

$$\langle W^{(2),\xi}_{0,e_i}, L^{(2)} g\rangle^{(2)}_{-1,\varrho} = \frac{1}{2} E^{\pi_\varrho}\left[\left(\gamma_2 + 1 - \frac{\eta_{e_i} + \eta_0}{2}\right)(\xi_0 - \xi_{e_i})\nabla^\xi_{0,e_i} \sum_x \tau_x g\right], \tag{5.3}$$

$$\langle L^{(2)} g, L^{(2)} g\rangle^{(2)}_{-1,\varrho} = \frac{1}{2}\sum_{j=1}^d E\left[\left(\gamma_1 + \frac{\xi_{e_j} + \xi_0}{2}\right)\left(\nabla^\eta_{0,e_j} \sum_x \tau_x g\right)^2\right. \tag{5.4}$$

$$\left. + \left(\gamma_2 + 1 - \frac{\eta_{e_j} + \eta_0}{2}\right)\left(\nabla^\xi_{0,e_j} \sum_x \tau_x g\right)^2\right], \tag{5.5}$$

$$\langle \nabla\zeta, L^{(2)} g\rangle^{(2)}_{-1,\varrho} = 0, \tag{5.6}$$

$$\langle W, W\rangle^{(2)}_{-1,\varrho} = \begin{pmatrix} (\gamma_1 + v) I_d & 0 \\ 0 & (\gamma_2 + 1 - u) I_d \end{pmatrix} \chi(\varrho), \tag{5.7}$$

$$\langle W, \nabla\zeta\rangle^{(2)}_{-1,\varrho} = \chi(\varrho). \tag{5.8}$$

THEOREM 5.1. (1). $D^{(2)}(\varrho)\langle \nabla\zeta, \nabla\zeta\rangle_{-1,\varrho} = \chi(\varrho)$.
(2). *For any* $\mathbf{r} \in \mathbf{R}^d \times \mathbf{R}^d$,

$$\mathbf{r}\chi(\varrho)D^{(2)}(\varrho)\mathbf{r}' = \inf_g \sum_{i=1}^d \mathcal{D}_{0,e_i}\left(\mathbf{r}\sum_x x\zeta_x - \sum_x \tau_x g\right). \tag{5.9}$$

The infimum is over local functions g. Note that $\sum_x \tau_x g$ makes no sense alone; however, since g is local, only finitely many terms in the sum are nonzero after applying the discrete gradients ∇_{0,e_i}.

Proof. Since $W^{(2)} - D^{(2)}(\varrho)\nabla\zeta$ is in the closure of $L\mathcal{F}$,

$$\langle W^{(2)} - D^{(2)}(\varrho)\nabla\zeta, \nabla\zeta\rangle_{-1,\varrho} = 0.$$

(1) then follows from (5.8). (2) then follows from (3.9) and (5.2), (5.3), and (5.5). □

We still need to obtain the simpler formula (2.11) for Model 2. Note that if $\mathbf{r} = (r_1^1, \dots, r_d^1, r_1^2, \dots, r_d^2)$, then $\mathcal{D}_{0,e_i}(\mathbf{r}\sum_x x\zeta_x - \sum_x \tau_x g)$ is given explicitly by

$$\frac{1}{2}E\left[\left(\gamma_1 + \frac{\xi_{e_i}+\xi_0}{2}\right)\left(r_i^1(\eta_{e_i}-\eta_0) - \nabla^\eta_{0,e_i}\sum_x \tau_x g\right)^2 + \left(\gamma_2 + 1 - \frac{\eta_{e_i}+\eta_0}{2}\right)\left(r_i^2(\xi_{e_i}-\xi_0) - \nabla^\xi_{0,e_i}\sum_x \tau_x g\right)^2\right].$$

Now for any constants a, b,

$$(a(\eta_e-\eta_0) - \nabla^\eta_{0,e}\sum_x \tau_x g)^2 = \left(\frac{a-b}{2}\right)^2(\eta_e-\eta_0)^2 + \left(\frac{a+b}{2}(\eta_e-\eta_0) - \nabla^\eta_{0,e}\sum_x \tau_x g\right)^2 + \frac{a^2-b^2}{2}(\eta_e-\eta_0)\nabla^\eta_{0,e}\sum_x \tau_x g,$$

$$(b(\xi_e-\xi_0) - \nabla^\xi_{0,e}\sum_x \tau_x g)^2 = \left(\frac{a-b}{2}\right)^2(\xi_e-\xi_0)^2 + \left(\frac{a+b}{2}(\xi_e-\xi_0) - \nabla^\xi_{0,e}\sum_x \tau_x g\right)^2 + \frac{b^2-a^2}{2}(\xi_e-\xi_0)\nabla^\xi_{0,e}\sum_x \tau_x g.$$

Now we claim that for any local function g,

$$E^{\pi_\varrho}\left[\left(\frac{\xi_e+\xi_0}{2}\right)(\eta_e-\eta_0)\nabla^\eta_{0,e}\sum_x \tau_x g - \left(1-\frac{\eta_e+\eta_0}{2}\right)(\xi_e-\xi_0)\nabla^\xi_{0,e}\sum_x \tau_x g\right] = 0. \tag{5.10}$$

To prove it we transfer the $\nabla_{0,e}$ onto the $\nabla\zeta$ to obtain $E^{\pi_\varrho}\left[(\eta_e\xi_e - \eta_0\xi_0)\sum_{x\in\Lambda}\tau_x g\right]$, where the sum is over some large but finite box Λ. Now use the translation invariance of the measure to rewrite this as $E^{\pi_\varrho}\left[\sum_{x\in\Lambda}(\eta_{x+e}\xi_{x+e} - \eta_x\xi_x)g\right]$. The first term is a telescoping sum, and we end up with $E^{\pi_\varrho}[fg]$, where f is mean zero and does not depend on variables ζ_x in a box A around the origin, while g depends only on ζ_x, $x \in A$. Since π_ϱ is a product measure, $E^{\pi_\varrho}[fg] = E^{\pi_\varrho}[f]\,E^{\pi_\varrho}[g] = 0$, which proves (5.10). Then (2.11) follows from this and the explicit form of $\mathcal{D}_{0,e_i}(\mathbf{r}\sum_x x\zeta_x - \sum_x \tau_x g)$ after a little computation.

Model 3. We now show the last part of Theorem 2.4.

Proof. First, we state a general fact about matrices. Let L be an invertible matrix and $L_s = (L^* + L)/2$ its symmetrization. One can check that $[(L^{-1})_s]^{-1} = L^* L_s^{-1} L$, or, in variational form,

$$\langle f, (-L)^{-1} f\rangle = \sup_g \inf_h \left\{2\langle f - Lh, g\rangle - \langle h, Lh\rangle\right\}.$$

In particular, taking $h = -g$ we have

$$\langle f, (-L)^{-1} f\rangle \le \langle f, (-L_s)^{-1} f\rangle.$$

We will apply this in our particular situation where $L_\ell^{(2)} = (L_\ell^{(3)})_s$. Let $T_\ell = L_\ell^{(2)}(L_\ell^{(3)})^{-1}$. This makes sense for any mean-zero function of configurations on $|x| \le \ell$ with n particles of type η and m particles of type ξ, and $\langle T_\ell f, (-L_\ell^{(2)})^{-1} T_\ell f\rangle = \langle f(-L_\ell^{(3)})^{-1} f\rangle \le \langle f(-L_\ell^{(2)})^{-1} f\rangle$, since on such hyperplanes the operators $L_l^{(i)}$ are invertible. Hence T_ℓ is bounded and therefore has a limit T defined on $\mathcal{H}_{-1,\varrho}^{(2)}$ which has the property that $T\overline{L^{(3)}\mathcal{F}} = \overline{L^{(2)}\mathcal{F}}$ and $TW^{(3)} = W^{(2)}$ and whose norm is bounded by 1. The diffusion coefficient is defined by $W^{(3)} - D^{(3)}(\varrho)\nabla\zeta \in \bigotimes_{i=1}^{2d} \overline{L^{(3)}\mathcal{F}}$ in $\mathcal{H}_{-1,\varrho}^{(2)}$. Applying T gives

$$W^{(2)} - D^{(3)}(\varrho)\mathbf{T}\nabla\zeta \in \bigotimes_{i=1}^{2d} \overline{L^{(2)}\mathcal{F}}, \tag{5.11}$$

which implies that $[D^{(3)}(\varrho)\mathbf{T} - D^{(2)}(\varrho)]\nabla\zeta \in \bigotimes_{i=1}^{2d} \overline{L^{(2)}\mathcal{F}}$ or

$$D^{(3)}(\varrho)\langle \mathbf{T}\nabla\zeta, \nabla\zeta\rangle_{-1,\varrho}^{(2)} = D^{(2)}(\varrho)\langle \nabla\zeta, \nabla\zeta\rangle_{-1,\varrho}^{(2)}. \tag{5.12}$$

For $\mathbf{r} \in \mathbf{R}^d \times \mathbf{R}^d$,

$$\begin{aligned}\left\langle \mathbf{T}\sum_i r_i\nabla\zeta_i, \sum_j r_j\nabla\zeta_j\right\rangle_{-1,\varrho}^{(2)} &= \sum_i\sum_j \langle \mathbf{T}\nabla\zeta_i, \nabla\zeta_j\rangle_{-1,\varrho}^{(2)} r_i r_j \\ &\le \sum_i\sum_j \langle \nabla\zeta_i, \nabla\zeta_j\rangle_{-1,\varrho}^{(2)} r_i r_j\end{aligned}$$

due to the bound $\|\mathbf{T}\| \le 1$. We also notice that the upper bound (in the sense of quadratic forms) is achieved if and only if $\mathbf{T}\nabla\zeta = \nabla\zeta$. This would imply that $\mathbf{T}\nabla\zeta \in \bigotimes_{i=1}^{2d} \overline{L^{(2)}\mathcal{F}}$. This leads to a contradiction, due to the property that $\langle \mathbf{T}\nabla\zeta, L^{(2)} g\rangle \ne 0$ (see [2, section 5]). This fact implies that $D_s^{(3)} \ne D^{(2)}$. □

Remark 5. The relation between the asymmetric and symmetric diffusion coefficients $D^{(3)}$ and $D^{(2)}$ is discussed in the context of (one-type particles) asymmetric simple exclusion in section 5 of [8] and the references within. In that context $D^{(2)}$ is diagonal and becomes a multiple of the identity in the *isotropic* case, when the transition probabilities to the neighboring sites in the exclusion process are identical along all possible axes and not just direction-wise (which is the symmetric case). The two properties coincide in dimension $d = 1$. Only in this special situation can one derive that $D^{(3)} \ge \text{const } I$ in the sense of quadratic forms. The property is significant because it shows that the hydrodynamic limit exhibits diffusivity in excess of the one introduced by the random walk (Laplacian). Even though we are able to prove (5.12) we cannot derive that $[D^{(3)}]_{sym} \ge D^{(2)}$ except in *weak sense*, as in Theorem 2.4, meaning that there exist matrices Q (not necessarily symmetric) and V (symmetric) such that $D^{(3)}Q = D^{(2)}V$ and $V > Q_{sym}$.

The difficulties in our model come from two sources. First, we have two types of particles, which have distinct densities in equilibrium; henceforth the compressibility matrix $\chi(\varrho)$ (see Theorem 5.1) is diagonal but not proportional to the identity. Second,

we do not have the option of proving the result in one dimension (as in the one-type particle models).

Under general conditions, without further knowledge of the properties of the matrices $Q = \langle \mathbf{T}\nabla\zeta, \nabla\zeta\rangle$ and $V = \langle \nabla\zeta, \nabla\zeta\rangle$, equation (5.12) *implies* $[D^{(3)}]_{sym} \geq D^{(2)}$ is false. In order to preserve the inequality sign between two matrices in the sense of quadratic forms we would need, for example, that the factors be commutative with the terms of the inequality, at a minimum that $QV = VQ$ in this case, which is not available to us.

REFERENCES

[1] A. De Masi and E. Presutti, *Mathematical Methods for Hydrodynamic Limits*, Lecture Notes in Math. 1501, Springer-Verlag, Berlin, 1991.

[2] R. Esposito, R. Marra, and H. T. Yau, *Diffusive limit of asymmetric simple exclusion*, Rev. Math. Phys., 6 (1994), pp. 1233–1267.

[3] J. Fritz, *On the diffusive nature of entropy flow in infinite systems: Remarks to a paper by Guo–Papanicolaou–Varadhan*, Comm. Math. Phys., 133 (1990), pp. 331–352.

[4] C. Kipnis and C. Landim, *Scaling Limits of Interacting Particle Systems*, Springer-Verlag, New York, 1999.

[5] C. Kipnis, S. Olla, and S. R. S. Varadhan, *Hydrodynamics and large deviation for simple exclusion processes*, Comm. Pure Appl. Math., 42 (1989), pp. 115–137.

[6] K. Komoriya, *Hydrodynamic limit for asymmetric mean zero exclusion processes with speed change*, Ann. Inst. H. Poincaré Probab. Statist., 34 (1998), pp. 767–797.

[7] K. Komoriya, *An asymmetric exclusion process related to vortex flow in viscous planar fluid*, in Probability Theory and Mathematical Statistics, World Scientific, River Edge, NJ, 1996, pp. 220–228.

[8] C. Landim, S. Olla, and H. T. Yau, *First-order correction for the hydrodynamic limit of asymmetric simple exclusion processes in dimension $d \geq 3$*, Comm. Pure Appl. Math., 50 (1997), pp. 149–203.

[9] W.-M. Ni, *Diffusion, cross-diffusion, and their spike-layer steady states*, Notices Amer. Math. Soc., 45 (1998), pp. 9–18.

[10] J. Quastel, *Diffusion of color in the simple exclusion process*, Comm. Pure Appl. Math., 45 (1998), pp. 321–379.

[11] H. Spohn, *Large Scale Dynamics of Interacting Particles*, Springer-Verlag, Berlin, 1991.

[12] D. W. Stroock and S. R. S. Varadhan, *Multidimensional Diffusion Processes*, Grundlehren Math. Wiss. 233, Springer-Verlag, Berlin, New York, 1979.

[13] S. R. S. Varadhan, *Nonlinear diffusion limit for a system with nearest neighbor interactions-II*, in Asymptotic Problems in Probability Theory: Stochastic Models and Diffusions on Fractals, K. D. Elworthy and N. Ikeda, eds., Longman, Harlow, UK, 1993, pp. 75–128.

[14] S. R. S. Varadhan, *Lectures on hydrodynamic scaling*, in Hydrodynamic Limits and Related Topics, Fields Inst. Commun. 26, S. Feng, A. T. Lawniczak, and S. R. S. Varadhan, eds., AMS, Providence, RI, 2000, pp. 3–40.

[15] L. Xu, *Hydrodynamics for Asymmetric Mean Zero Simple Exclusion*, Ph.D. thesis, New York University, New York, 1993.

SIAM J. MATH. ANAL.
Vol. 35, No. 6, pp. 1534–1560

ON AVERAGING PRINCIPLES: AN ASYMPTOTIC EXPANSION APPROACH*

R. Z. KHASMINSKII† AND G. YIN†

Abstract. This work is concerned with diffusion processes having fast and slow components. It was known that under suitable assumptions the slow component can be approximated by the Markov process with averaged characteristics. In this work, asymptotic expansions for the solutions of the Kolmogorov backward equations are constructed and justified. Certain probabilistic conclusions and examples are also provided.

Key words. singular perturbation, diffusion, Kolmogorov backward equation, asymptotic expansion

AMS subject classifications. 34E05, 60J27, 60F05

DOI. 10.1137/S0036141002403973

1. Introduction. This work is concerned with asymptotic properties of $(X_\varepsilon(t), Y_\varepsilon(t))$, a pair of singularly perturbed diffusions with a small parameter $\varepsilon > 0$, where $X_\varepsilon(t)$ is an $\mathbb{R}^r$-valued diffusion and $Y_\varepsilon(t)$ is an $\mathbb{R}^d$-valued diffusion. There are weak and strong interactions between these processes such that the evolution of $X_\varepsilon(t)$ is fast changing, whereas that of $Y_\varepsilon(t)$ is slowly varying. The system of diffusions takes the form

$$\begin{cases} dX_\varepsilon = \dfrac{1}{\varepsilon} b_1(X_\varepsilon, Y_\varepsilon)dt + \dfrac{1}{\sqrt{\varepsilon}}\sigma_1(X_\varepsilon, Y_\varepsilon)dw_1(t), \\ dY_\varepsilon = b_2(X_\varepsilon, Y_\varepsilon)dt + \sigma_2(X_\varepsilon, Y_\varepsilon)dw_2(t), \end{cases} \tag{1.1}$$

where $b_1(\cdot) : \mathbb{R}^r \times \mathbb{R}^d \mapsto \mathbb{R}^r$, $b_2(\cdot) : \mathbb{R}^r \times \mathbb{R}^d \mapsto \mathbb{R}^d$, $\sigma_1(\cdot) : \mathbb{R}^r \times \mathbb{R}^d \mapsto \mathbb{R}^r \times \mathbb{R}^r$, and $\sigma_2(\cdot) : \mathbb{R}^r \times \mathbb{R}^d \mapsto \mathbb{R}^d \times \mathbb{R}^d$ are appropriate functions, and $w_1(\cdot)$ and $w_2(\cdot)$ are independent standard Brownian motions taking values in $\mathbb{R}^r$ and $\mathbb{R}^d$, respectively.

In [11], equation (1.1) and more general systems were analyzed. It was proved that when the fast varying component is positive recurrent, the slow component $Y_\varepsilon(\cdot)$ converges weakly to a Markov diffusion process $Y(\cdot)$. Owing to the importance in the modeling and analysis of various stochastic systems arising from mechanics, climatology, wireless communication, signal processing, manufacturing, and production planning, such systems have received renewed interest in recent years; see [18, 20, 24, 28]. In fact, as was mentioned in [25], all physical systems have a certain hierarchy in which not all components, parts, or subsystems vary at the same rate. Some of them change rapidly, and others evolve slowly. A convenient way of modeling such phenomena is to note the high contrast of the fast-slow nature and to formulate the problem as a singularly perturbed system with a small parameter such as the one given in (1.1). The singular perturbation formulation then provides a viable alternative for treating complex systems. For example, suppose that one wishes to control a large and

*Received by the editors March 12, 2002; accepted for publication (in revised form) June 13, 2003; published electronically March 11, 2004.
http://www.siam.org/journals/sima/35-6/40397.html
†Department of Mathematics, Wayne State University, Detroit, MI 48202 (rafail@math.wayne.edu, gyin@math.wayne.edu). The research of the first author was supported in part by the Office of Naval Research under grant N00014-95-1-0793 and in part by the National Science Foundation under grant DMS-9971608. The research of the second author was supported in part by the National Science Foundation under grants DMS-9971608 and DMS-0304928.

complex system. Using the singular perturbation formulation, in lieu of treating the system directly, one can consider its limit, whose complexity is much reduced compared with the original system. Based on the optimal control of the reduced system, one will be able to construct near-optimal control of the original systems. As demonstrated in [18], many problems arising in engineering systems are too complicated to handle, so it is common to simplify the systems by eliminating or averaging out some "transient" or "quickly stabilized" components. A two-time-scale controlled diffusion has the form

$$\begin{cases} dX_\varepsilon = \dfrac{1}{\varepsilon} b_1(X_\varepsilon, Y_\varepsilon, u)dt + \dfrac{1}{\sqrt{\varepsilon}}\sigma_1(X_\varepsilon, Y_\varepsilon, u)dw_1(t), \quad X_\varepsilon(0) = x, \\ dY_\varepsilon = b_2(X_\varepsilon, Y_\varepsilon, u)dt + \sigma_2(X_\varepsilon, Y_\varepsilon, u)dw_2(t), \quad Y_\varepsilon(0) = y, \end{cases} \tag{1.2}$$

where $b_1(\cdot) : \mathbb{R}^r \times \mathbb{R}^d \times \mathbb{R}^{d_1} \mapsto \mathbb{R}^r$, $b_2(\cdot) : \mathbb{R}^r \times \mathbb{R}^d \times \mathbb{R}^{d_1} \mapsto \mathbb{R}^d$, $\sigma_1(\cdot) : \mathbb{R}^r \times \mathbb{R}^d \times \mathbb{R}^{d_1} \mapsto \mathbb{R}^r \times \mathbb{R}^r$, $\sigma_2(\cdot) : \mathbb{R}^r \times \mathbb{R}^d \times \mathbb{R}^{d_1} \mapsto \mathbb{R}^d \times \mathbb{R}^d$, $u(\cdot)$ is the control taking values in some compact subset of $\mathbb{R}^{d_1}$, and $w_1(t)$ and $w_2(t)$ are independent Brownian motions. The objective is to find admissible controls such that the expected cost function,

$$J(x, y, u(\cdot)) = E_{x,y} \int_0^T \widehat{c}_1(X_\varepsilon(s), Y_\varepsilon(s), u(s))ds + E_{x,y}\widehat{c}_2(X_\varepsilon(T), Y_\varepsilon(T), u(T)),$$

is minimized, where $E_{x,y}$ denotes the expectation taken with respect to $X_\varepsilon(0) = x$ and $Y_\varepsilon(0) = y$. Weak convergence methods were used in [18] to develop a framework to handle such systems. The analytic techniques presented in this paper may provide an alternative, although many more details have to be worked out.

In studying singularly perturbed Markov processes, weak convergence methods were used in [11]; see also the recent work [23]. Asymptotic expansions of solutions for the forward equations were obtained in [13] and [14]. This work considers the adjoint problem, namely, asymptotic properties of the solutions of the Kolmogorov backward equations. In the literature, the study of certain asymptotic properties of the backward equations was carried out in, for example, [21, 22], with the main focus on the leading term of the asymptotic expansion. In a related reference, multiple scale asymptotic expansions in the nonseparable form $\sum \varepsilon^i g_i(t, t/\varepsilon, x, y)$ were considered in [3] (see, in particular, the appendix of that paper). This work is devoted to the study of developing the full asymptotic expansions of the solutions of backward equations. As will be seen, the full asymptotic expansion is a delicate case to handle. We derive the asymptotic expansions in a separable form, namely, the additions of the outer expansions and initial layer corrections, which is easily applicable to various applications (e.g., consideration of functional central limit theorem results). Our asymptotic expansions reveal the averaging principles in a more illuminating way.

The rest of the paper is arranged as follows. Section 2 presents the precise problem formulation. Using the ideas of singular perturbation theory, section 3 proceeds with the construction of the matched asymptotic expansions such that the outer expansions are smooth and the initial layer corrections decay exponentially fast. Then the asymptotic expansions are fully justified and uniform error bounds are obtained in section 4. We present some probabilistic interpretation in section 5 together with an example. Included in this section are also further remarks and extensions. Finally, the proofs of a couple of lemmas are given in the appendix. Throughout the paper, we often use c_i to denote generic positive constants independent of ε, whose value may change for different usages; for a diffusion process $Y(t)$, $Y^y(t)$ means that $Y(0) = y$.

2. Formulation. Let $\mathbf{K}_r$ and $\mathbf{K}_d$ be r-dimensional and d-dimensional Riemann C^∞-manifolds, respectively. Denote $\mathbf{K} = \mathbf{K}_r \times \mathbf{K}_d$, and denote the corresponding totality of C^∞-vector fields by $\mathcal{X}(\mathbf{K}_r)$ and $\mathcal{X}(\mathbf{K}_d)$, respectively. Following the notation of [8, p. 231], we assume that these manifolds are connected and compact throughout the paper. Let $A_0(x,y),\dots,A_{n_1}(x,y) \in \mathcal{X}(\mathbf{K}_r)$ and $B_0(x,y),\dots,B_{n_2}(x,y) \in \mathcal{X}(\mathbf{K}_d)$. Suppose that $x = (x_1,\dots,x_r) \in \mathbf{K}_r$ and $y = (y_1,\dots,y_d) \in \mathbf{K}_d$ are the local coordinates of x and y, respectively. For a small parameter $\varepsilon > 0$, consider the Markov diffusion process on $\mathbf{K}$, whose generator is given by

$$\begin{aligned} &A_\varepsilon g(x,y) \\ &= \frac{1}{\varepsilon}\left[\frac{1}{2}\sum_{\alpha=1}^{n_1} A_\alpha(A_\alpha g)(x,y) + (A_0 g)(x,y)\right] + \frac{1}{2}\sum_{\beta=1}^{n_2} B_\beta(B_\beta g)(x,y) + (B_0 g)(x,y) \\ &\stackrel{\text{def}}{=} \frac{1}{\varepsilon}\mathcal{L}_1 g(x,y) + \mathcal{L}_2 g(x,y). \end{aligned} \tag{2.1}$$

Treating $y \in \mathbf{K}_d$ as a parameter, we consider also a family of Markov diffusion processes $X(t|y)$ on $\mathbf{K}_r$ with the generator $\mathcal{L}_1(x,y)$. We assume that $\mathcal{L}_1(x,y)$ is nondegenerate on $\mathbf{K}_r$ for all $y \in \mathbf{K}_d$ (see [8, section 5.4]). Then it is known (see [8, Proposition 4.5, p. 278]) that for the process $X(t|y)$ there exists a unique stationary density that is a solution of

$$\mathcal{L}_1^*(x,y)\mu(x|y) = 0, \qquad \int_{\mathbf{K}_r} \mu(x|y)dx = 1, \tag{2.2}$$

where $\mathcal{L}_1^*(x,y)$ is the adjoint of $\mathcal{L}_1(x,y)$ with respect to the inner product in $\mathbf{K}_r$,

$$\begin{aligned} \langle f,g\rangle_{\mathbf{K}_r} &= \int_{\mathbf{K}_r} f(x)g(x)dx, \\ dx &= \sqrt{\det G}\,dx_1dx_2\cdots dx_r, \end{aligned}$$

and $G = \|g_{ij}\|$ is a Riemannian metric on $\mathbf{K}_r$. Our goal is to construct and justify the asymptotic expansion in powers of ε for the solution of the Cauchy problem

$$\frac{\partial u_\varepsilon}{\partial t} = \left[\frac{1}{\varepsilon}\mathcal{L}_1(x,y) + \mathcal{L}_2(x,y)\right]u_\varepsilon + c(x,y)u_\varepsilon + f(x,y), \tag{2.3}$$

where $c(x,y)$ and $f(x,y)$ are C^∞ functions on $\mathbf{K}$.

Using the local coordinates, we can write the Cauchy problem for the Kolmogorov backward equation as

$$\begin{cases} \dfrac{\partial u_\varepsilon}{\partial t} = \left(\dfrac{1}{\varepsilon}\mathcal{L}_1(x,y) + \mathcal{L}_2(x,y)\right)u_\varepsilon + c(x,y)u_\varepsilon + f(x,y), \\ u_\varepsilon(0,x,y) = \varphi(x,y), \end{cases} \tag{2.4}$$

where

$$\begin{aligned} \mathcal{L}_1(x,y) &= \sum_{i,j=1}^{r} a_{1,ij}(x,y)\frac{\partial^2}{\partial x_i\partial x_j} + \sum_{i=1}^{r} b_{1,i}(x,y)\frac{\partial}{\partial x_i}, \\ \mathcal{L}_2(x,y) &= \sum_{i,j=1}^{d} a_{2,ij}(x,y)\frac{\partial^2}{\partial y_i\partial y_j} + \sum_{i=1}^{d} b_{2,i}(x,y)\frac{\partial}{\partial y_i}, \end{aligned} \tag{2.5}$$

and $a_\ell(x,y) = (a_{\ell,ij}(x,y)) = \sigma_\ell(x,y)\sigma_\ell'(x,y)/2$ for $\ell = 1,2$. We will use the following condition throughout the paper.

(A) The $c(\cdot)$, $f(\cdot)$, $\varphi(\cdot)$, $a_\ell(\cdot)$, and $b_\ell(\cdot)$ (for $\ell = 1,2$) are C^∞ functions on $\mathbf{K}$. Moreover, for any $\lambda \in \mathbb{R}^r$ with $|\lambda| > 0$,

$$\begin{aligned} &\sum_{i,j=1}^{r} a_{1,ij}(x,y)\lambda_i\lambda_j > 0, \\ &\sum_{i,j=1}^{r} a_{2,ij}(x,y)\lambda_i\lambda_j \geq 0. \end{aligned} \tag{2.6}$$

Although it is possible to consider diffusions on the entire space, it is more convenient to work with processes that are defined on a compact subset; otherwise, one has to impose conditions to ensure the desired ergodicity. Since our main concern here is the asymptotic expansions, it appears to be more instructive to concentrate on the constructive methods rather than to be concerned with the existence of solutions of the equation $\mathcal{L}_1 u = f$ in a noncompact domain.

Remark 2.1. The assumption $\mathbf{K}_d$ being compact allows to us obtain error bounds for the asymptotic expansions uniform in y. For future use, for a suitable function $h(\cdot)$, for $(x,y) \in \mathbf{K}$, define

$$\begin{aligned} \overline{h}(y) &\stackrel{\text{def}}{=} \int_{\mathbf{K}_r} h(x,y)\mu(x,y)dx \\ &\stackrel{\text{def}}{=} \langle h(\cdot,y), \mu(\cdot,y)\rangle . \end{aligned} \tag{2.7}$$

For references on ergodicity of diffusion processes and related problems, we refer the reader to [12, 26] and the references therein. Since we are working with a compact set $\mathbf{K}$, the smoothness of the functions given implies that they are bounded uniformly on $\mathbf{K}$.

In what follows, we aim to construct two sequences $\{u_i(t,x,y)\}$ and $\{v_i(t/\varepsilon,x,y)\}$ such that $u_i(t,x,y)$ are the outer expansions, $v_i(t/\varepsilon,x,y)$ are the initial layer corrections, and the asymptotic expansion

$$u_0(t,y) + \sum_{i=1}^{n} \varepsilon^i u_i(t,x,y) + \sum_{i=0}^{n} \varepsilon^i v_i(t/\varepsilon,x,y) \tag{2.8}$$

well approximates $u_\varepsilon(t,x,y)$, the solution of (2.4). For $k = 0,1,\ldots,n$, define a sequence of approximation errors by

$$\begin{aligned} e_{\varepsilon,0}(t,x,y) &= u_0(t,y) + v_0(t/\varepsilon,x,y) - u_\varepsilon(t,x,y), \\ e_{\varepsilon,k}(t,x,y) &= u_0(t,y) + \sum_{i=1}^{k} \varepsilon^i u_i(t,x,y) + \sum_{i=0}^{k} \varepsilon^i v_i(t/\varepsilon,x,y) - u_\varepsilon(t,x,y). \end{aligned} \tag{2.9}$$

We will show that the error term is of the order $O(\varepsilon^{n+1})$ uniformly in $(t,x,y) \in [0,T] \times \mathbf{K}$ for some $T > 0$.

The technique that we are using is singular perturbation theory; see [2, 27] and [9, 22]. However, the realizations of the construction procedure are rather involved. We blend probabilistic methods with those of analytic techniques. The solution method is interesting in its own right. It may shed some light on other problems involving singularly perturbed diffusions in which the approach that we develop in this paper can be adopted.

3. Asymptotic expansions. This section is divided into several parts. First, we set forth the constructive method by presenting the equations involved. Next, we present a couple of technical lemmas to be used in the subsequent development. Then we construct the leading term in the asymptotic expansion. The more difficult part appears in the next stage, namely, constructions of higher order expansions. In addition to the construction of the asymptotic expansions, we also demonstrate that the initial layer corrections decay exponentially fast.

3.1. Differential equations satisfied by $u_i(t,x,y)$ and $v_i(\tau,x,y)$. We seek the asymptotic expansion (2.8). To get the desired error estimates (see Lemma 4.2), we need a couple of extra terms, $u_{n+1}(\cdot)$ and $v_{n+1}(\cdot)$. To derive the differential equations satisfied by $u_i(\cdot)$ and $v_i(\cdot)$, using the singular perturbation techniques, substituting the outer expansion terms $u_0(t,y)+\sum_{i=1}^{n+1}\varepsilon^i u_i(t,x,y)$ into (2.4), and equating coefficients of like power of ε^i yield

$$\frac{\partial}{\partial t}u_0(t,y)=\mathcal{L}_1(x,y)u_1(t,x,y)+\mathcal{L}_2(x,y)u_0(t,y)+c(x,y)u_0(t,y)+f(x,y) \tag{3.1}$$

and

$$\begin{aligned}\frac{\partial}{\partial t}u_k(t,x,y)&=\mathcal{L}_1(x,y)u_{k+1}(t,x,y)+\mathcal{L}_2(x,y)u_k(t,x,y)\\&\quad+c(x,y)u_k(t,x,y),\qquad k=1,\dots,n+1.\end{aligned} \tag{3.2}$$

Likewise, denoting the fast time variable by $\tau=t/\varepsilon$, we obtain

$$\frac{\partial}{\partial \tau}v_0(\tau,x,y)=\mathcal{L}_1(x,y)v_0(\tau,x,y) \tag{3.3}$$

and

$$\begin{aligned}\frac{\partial}{\partial \tau}v_k(\tau,x,y)&=\mathcal{L}_1(x,y)v_k(\tau,x,u)+\mathcal{L}_2(x,y)v_{k-1}(\tau,x,y)\\&\quad+c(x,y)v_{k-1}(\tau,x,y),\qquad k=1,\dots,n+1.\end{aligned} \tag{3.4}$$

The initial conditions are chosen so that

$$\begin{aligned}u_0(0,y)+v_0(0,x,y)&=\phi(x,y),\\u_k(0,x,y)+v_k(0,x,y)&=0,\quad k=1,\dots,n+1.\end{aligned} \tag{3.5}$$

In addition, we will require that $v_k(\tau,x,y)\to 0$ as $\tau\to\infty$.

Our task to follow is to find the functions $\{u_i(t,x,y)\}$ and $\{v_i(\tau,x,y)\}$ in a constructive manner. We proceed to carry out the constructions "recursively" as it should be done in the actual computation. That is, we first determine the leading terms $u_0(t,y)$ and $v_0(\tau,x,y)$ and obtain the decay property of $v_0(\tau,x,y)$. Then we find $u_1(t,x,y)$ and $v_1(\tau,x,y)$ and verify the decay property of $v_1(\tau,x,y)$. Subsequently, we carry out the procedure inductively to get $u_i(t,x,y)$ and $v_i(\tau,x,y)$.

3.2. Auxiliary results. With the y suppressed, we write $\mathcal{L}(x)$ in this section. It is known that if $\mathcal{L}(x)$ is an elliptic operator [7] on $\mathbf{K}_r$, then (2.2) has a unique solution. Note that (2.2) is the Kolmogorov–Fokker–Planck equation for the density of invariant measure of the Markov process $X(t)$ with generator $\mathcal{L}(x)$. The probability density function of the diffusion process (the Green function for the equation

$(\partial/\partial t)u = \mathcal{L}(x)u)$ verifies the so-called Doeblin condition (see [4]): For $t \geq t_0 > 0$, $\inf_{x,x_1 \in \mathbf{K}_r} p(x,t,x_1) > 0$. This condition implies the exponential convergence of $p(x,t,x_1)$ to $\mu(x_1)$ in the sense that

$$|p(x,t,x_1) - \mu(x_1)| \leq c_1 \exp(-c_2 t) \quad \text{as } t \to \infty. \tag{3.6}$$

Analogues to (2.7) denote

$$\overline{\psi} = \int_{\mathbf{K}_r} \psi(x)\mu(x)dx = \langle \psi, \mu \rangle .$$

Using (3.6), we can assert that for any bounded measurable function $\psi : \mathbf{K}_r \mapsto \mathbb{R}$,

$$|E\psi(X^x(\tau)) - \overline{\psi}| \leq c_1 \exp(-c_2 \tau). \tag{3.7}$$

We present some preliminary results. The following lemmas will be needed in the construction of the asymptotic expansions. Let $\mathcal{L}(x)$ be an elliptic operator in $\mathbf{K}_r$. It can be given in local coordinates by

$$\mathcal{L}(x) = \sum_{i,j=1}^{r} a_{ij}(x) \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{n} b_i(x) \frac{\partial}{\partial x_i} \tag{3.8}$$

with $\sum_{i,j=1}^{r} a_{ij}(x)\lambda_i\lambda_j > 0$ for all $x \in \mathbf{K}_r, |\lambda| > 0$.

LEMMA 3.1. *Let $\mathcal{L}(x)$ be given by* (3.8) *and $x \in \mathbf{K}_r$. Let $\mu(x)$ be the unique solution of* (2.2), *and let $\psi(x)$ be a continuous function satisfying $\overline{\psi} = 0$. Then there exists a solution of the equation*

$$\mathcal{L}(x)V = \psi(x).$$

The proof of Lemma 3.1 can be found, for instance, in [1, Theorem 2.3.12] in a more abstract form. However, for the reader's convenience, we provide an alternative proof in the appendix. The following lemma gives the uniqueness of the solution (unique up to a constant), whose proof is also in the appendix.

LEMMA 3.2. *Let $\mathcal{L}(x)$ be given by* (3.8). *Then any solution of*

$$\mathcal{L}(x)W(x) = 0, \quad x \in \mathbf{K}_r, \tag{3.9}$$

is a constant.

Remark 3.3. Note that

$$\mathcal{L}(x)u = \psi(x) \tag{3.10}$$

has a solution only if $\overline{\psi} = 0$ since

$$\langle \mathcal{L}u, \mu \rangle = \langle u, \mathcal{L}^*\mu \rangle = 0. \tag{3.11}$$

Moreover, it follows from Lemma 3.2 that any solution of (3.10) can be written in the form

$$u(x) = c + \widetilde{u}(x), \tag{3.12}$$

where c is an arbitrary constant and $\widetilde{u}(x)$ is a particular solution of (3.10).

LEMMA 3.4. *Suppose that $x \in \mathbf{K}_r$ and $V(\tau, x)$ is a solution of*

$$\frac{\partial}{\partial \tau} V(\tau, x) = \mathcal{L}(x)V(\tau, x) + F(\tau, x), \qquad V(0, x) = \psi(x), \tag{3.13}$$

where $\mathcal{L}(x)$ is elliptic and $F(\tau, x)$ decays exponentially fast; i.e.,

$$\sup_{x \in \mathbf{K}_r} |F(\tau, x)| \le c_1 \exp(-c_2 \tau) \quad \text{for some } c_i > 0,\ i = 1, 2. \tag{3.14}$$

Let $\mu(x)$ be the solution of (2.2). *Then*

$$\left| V(\tau, x) - \overline{\psi} - \int_0^\infty ds \int_{\mathbf{K}_r} F(s, x)\mu(x)dx \right| \le c_1 \exp(-c_2 \tau). \tag{3.15}$$

Proof. See the appendix. □

3.3. Leading terms. Let us start with the determination of $u_0(t, y)$ and $v_0(\tau, x, y)$. Using (2.4) and (3.1), choose the initial condition

$$u_0(0, y) = \overline{\varphi}(y) = \int_{\mathbf{K}_r} \varphi(x, y)\mu(x, y)dx. \tag{3.16}$$

To determine $u_0(t, y)$, multiplying (3.1) by $\mu(x, y)$, integrating with respect to x, and using $\langle u_1(t, \cdot, y), \mathcal{L}_1^*(x, y)\mu(x, y)\rangle = 0$ lead to

$$\begin{aligned} \frac{\partial}{\partial t} u_0(t, y) &= \langle \mathcal{L}_1(x, y)u_1(t, \cdot, y), \mu(\cdot, y)\rangle + \overline{\mathcal{L}}_2(y)u_0(t, y) + \overline{c}(y)u_0(t, y) + \overline{f}(y) \\ &= \overline{\mathcal{L}}_2(y)u_0(t, y) + \overline{c}(y)u_0(t, y) + \overline{f}(y), \end{aligned} \tag{3.17}$$

where

$$\begin{aligned} \overline{\mathcal{L}}_2(y) &= \int_{\mathbf{K}_r} \mathcal{L}_2(x, y)\mu(x, y)dx, \\ \overline{c}(y) = \int_{\mathbf{K}_r} c(x, y)\mu(x, y)dx, &\qquad \overline{f}(y) = \int_{\mathbf{K}_r} f(x, y)\mu(x, y)dx. \end{aligned} \tag{3.18}$$

By virtue of the well-known results in parabolic partial differential equations [6, 19], the Cauchy problem given by (3.16)–(3.18) has a unique solution. Thus $u_0(t, y)$ has been found.

To proceed, the Cauchy problem of (3.3), together with the initial condition

$$v_0(0, x, y) = \varphi(x, y) - \overline{\varphi}(y) \stackrel{\text{def}}{=} \Phi_0(x, y), \tag{3.19}$$

has a unique solution $v_0(\tau, x, y)$. Up to now, we have found both $u_0(t, y)$and $v_0(\tau, x, y)$. We demonstrate that $v_0(\tau, x, y)$ decays exponentially fast.

PROPOSITION 3.5. *The initial layer term $v_0(\tau, x, y)$ satisfies*

$$|v_0(\tau, x, y)| \le c_1 \exp(-c_2 \tau). \tag{3.20}$$

Proof. It is apparent that $v_0(0, x, y)$ is orthogonal to $\mu(x, y)$ (i.e., $\overline{v}_0(0, y) = 0$). It follows from (3.7) that

$$|v_0(\tau, x, y)| = |Ev_0(0, X^{x,y}(\tau), y) - \overline{v}_0(0, y)| \le c_1 \exp(-c_2 \tau).$$

The proposition thus follows. □

At first glance, the constants c_1 and c_2 could be y-dependent. However, the compactness of $\mathbf{K}_d$ implies that these constants can be taken to be uniformly in y, thus independent of y. For the subsequent use, we derive another lemma that is on the bounds of the mixed partial derivatives (with respect to the variable y) of $v_0(t,x,y)$ up to the fourth order. Using the multi-index convention (see, e.g., [5, p. 3]), let $\nu=(\nu_1,\dots,\nu_d)$ be a d-tuple of nonnegative integers, which is referred to as a multi-index with $|\nu|=\sum_{i=1}^d \nu_i$. For $y\in\mathbf{K}_d$, write

$$\frac{\partial^{|\nu|}}{\partial y_1^{\nu_1}\cdots\partial y_d^{\nu_d}} \qquad \text{or} \qquad \frac{\partial^{\nu_1+\nu_2+\cdots+\nu_d}}{\partial y_1^{\nu_1}\cdots\partial y_d^{\nu_d}}.$$

LEMMA 3.6. *The following bounds hold:*

$$\left|\frac{\partial^{|\nu|}v_0(\tau,x,y)}{\partial y_1^{\nu_1}\cdots\partial y_d^{\nu_d}}\right|\le c_1\exp(-c_2\tau),\quad |\nu|=1,\dots,4. \tag{3.21}$$

Remark 3.7. First note that c_i for $i=1,2$ are generic positive constants according to our convention. Their values may be different for different appearances. The estimate (3.21) is needed in the construction of $v_1(\tau,x,y)$ and obtaining the exponential bounds for $v_1(\tau,x,y)$ and $\mathcal{L}_2(x,y)v_1(\tau,x,y)$.

Proof of Lemma 3.6. It suffices to verify (3.21) for each $1\le|\nu|\le 4$.

Step 1. Verify (3.21) for $|\nu|=1$. It suffices to show that for each $i_1=1,\dots,d$,

$$\left|\frac{\partial}{\partial y_{i_1}}v_0(\tau,x,y)\right|\le c_1\exp(-c_2\tau). \tag{3.22}$$

Denote $v_{0,1}^{(i_1)}(\tau,x,y)=(\partial/\partial y_{i_1})v_0(\tau,x,y)$, where the superscript (i_1) denotes the dependence on i_1. Then

$$\begin{aligned}\frac{\partial}{\partial\tau}v_{0,1}^{(i_1)}(\tau,x,y)&=\mathcal{L}_1(x,y)v_{0,1}^{(i_1)}(\tau,x,y)+\left(\frac{\partial}{\partial y_{i_1}}\mathcal{L}_1(x,y)\right)v_0(\tau,x,y),\\ v_{0,1}^{(i_1)}(0,x,y)&=\frac{\partial}{\partial y_{i_1}}\Phi_0(x,y).\end{aligned} \tag{3.23}$$

Recall that $\Phi_0(x,y)=\varphi(x,y)-\overline{\varphi}(y)$. To derive the exponential decay property of $v_{0,1}^{(i_1)}(\tau,x,y)$ by applying Lemma 3.4, we need to verify that

$$\left|\left(\frac{\partial}{\partial y_{i_1}}\mathcal{L}_1(x,y)\right)v_0(\tau,x,y)\right|\le c_1\exp(-c_2\tau). \tag{3.24}$$

Owing to the boundedness of $\mathbf{K}$, all coefficients of $\mathcal{L}_1(x,y)$ and $\mathcal{L}_2(x,y)$ are bounded together with their derivatives. Denote by $G(\tau,x,\xi|y)$ the Green function for (3.3) with y being a parameter. Then

$$v_0(\tau,x,y)=\int_{\mathbf{K}_r}G(\tau/2,x,\xi|y)v_0(\tau/2,\xi,y)d\xi. \tag{3.25}$$

It is well known (see [6]) that as long as s is away from 0, i.e., $s\ge s_0>0$, $(\partial/\partial x)G(s,x,\xi|y)$ and $(\partial^2/\partial x^2)G(s,x,\xi|y)$ are bounded, where $(\partial/\partial x)$ and $(\partial^2/\partial x^2)$

are the gradient and Hessian with respect to the variable x, respectively. Differentiating (3.25) with respect to x (twice) and using (3.20), we obtain

$$\left|\frac{\partial}{\partial x}v_0(\tau,x,y)\right| + \left|\frac{\partial^2}{\partial x^2}v_0(\tau,x,y)\right| \le c_1\exp(-c_2\tau), \tag{3.26}$$

and (3.24) follows.

Next, applying Lemma 3.4 to (3.23) leads to

$$\left|v_{0,1}^{(i_1)}(\tau,x,y) - A_1\right| = \left|\frac{\partial}{\partial y_{i_1}}v_0(\tau,x,y) - A_1\right| \le c_1\exp(-c_2\tau),$$

where A_1 is given by

$$A_1 = \overline{\frac{\partial}{\partial y_{i_1}}\Phi_0(y)} + \int_0^\infty\int_{\mathbf{K}_r}\mu(x,y)\left(\frac{\partial}{\partial y_{i_1}}\mathcal{L}_1(x,y)\right)v_0(s,x,y)dxds.$$

Note that A_1 in fact depends on i_1, but for notational simplicity we suppress the i_1-dependence. We need only show that $A_1 \equiv 0$. Integrating (3.3) with respect to τ and sending $\tau\to\infty$ yield

$$\Phi_0(x,y) + \int_0^\infty \mathcal{L}_1(x,y)v_0(s,x,y)ds = 0. \tag{3.27}$$

Thus,

$$\begin{aligned}A_1 &= \int_0^\infty ds\int_{\mathbf{K}_r}\mu(x,y)dx\Bigg[\left(\frac{\partial}{\partial y_{i_1}}\mathcal{L}_1(x,y)\right)v_0(s,x,y) - \left(\frac{\partial}{\partial y_{i_1}}\mathcal{L}_1(x,y)\right)v_0(s,x,y)\\ &\qquad - \mathcal{L}_1(x,y)\frac{\partial}{\partial y_{i_1}}v_0(s,x,y)\Bigg]\\ &= -\int_0^\infty ds\left\langle\mathcal{L}_1(\cdot,y)\frac{\partial}{\partial y_{i_1}}v_0(s,\cdot,y),\mu(\cdot,y)\right\rangle = 0.\end{aligned}$$

In the above, we have used (3.27) and the fact $\langle\mathcal{L}_1 w,\mu\rangle = 0$. Thus (3.21) is proven for $|\nu| = 1$.

Step 2. Verify (3.21) for $|\nu| = 2$. For each $i_2 = 1,\dots,d$, define $v_{0,2}^{(i_1,i_2)}(\tau,x,y) = (\partial/\partial y_{i_2})v_{0,1}^{(i_1)}(\tau,x,y)$. Then $v_{0,2}^{(i_1,i_2)}(\tau,x,y)$ satisfies

$$\begin{cases}\dfrac{\partial}{\partial\tau}v_{0,2}^{(i_1,i_2)}(\tau,x,y) = \mathcal{L}_1(x,y)v_{0,2}^{(i_1,i_2)}(\tau,x,y) + \left[\dfrac{\partial}{\partial y_{i_1}}\mathcal{L}_1(x,y)\right]v_{0,1}^{(i_2)}(\tau,x,y)\\ \qquad\qquad + \left[\dfrac{\partial}{\partial y_{i_2}}\mathcal{L}_1(x,y)\right]v_{0,1}^{(i_1)}(t,x,y) + \left[\dfrac{\partial^2}{\partial y_{i_1}y_{i_2}}\mathcal{L}_1(x,y)\right]v_0(\tau,x,y),\\ v_{0,2}^{(i_1,i_2)}(0,x,y) \quad = \dfrac{\partial^2}{\partial y_{i_1}\partial y_{i_2}}\Phi_0(x,y).\end{cases} \tag{3.28}$$

Noting the smoothness of $a_{1,ij}(x,y)$ and $b_{1,ij}(x,y)$,

$$\left|\left(\frac{\partial^2}{\partial y_{i_1}\partial y_{i_2}}\mathcal{L}_1(x,y)\right)v_0(\tau,x,y)\right| \le c_1\exp(-c_2\tau) \tag{3.29}$$

by (3.26). Considering (3.23) and taking into account (3.21) for $|\nu| = 1$ and (3.24), by virtue of [17, Theorem 8.11.1],

$$\left|\frac{\partial}{\partial x}v_{0,1}^{(i_1)}(\tau,x,y)\right| + \left|\frac{\partial^2}{\partial x^2}v_{0,1}^{(i_1)}(\tau,x,y)\right| \le c_1\exp(-c_2\tau). \tag{3.30}$$

Combining (3.29) and (3.30),

$$\begin{aligned}&\left|\left[\frac{\partial}{\partial y_{i_1}}\mathcal{L}_1(x,y)\right]v_{0,1}^{(i_2)}(\tau,x,y)\right| + \left|\left[\frac{\partial}{\partial y_{i_2}}\mathcal{L}_1(x,y)\right]v_{0,1}^{(i_1)}(\tau,x,y)\right| \\ &\quad + \left|\left[\frac{\partial^2}{\partial y_{i_1}\partial y_{i_2}}\mathcal{L}_1(x,y)\right]v_0(\tau,x,y)\right| \le c_1\exp(-c_2\tau),\end{aligned} \tag{3.31}$$

so Lemma 3.4 is applicable.

By Lemma 3.4, $|v_{0,2}^{(i_1,i_2)}(\tau,x,y) - A_2| \le c_1\exp(-c_2\tau)$, where

$$\begin{aligned}A_2 =& \int_0^\infty\int_{\mathbf{K}_r}\left[\left[\frac{\partial}{\partial y_{i_1}}\mathcal{L}_1(x,y)\right]v_{0,1}^{(i_2)}(s,x,y) + \left[\frac{\partial}{\partial y_{i_2}}\mathcal{L}_1(x,y)\right]v_{0,1}^{(i_1)}(s,x,y)\right]\mu(x,y)dxds \\ &+ \int_0^\infty\int_{\mathbf{K}_r}\left[\frac{\partial^2}{\partial y_{i_1}\partial y_{i_2}}\mathcal{L}_1(x,y)\right]v_0(s,x,y)\mu(x,y)dxds + \overline{\frac{\partial^2}{\partial y_{i_1}\partial y_{i_2}}\Phi_0(y)}.\end{aligned}$$

We proceed to prove that $A_2 = 0$.

Integrating (3.23), we have

$$-\frac{\partial}{\partial y_{i_1}}\Phi_0(x,y) = \int_0^\infty \mathcal{L}_1(x,y)v_{0,1}^{(i_1)}(s,x,y)ds + \int_0^\infty\left[\frac{\partial}{\partial y_{i_1}}\mathcal{L}_1(x,y)\right]v_0(s,x,y)ds.$$

Thus,

$$\begin{aligned}A_2 =& \int_0^\infty\int_{\mathbf{K}_r}\left[\left[\frac{\partial}{\partial y_{i_1}}\mathcal{L}_1(x,y)\right]v_{0,1}^{(i_2)}(s,x,y) + \left[\frac{\partial}{\partial y_{i_2}}\mathcal{L}_1(x,y)\right]v_{0,1}^{(i_1)}(s,x,y)\right]\mu(x,y)dxds \\ &+ \int_0^\infty\int_{\mathbf{K}_r}\left(\frac{\partial^2}{\partial y_{i_1}\partial y_{i_2}}\mathcal{L}_1(x,y)\right)v_0(s,x,y)\mu(x,y)dxds \\ &- \int_{\mathbf{K}_r}\frac{\partial}{\partial y_{i_2}}\left[\int_0^\infty\left[\mathcal{L}_1(x,y)v_{0,1}^{(i_1)}(s,x,y) + \left[\frac{\partial}{\partial y_{i_1}}\mathcal{L}_1(x,y)\right]v_0(s,x,y)\right]ds\right]\mu(x,y)dx \\ =& \int_{\mathbf{K}_r}\mu(x,y)dx\int_0^\infty ds\left\{\left[\frac{\partial}{\partial y_{i_1}}\mathcal{L}_1(x,y)\right]v_{0,1}^{(i_2)}(s,x,y) + \left[\frac{\partial}{\partial y_{i_2}}\mathcal{L}_1(x,y)\right]v_{0,1}^{(i_1)}(s,x,y)\right. \\ &+ \left(\frac{\partial^2}{\partial y_{i_1}\partial y_{i_2}}\mathcal{L}_1(x,y)\right)v_0(s,x,y) \\ &\left.- \frac{\partial}{\partial y_{i_2}}\left(\mathcal{L}_1(x,y)v_{0,1}^{(i_1)}(s,x,y)\right) - \frac{\partial}{\partial y_{i_2}}\left[\left(\frac{\partial}{\partial y_{i_1}}\mathcal{L}_1(x,y)\right)v_0(s,x,y)\right]\right\} \\ =& -\int_0^\infty ds\int_{\mathbf{K}_r}\mathcal{L}_1(x,y)\left(\frac{\partial}{\partial y_{i_2}}v_{0,1}^{(i_1)}(s,x,y)\right)\mu(x,y)dx \\ =& -\int_0^\infty ds\left\langle\mathcal{L}_1\frac{\partial v_{0,1}^{(i_1)}}{\partial y_{i_2}},\mu\right\rangle = 0.\end{aligned}$$

Thus, the exponential decay in (3.21) is verified for $|\nu| = 2$.

Step 3. Verify (3.21) for $|\nu| = 3$ and 4. For each $i_3 = 1, \dots, d$, let $v_{0,3}^{(i_1,i_2,i_3)}(\tau, x, y) = (\partial/\partial y_{i_3}) v_{0,2}^{(i_1,i_2)}(\tau, x, y)$. Then $v_{0,3}^{(i_1,i_2,i_3)}(\tau, x, y)$ satisfies

$$\begin{cases} \dfrac{\partial}{\partial \tau} v_{0,3}^{(i_1,i_2,i_3)}(\tau, x, y) &= \mathcal{L}_1(x, y) v_{0,3}^{(i_1,i_2,i_3)}(\tau, x, y) + V_3(\tau, x, y), \\ v_{0,3}^{(i_1,i_2,i_3)}(0, x, y) &= \dfrac{\partial^3}{\partial y_{i_1} \partial y_{i_2} \partial y_{i_3}} \Phi_0(x, y), \end{cases} \tag{3.32}$$

where

$$V_3(\tau, x, y) = \frac{\partial}{\partial y_{i_3}} \Bigg\{ \left[\frac{\partial}{\partial y_{i_1}} \mathcal{L}_1(x, y) \right] v_{0,1}^{(i_2)}(\tau, x, y) + \left[\frac{\partial}{\partial y_{i_2}} \mathcal{L}_1(x, y) \right] v_{0,1}^{(i_1)}(t, x, y) + \left[\frac{\partial^2}{\partial y_{i_1} y_{i_2}} \mathcal{L}_1(x, y) \right] v_0(\tau, x, y) \Bigg\} + \left[\frac{\partial}{\partial y_{i_3}} \mathcal{L}_1(x, y) \right] v_{0,2}^{(i_1,i_2)}(\tau, x, y).$$

By virtue of [17, Theorem 8.11.1], we have $|V_3(\tau, x, y)| \le c_1 \exp(-c_2 \tau)$. Then Lemma 3.4 implies that $|v_{0,3}^{(i_1,i_2,i_3)}(\tau, x, y) - A_3| \le c_1 \exp(-c_2 \tau)$, where

$$A_3 = \overline{\frac{\partial^3}{\partial y_{i_1} \partial y_{i_2} \partial y_{i_3}} \Phi_0(y)} + \int_0^\infty ds \int_{\mathbf{K}_r} V_3(s, x, y) \mu(x, y) dx ds.$$

Similar to Steps 1 and 2, we can establish $A_3 = 0$.

Likewise, for each $i_4 = 1, \dots, d$, define $v_{0,4}^{(i_1,i_2,i_3,i_4)}(\tau, x, y) = (\partial/\partial y_{i_4}) v_{0,3}^{(i_1,i_2,i_3)}(\tau, x, y)$. Then $v_{0,4}^{(i_4)}(\tau, x, y)$ satisfies

$$\begin{cases} \dfrac{\partial}{\partial \tau} v_{0,4}^{(i_1,i_2,i_3,i_4)}(\tau, x, y) &= \mathcal{L}_1(x, y) v_{0,4}^{(i_1,i_2,i_3,i_4)}(\tau, x, y) + V_4(\tau, x, y), \\ v_{0,4}^{(i_1,i_2,i_3,i_4)}(0, x, y) &= \dfrac{\partial^4}{\partial y_{i_1} \partial y_{i_2} \partial y_{i_3} \partial y_{i_4}} \Phi_0(x, y), \end{cases} \tag{3.33}$$

where

$$V_4(\tau, x, y) = \frac{\partial}{\partial y_{i_4}} V_3(\tau, x, y) + \left[\frac{\partial}{\partial y_{i_4}} \mathcal{L}_1(x, y) \right] v_{0,3}^{(i_1,i_2,i_3)}(\tau, x, y).$$

By virtue of [17, Theorem 8.11.1], we have $|V_4(\tau, x, y)| \le c_1 \exp(-c_2 \tau)$. Then Lemma 3.4 implies that $|v_{0,4}^{(i_1,i_2,i_3,i_4)}(\tau, x, y) - A_4| \le c_1 \exp(-c_2 \tau)$, where

$$A_4 = \overline{\frac{\partial^4}{\partial y_{i_1} \partial y_{i_2} \partial y_{i_3} \partial y_{i_4}} \Phi_0(y)} + \int_0^\infty ds \int_{\mathbf{K}_r} V_4(s, x, y) \mu(x, y) dx ds.$$

As in Steps 1 and 2, we can establish $A_4 = 0$. Hence the lemma follows. □

3.4. Construction of $u_1(t, x, y)$ and $v_1(\tau, x, y)$. Next, subtracting (3.17) from (3.1), we arrive at

$$\mathcal{L}_1(x, y) u_1(t, x, y) = \Psi_0(t, x, y), \tag{3.34}$$

where

$$\Psi_0(t, x, y) \stackrel{\text{def}}{=} (\overline{\mathcal{L}}_2(y) - \mathcal{L}_2(x, y)) u_0(t, y) + (\overline{c}(y) - c(x, y)) u_0(t, y) + (\overline{f}(y) - f(x, y)).$$

Note that $\Psi_0(t,x,y)$ satisfies the condition $\overline{\Psi_0(t,\cdot,y)}=0$. Thus, Lemmas 3.1 and 3.2 imply that

$$u_1(t,x,y)=U_1(t,y)+\widetilde{u}_1(t,x,y), \tag{3.35}$$

where $\widetilde{u}_1(t,x,y)$ is a particular solution of (3.34) satisfying

$$\int_{\mathbf{K}_r}\widetilde{u}_1(t,x,y)\mu(x,y)dx=0. \tag{3.36}$$

We proceed to determine $U_1(t,y)$, which has to be obtained via the match of the initial layer term $v_1(\tau,x,y)$. Using (3.4) with $k=1$, we solve the Cauchy problem

$$\begin{aligned}\frac{\partial}{\partial\tau}v_1(\tau,x,y)&=\mathcal{L}_1(x,y)v_1(\tau,x,y)+\mathcal{L}_2(x,y)v_0(\tau,x,y)+c(x,y)v_0(\tau,x,y),\\ v_1(0,x,y)&=-u_1(0,x,y)=-U_1(0,y)-\widetilde{u}_1(0,x,y).\end{aligned} \tag{3.37}$$

By requiring $v_1(\tau,x,y)\to 0$ as $\tau\to\infty$, Lemmas 3.6 and 3.4 yield

$$\begin{aligned}&\int_{\mathbf{K}_r}v_1(0,x,y)\mu(x,y)dx+\int_0^\infty\int_{\mathbf{K}_r}\mathcal{L}_2(x,y)v_0(s,x,y)\mu(x,y)dxds\\ &\quad+\int_0^\infty\int_{\mathbf{K}_r}c(x,y)v_0(s,x,y)\mu(x,y)dxds=0.\end{aligned} \tag{3.38}$$

Using the initial condition given in (3.37), we rewrite (3.38) as

$$\begin{aligned}\int_{\mathbf{K}_r}u_1(0,x,y)\mu(x,y)dx=&\int_0^\infty ds\int_{\mathbf{K}_r}\mathcal{L}_2(x,y)v_0(s,x,y)\mu(x,y)dx\\ &+\int_0^\infty ds\int_{\mathbf{K}_r}c(x,y)v_0(s,x,y)\mu(x,y)dx.\end{aligned} \tag{3.39}$$

Using (3.35), (3.37) can be written as

$$\begin{aligned}\frac{\partial}{\partial t}U_1(t,y)+\frac{\partial}{\partial t}\widetilde{u}_1(t,x,y)=&\mathcal{L}_1(x,y)u_2(t,x,y)+\mathcal{L}_2(x,y)U_1(t,y)+\mathcal{L}_2(x,y)\widetilde{u}_1(t,x,y)\\ &+c(x,y)U_1(t,y)+c(x,y)\widetilde{u}_1(t,x,y).\end{aligned}$$

Again, multiplying through by $\mu(x,y)$ and integrating with respect to $x\in\mathbf{K}_r$, we arrive at

$$\frac{\partial}{\partial t}U_1(t,y)=\overline{\mathcal{L}}_2(y)U_1(t,y)+\overline{c}(y)U_1(t,y)+\widetilde{\Psi}_1(t,y), \tag{3.40}$$

where $\widetilde{\Psi}_1(t,y)$ is a known function given by

$$\widetilde{\Psi}_1(t,y)\overset{\text{def}}{=}\int_{\mathbf{K}_r}\mathcal{L}_2(x,y)\widetilde{u}_1(t,x,y)\mu(x,y)dx+\int_{\mathbf{K}_r}c(x,y)\widetilde{u}_1(t,x,y)\mu(x,y)dx. \tag{3.41}$$

In the above, we used $\int_{\mathbf{K}_r}\frac{\partial}{\partial t}\widetilde{u}_1(t,x,y)\mu(x,y)dx=0$. To determine $U_1(t,y)$, we need to find the initial condition $U_1(0,y)$.

Letting $t=0$ in (3.34) gives us

$$U_1(0,y)=u_1(0,x,y)-\widetilde{u}_1(0,x,y).$$

Multiplying through by the invariant density $\mu(x,y)$ and integrating over $\mathbf{K}_r$, together with (3.36) and (3.38), we have

$$\begin{aligned} U_1(0,y) &= \int_0^\infty ds \int_{\mathbf{K}_r} \mathcal{L}_2(x,y) v_0(s,x,y)\mu(x,y)dx \\ &\quad + \int_0^\infty ds \int_{\mathbf{K}_r} c(x,y) v_0(s,x,y)\mu(x,y)dx. \end{aligned} \tag{3.42}$$

Thus, the solution of the Cauchy problem given by (3.40) and (3.42) is uniquely determined. As a result, $u_1(t,x,y)$ is determined; so is $v_1(\tau,x,y)$.

The choice of the initial condition and (3.37)–(3.42) imply

$$|v_1(\tau,x,y)| \le c_1 \exp(-c_2\tau).$$

We proceed to verify

$$\left| \frac{\partial^{|\nu|} v_1(\tau,x,y)}{\partial y_1^{\nu_1} \cdots \partial y_d^{\nu_d}} \right| \le c_1 \exp(-c_2\tau), \quad |\nu| = 1,\dots,4. \tag{3.43}$$

Similar to the proof of (3.21), for each $i_1 = 1,\dots,d$, define $v_{1,1}^{(i_1)}(\tau,x,y) = (\partial/\partial y_{i_1}) v_1(\tau,x,y)$. It follows from (3.37) that

$$\begin{aligned} \frac{\partial}{\partial\tau} v_{1,1}^{(i_1)}(\tau,x,y) &= \mathcal{L}_1(x,y) v_{1,1}^{(i_1)}(\tau,x,y) + \left(\frac{\partial}{\partial y_{i_1}} \mathcal{L}_1(x,y) \right) v_1(\tau,x,y) \\ &\quad + \frac{\partial}{\partial y_{i_1}} (\mathcal{L}_2(x,y) v_0(\tau,x,y)) + \frac{\partial}{\partial y_{i_1}} (c(x,y) v_0(\tau,x,y)), \\ v_{1,1}^{(i_1)}(0,x,y) &= -\left[\frac{\partial}{\partial y_{i_1}} U(0,y) + \frac{\partial}{\partial y_{i_1}} \widetilde{u}_1(0,x,y) \right]. \end{aligned}$$

By virtue of Lemma 3.6,

$$\left| \frac{\partial}{\partial y_{i_1}} (\mathcal{L}_2(x,y) v_0(\tau,x,y)) + \frac{\partial}{\partial y_{i_1}} (c(x,y) v_0(\tau,x,y)) \right| \le c_1 \exp(-c_2\tau).$$

To apply Lemma 3.4, it suffices to verify

$$\left| \left(\frac{\partial}{\partial y_{i_1}} \mathcal{L}_1(x,y) \right) v_1(\tau,x,y) \right| \le c_1 \exp(-c_2\tau). \tag{3.44}$$

Parallel to the argument used in Step 1 of the proof of Lemma 3.6, denoting the associated Green's function for (3.37) as $G(\tau,x,\xi|y)$, we obtain

$$\begin{aligned} v_1(\tau,x,y) &= \int_{\mathbf{K}_r} G(\tau/2,x,\xi|y) v_1(\tau/2,\xi,y)dy \\ &\quad + \int_{\tau/2}^\tau ds \int_{\mathbf{K}_r} G(s,x,\xi|y) \widetilde{F}_1(s,\xi,y)dy, \end{aligned}$$

where $\widetilde{F}_1(\tau,x,y) = \mathcal{L}_2(x,y) v_0(\tau,x,y) + c(x,y) v_0(\tau,x,y)$. As in the argument in the paragraph containing (3.25)–(3.26), we conclude that

$$\left| \frac{\partial}{\partial x} v_1(\tau,x,y) \right| + \left| \frac{\partial^2}{\partial x^2} v_1(\tau,x,y) \right| \le c_1 \exp(-c_2\tau),$$

and hence (3.44) follows.

Redefine A_1 (again with the i_1-dependence suppressed) as

$$\begin{aligned}A_1 = {} & \overline{\frac{\partial}{\partial y_{i_1}} v_{1,1}^{(i_1)}(0,x,y)} \\ & + \int_0^\infty ds \int_{\mathbf{K}_r} \mu(x,y) \Bigg[\left(\frac{\partial}{\partial y_{i_1}} \mathcal{L}_1(x,y) \right) v_1(s,x,y) \\ & \qquad + \frac{\partial}{\partial y_{i_1}} (\mathcal{L}_2(x,y) v_0(s,x,y)) + \frac{\partial}{\partial y_{i_1}} (c(x,y) v_0(s,x,y)) \Bigg] dx.\end{aligned}$$

Lemma 3.4 implies that $|v_{1,1}^{(i_1)}(\tau,x,y) - A_1| \le c_1 \exp(-c_2\tau)$. We next show that $A_1 \equiv 0$.

Integrating (3.37) over τ yields

$$-v_1(0,x,y) = \int_0^\infty (\mathcal{L}_1(x,y) v_1(s,x,y) + \mathcal{L}_2(x,y) v_0(s,x,y) + c(x,y) v_0(s,x,y))\, ds,$$

and hence

$$\begin{aligned}& \int_{\mathbf{K}_r} \frac{\partial}{\partial y_{i_1}} v_1(0,x,y) \mu(x,y) dx \\ & \quad = - \int_{\mathbf{K}_r} \mu(x,y) \frac{\partial}{\partial y_{i_1}} \Bigg[\int_0^\infty [\mathcal{L}_1(x,y) v_1(s,x,y) \\ & \qquad + \mathcal{L}_2(x,y) v_0(s,x,y) + c(x,y) v_0(s,x,y)] ds \Bigg] dx.\end{aligned}$$

Replacing $\overline{(\partial/\partial y_{i_1}) v_{1,1}^{(i_1)}(0,x,y)}$ in A_1 by the right-hand side above, upon cancellation, we obtain

$$\begin{aligned}A_1 = {} & - \int_{\mathbf{K}_r} \mu(x,y) \frac{\partial}{\partial y_{i_1}} \Bigg[\int_0^\infty (\mathcal{L}_1(x,y) v_1(s,x,y) + \mathcal{L}_2(x,y) v_0(s,x,y) \\ & \qquad + c(x,y) v_0(s,x,y)) ds \Bigg] dx \\ & + \int_{\mathbf{K}_r} \mu(x,y) \int_0^\infty \Bigg[\left(\frac{\partial}{\partial y_{i_1}} \mathcal{L}_1(x,y) \right) v_1(s,x,y) + \frac{\partial}{\partial y_{i_1}} (\mathcal{L}_2(x,y) v_0(s,x,y)) \\ & \qquad + \frac{\partial}{\partial y_{i_1}} (c(x,y) v_0(s,x,y)) \Bigg] dx \\ = {} & - \int_{\mathbf{K}_r} \mu(x,y) \int_0^\infty \mathcal{L}_1(x,y) v_{1,1}^{(i_1)}(s,x,y) ds dx \\ = {} & 0,\end{aligned}$$

due to the fact that $\langle \mathcal{L}_1(x,y) v_{1,1}^{(i_1)}, \mu \rangle = 0$. Thus (3.43) is verified for $|\nu| = 1$. Likewise, we can define $v_{1,2}^{(i_1,i_2)}(\tau,x,y)$, $v_{1,3}^{(i_1,i_2,i_3)}(\tau,x,y)$, and $v_{1,4}^{(i_1,i_2,i_3,i_4)}(\tau,x,y)$ and proceed to verify (3.43) as in the proof of Lemma 3.6 for $|\nu| = 2,3,4$.

3.5. Construction of $u_k(t,x,y)$ and $v_k(\tau,x,y)$ for $k \ge 2$. Using similar methods as in the last section, we can obtain outer expansions and initial layer corrections $u_k(t,x,y)$ and $v_k(\tau,x,y)$ for $k = 2,\dots,n+1$. We do this inductively.

Suppose that for each $k = 2,\dots,n+1$, $u_{k-1}(t,x,y)$ and $v_{k-1}(\tau,x,y)$ have been determined such that

$$u_{k-1}(t,x,y) = U_{k-1}(t,y) + \widetilde{u}_{k-1}(t,x,y),$$

where $\widetilde{u}_{k-1}(t,x,y)$ satisfying

$$\int_{\mathbf{K}_r} \widetilde{u}_{k-1}(t,x,y)\mu(x,y)dx = 0 \tag{3.45}$$

is a particular solution of

$$\mathcal{L}_1(x,y)u_{k-1}(t,x,y) = \Psi_{k-2}(t,x,y) \tag{3.46}$$

and where

$$\begin{aligned}\Psi_{k-2}(t,x,y) = \frac{\partial}{\partial t}u_{k-2}(t,x,y) - \overline{\frac{\partial}{\partial t}u_{k-2}(t,\cdot,y)}\\ + \overline{\mathcal{L}_2(\cdot,y)u_{k-2}(t,\cdot,y)} - \mathcal{L}_2(x,y)u_{k-2}(t,x,y)\\ + \overline{c(\cdot,y)u_{k-2}(t,\cdot,y)} - c(x,y)u_{k-2}(t,x,y)\end{aligned}$$

and

$$\overline{F}(t,\cdot,y) = \int_{\mathbf{K}_r} F(t,x,y)\mu(x,y)dx$$

for an appropriate function $F(\cdot)$. Moreover, $v_{k-1}(\tau,x,y)$ is a solution of

$$\frac{\partial}{\partial\tau}v_{k-1}(\tau,x,y) = \mathcal{L}_1(x,y)v_{k-1}(\tau,x,y) + \mathcal{L}_2(x,y)v_{k-2}(\tau,x,y) + c(x,y)v_{k-2}(\tau,x,y)$$

such that

$$|v_{k-1}(\tau,x,y)| \le c_1\exp(-c_2\tau) \tag{3.47}$$

and

$$\left|\frac{\partial^{|\nu|}v_{k-1}(\tau,x,y)}{\partial y_1^{\nu_1}\cdots\partial y_d^{\nu_d}}\right| \le c_1\exp(-c_2\tau), \quad |\nu| = 1,\dots,4. \tag{3.48}$$

We proceed to obtain $u_k(t,x,y)$ and $v_k(\tau,x,y)$. Multiplying

$$\begin{aligned}&\frac{\partial}{\partial t}u_{k-1}(t,x,y)\\ &\quad = \mathcal{L}_1(x,y)u_k(t,x,y) + \mathcal{L}_2(x,y)u_{k-1}(t,x,y) + c(x,y)u_{k-1}(t,x,y)\end{aligned} \tag{3.49}$$

by $\mu(x,y)$, integrating with respect to x, and noting that $\mathcal{L}_1(x,y)u_k(t,x,y)$ is orthogonal to $\mu(x,y)$, we obtain

$$\overline{\frac{\partial}{\partial t}u_{k-1}(t,\cdot,y)} = \overline{\mathcal{L}_2(\cdot,y)u_{k-1}(t,\cdot,y)} + \overline{c(\cdot,y)u_{k-1}(t,\cdot,y)}.$$

Next, in view of (3.2), we obtain

$$\mathcal{L}_1(x,y)u_k(t,x,y) = \Psi_{k-1}(t,x,y), \tag{3.50}$$

where

$$\begin{aligned}\Psi_{k-1}(t,x,y) = \frac{\partial}{\partial t}u_{k-1}(t,x,y) - \overline{\frac{\partial}{\partial t}u_{k-1}(t,\cdot,y)}\\ + \overline{\mathcal{L}_2(x,y)u_{k-1}(t,\cdot,y)} - \mathcal{L}_2(x,y)u_{k-1}(t,x,y)\\ + \overline{c(\cdot,y)u_{k-2}(t,\cdot,y)} - c(x,y)u_{k-2}(t,x,y).\end{aligned}$$

It is readily seen that $\Psi_{k-1}(t,x,y)$ is orthogonal to $\mu(x,y)$. It follows from Lemmas 3.1 and 3.2 that (3.2) has a solution of the form

$$u_k(t,x,y) = U_k(t,y) + \widetilde{u}_k(t,x,y), \tag{3.51}$$

where $\widetilde{u}_k(t,x,y)$ is a particular solution of (3.50) satisfying

$$\int_{\mathbf{K}_r} \widetilde{u}_k(t,x,y)\mu(x,y)dx = 0. \tag{3.52}$$

To ensure the desired match, consider the solution of the initial layer correction

$$\begin{aligned} \frac{\partial}{\partial\tau}v_k(\tau,x,y) &= \mathcal{L}_1(x,y)v_k(\tau,x,y) + \mathcal{L}_2(x,y)v_{k-1}(\tau,x,y) + c(x,y)v_{k-1}(\tau,x,y),\\ v_k(0,x,y) &= -u_k(0,x,y). \end{aligned} \tag{3.53}$$

We proceed with the estimate on $v_k(\tau,x,y)$. By requiring $v_k(\tau,x,y)\to 0$ as $\tau\to\infty$, using (3.47), (3.48), and Lemma 3.4, we obtain

$$\begin{aligned} &\int_{\mathbf{K}_r} v_k(0,x,y)\mu(x,y)dx + \int_{\mathbf{K}_r}\mathcal{L}_2(x,y)\left[\int_0^\infty v_{k-1}(s,x,y)ds\right]\mu(x,y)dx\\ &\quad + \int_{\mathbf{K}_r} c(x,y)\left(\int_0^\infty v_{k-1}(s,x,y)ds\right)\mu(x,y)dx = 0, \end{aligned} \tag{3.54}$$

and $|v_k(\tau,x,y)|\le c_1\exp(-c_2\tau)$. Next, similar to the proof of Lemma 3.6, define $\widehat{v}_k(\tau,x,y) = (\partial/\partial y)v_k(\tau,x,y)$. We obtain from (3.53) that

$$\frac{\partial}{\partial\tau}\widehat{v}_k(\tau,x,y) = \mathcal{L}_1(x,y)\widehat{v}_k(\tau,x,y) + \widehat{F}_k(\tau,x,y),$$

where

$$\begin{aligned} \widehat{F}_k(\tau,x,y) = &\left(\frac{\partial}{\partial y}\mathcal{L}_1(x,y)\right)v_k(\tau,x,y) + \frac{\partial}{\partial y}\left(\mathcal{L}_2(x,y)v_{k-1}(\tau,x,y)\right)\\ &+ \frac{\partial}{\partial y}\left(c(x,y)v_{k-1}(\tau,x,y)\right). \end{aligned}$$

Using similar estimates as in Lemma 3.6,

$$|(\partial/\partial y)v_{k-1}(\tau,x,y)|\le c_1\exp(-c_2\tau).$$

By induction hypothesis,

$$\begin{aligned} \left|\frac{\partial}{\partial y}\left(\mathcal{L}_2(x,y)v_{k-1}(\tau,x,y)\right)\right| &\le c_1\exp(-c_2\tau),\\ \left|\frac{\partial}{\partial y}\left(c(x,y)v_{k-1}(\tau,x,y)\right)\right| &\le c_1\exp(-c_2\tau). \end{aligned}$$

Thus $\widehat{F}_k(\tau,x,y)$ decays exponentially fast. An application of Lemma 3.4 then yields that $|(\partial/\partial y)v_k(\tau,x,y)|\le c_1\exp(-c_2\tau)$. Similar to the proof of Lemma 3.6, we obtain the following proposition.

PROPOSITION 3.8. *For $k=2,\dots,n+1$,*

$$|v_k(\tau,x,y)|\le c_1\exp(-c_2\tau) \tag{3.55}$$

and

$$\left|\frac{\partial^{|\nu|}v_k(\tau,x,y)}{\partial y_1^{\nu_1}\cdots\partial y_d^{\nu_d}}\right| \le c_1\exp(-c_2\tau), \quad |\nu|=1,\dots,4. \tag{3.56}$$

Using (3.53), rewrite (3.54) as

$$\begin{aligned}\int_{\mathbf{K}_r} u_k(0,x,y)\mu(x,y)dx &= \int_0^\infty ds\int_{\mathbf{K}_r}\mathcal{L}_2(x,y)v_{k-1}(s,x,y)\mu(x,y)dx\\ &\quad+\int_0^\infty ds\int_{\mathbf{K}_r} c(x,y)v_{k-1}(s,x,y)\mu(x,y)dx.\end{aligned} \tag{3.57}$$

Owing to (3.51), we have

$$\begin{aligned}&\frac{\partial}{\partial t}U_k(t,y)+\frac{\partial}{\partial t}\widetilde{u}_k(t,x,y)\\ &\quad=\mathcal{L}_1(x,y)u_{k+1}(t,x,y)+\mathcal{L}_2(x,y)U_k(t,y)+\mathcal{L}_2(x,y)\widetilde{u}_k(t,x,y)\\ &\qquad+c(x,y)U_k(t,y)+c(x,y)\widetilde{u}_k(t,x,y).\end{aligned}$$

Again, multiplying through by $\mu(x,y)$ and integrating with respect to $x\in\mathbf{K}_r$, we arrive at

$$\frac{\partial}{\partial t}U_k(t,y)=\overline{\mathcal{L}}_2(y)U_k(t,y)+\overline{c}(y)U_k(t,y)+\widetilde{\Psi}_k(t,y), \tag{3.58}$$

where $\widetilde{\Psi}_k(t,y)$ is a known function given by

$$\widetilde{\Psi}_k(t,y)\stackrel{\text{def}}{=}\int_{\mathbf{K}_r}\mathcal{L}_2(x,y)\widetilde{u}_k(t,x,y)\mu(x,y)dx+\int_{\mathbf{K}_r}c(x,y)\widetilde{u}_k(t,x,y)\mu(x,y)dx. \tag{3.59}$$

In the above, we used $\int_{\mathbf{K}_r}\frac{\partial}{\partial t}\widetilde{u}_k(t,x,y)\mu(x,y)dx=0$.

Using $t=0$ in (3.51),

$$U_k(0,y)=u_k(0,x,y)-\widetilde{u}_k(0,x,y).$$

Multiplying through by the invariant density $\mu(x,y)$ and integrating over $\mathbf{K}_r$, together with (3.57), we have

$$\begin{aligned}U_k(0,y)&=\int_{\mathbf{K}_r}u_k(0,x,y)\mu(x,y)dx\\ &=-\int_{\mathbf{K}_r}v_k(0,x,y)\mu(x,y)dx\\ &=\int_0^\infty ds\int_{\mathbf{K}_r}\mathcal{L}_2(x,y)v_{k-1}(s,x,y)\mu(x,y)dx\\ &\quad+\int_0^\infty\int_{\mathbf{K}_r}c(x,y)v_{k-1}(s,x,y)\mu(x,y)dx.\end{aligned} \tag{3.60}$$

Thus, $u_k(t,x,y)$ is determined; so is $v_k(\tau,x,y)$. We record this into the following theorem.

THEOREM 3.9. *Assume condition* (A). *Then for $i=0,\dots,n+1$, the functions $u_i(\cdot)$ and $v_i(\cdot)$ in the formal asymptotic expansions* (2.8) *can be constructed such that $u_i(\cdot)$ are continuously differentiable with respect to t and twice continuously differentiable with respect to x and y and such that $|v_i(\tau,x,y)|\le c_1\exp(-c_2\tau)$.*

Remark 3.10. For the asymptotic expansions, we need only $k=n$. The extra terms $u_{n+1}(t,x,y)$ and $v_{n+1}(\tau,x,y)$ are needed for the error estimates.

4. Error bounds. For a suitable smooth function $h(\cdot)$, define the operator $L_\varepsilon(t,x,y)$ as

$$\begin{aligned} L_\varepsilon(t,x,y)h(t,x,y) &= \frac{\partial}{\partial t}h(t,x,y) - \frac{1}{\varepsilon}\mathcal{L}_1(x,y)h(t,x,y) \\ &\quad - \mathcal{L}_2(x,y)h(t,x,y) - c(x,y)h(t,x,y). \end{aligned} \tag{4.1}$$

Note that $u_\varepsilon(t,x,y)$ as a solution of (2.4) is the same as $L_\varepsilon(t,x,y)u_\varepsilon(t,x,y) = f(x,y)$. Recall that the estimation error sequence $\{e_{\varepsilon,k}(t,x,y)\}$ was defined in (2.9). Note that for each $0 \le k \le n$,

$$\begin{aligned} L_\varepsilon(t,x,y)e_{\varepsilon,k}(t,x,y) = L_\varepsilon(t,x,y)\Bigg(& u_0(t,y) + \sum_{i=1}^{k}\varepsilon^i u_i(t,x,y) \\ & + \sum_{i=0}^{k}\varepsilon^i v_i(t/\varepsilon,x,y) - u_\varepsilon(t,x,y)\Bigg). \end{aligned} \tag{4.2}$$

We proceed to establish the following lemma.

LEMMA 4.1. *Under conditions in* (A)*, for each* $0 \le k \le n+1$,

$$\sup_{(t,x,y)\in[0,T]\times\mathbf{K}} |L_\varepsilon(t,x,y)e_{\varepsilon,k}(t,x,y)| = O(\varepsilon^k).$$

Proof. Let us begin with $k = 0$. By virtue of (3.1) and (3.3), it is easily seen that

$$\begin{aligned} & L_\varepsilon(t,x,y)e_{\varepsilon,0}(t,x,y) \\ &= L_\varepsilon(t,x,y)u_0(t,y) + L_\varepsilon(t,x,y)v_0\left(\frac{t}{\varepsilon},x,y\right) - L_\varepsilon(t,x,y)u_\varepsilon(t,x,y) \\ &= \frac{\partial}{\partial t}u_0(t,y) - \frac{1}{\varepsilon}\mathcal{L}_1(x,y)u_0(t,y) - \mathcal{L}_2(x,y)u_0(t,y) - c(x,y)u_0(t,y) - f(x,y) \\ &\quad + \frac{1}{\varepsilon}\left[\frac{\partial}{\partial\tau}v_0(\tau,x,y) - \mathcal{L}_1(x,y)v_0(\tau,x,y)\right] - \mathcal{L}_2(x,y)v_0(\tau,x,y) - c(x,y)v_0(\tau,x,y) \\ &= \frac{\partial}{\partial t}u_0(t,y) - \mathcal{L}_2(x,y)u_0(t,y) - c(x,y)u_0(t,y) \\ &\quad - f(x,y) - \mathcal{L}_2(x,y)v_0(\tau,x,y) - c(x,y)v_0(\tau,x,y). \end{aligned}$$

Note that the smoothness of $u_0(t,y)$ and that of the coefficients of $\mathcal{L}_1(x,y)$ and $\mathcal{L}_2(x,y)$ imply

$$\begin{gathered} \sup_{(t,x,y)\in[0,T]\times\mathbf{K}}\left|\frac{\partial}{\partial t}u_0(t,y)\right| = O(1), \quad \sup_{(t,x,y)\in[0,T]\times\mathbf{K}}|\mathcal{L}_2(x,y)u_0(t,y)| = O(1), \\ \sup_{(x,y)\in\mathbf{K}}|c(x,y)| = O(1), \qquad \text{and} \qquad \sup_{(x,y)\in\mathbf{K}}|f(x,y)| = O(1). \end{gathered}$$

By Lemma 3.6,

$$|v_0(\tau,x,y)| \le c_1\exp(-c_2\tau), \qquad |\mathcal{L}_2(x,y)v_0(\tau,x,y)| \le c_1\exp(-c_2\tau);$$

it then follows that

$$\sup_{(t,x,y)\in[0,T]\times\mathbf{K}}[|v_0(\tau,x,y)| + |\mathcal{L}_2(x,y)v_0(\tau,x,y)|] = O(1).$$

Thus, we arrive at

$$\sup_{(t,x,y)\in[0,T]\times\mathbf{K}} |L_\varepsilon(t,x,y)e_{\varepsilon,0}(t,x,y)| = O(1).$$

Likewise, for any $1 \le k \le n+1$, using (3.1)–(3.4) and upon cancellation,

$$\begin{aligned}
&L_\varepsilon(t,x,y)e_{\varepsilon,k}(t,x,y)\\
&\quad= L_\varepsilon(t,x,y)u_0(t,x,y) + \sum_{i=1}^{k}\varepsilon^i L_\varepsilon(t,x,y)u_i(t,x,y)\\
&\qquad+\sum_{i=0}^{k}\varepsilon^i L_\varepsilon(t,x,y)v_i(\tau,x,y) - L_\varepsilon(t,x,y)u_\varepsilon(t,x,y)\\
&\quad= \frac{\partial}{\partial t}u_0(t,y) - \mathcal{L}_2(x,y)u_0(t,y) - c(x,y)u_0(t,y) - f(x,y)\\
&\qquad+\sum_{i=1}^{k}\varepsilon^i\Big[\frac{\partial}{\partial t}u_i(t,x,y) - \frac{1}{\varepsilon}\mathcal{L}_1(x,y)u_i(t,x,y)\\
&\qquad\qquad -\mathcal{L}_2(x,y)u_i(t,x,y) - c(x,y)u_i(t,x,y)\Big]\\
&\qquad+\sum_{i=0}^{k}\varepsilon^{i-1}\Big[\frac{\partial}{\partial \tau}v_i(\tau,x,y) - \mathcal{L}_1(x,y)v_i(\tau,x,y)\\
&\qquad\qquad -\varepsilon\mathcal{L}_2(x,y)v_i(\tau,x,y) - \varepsilon c(x,y)v_i(\tau,x,y)\Big]\\
&\quad= \varepsilon^k\frac{\partial}{\partial t}u_k(t,x,y) - \varepsilon^k c(x,y)u_k(t,x,y) - \varepsilon^k\mathcal{L}_2(x,y)u_k(t,x,y)\\
&\qquad-\varepsilon^k\mathcal{L}_2(x,y)v_k(\tau,x,y) - \varepsilon^k c(x,y)v_k(\tau,x,y).
\end{aligned}$$

Then (3.47), (3.48), and the smoothness of the coefficients of $\mathcal{L}_2(x,y)$, $c(x,y)$, and $(\partial/\partial t)u_k(t,x,y)$ yield

$$\begin{aligned}
\sup_{(t,x,y)\in[0,T]\times\mathbf{K}}\Big|\varepsilon^k\frac{\partial}{\partial t}u_k(t,x,y)\Big| &= O(\varepsilon^k),\\
\sup_{(t,x,y)\in[0,T]\times\mathbf{K}}\big|\varepsilon^k c(x,y)u_k(t,x,y)\big| &= O(\varepsilon^k),\\
\sup_{(t,x,y)\in[0,T]\times\mathbf{K}}\big|\varepsilon^k\mathcal{L}_2(x,y)u_k(t,x,y)\big| &= O(\varepsilon^k).
\end{aligned}$$

By virtue of Lemma 3.8,

$$\begin{aligned}
\sup_{(t,x,y)\in[0,T]\times\mathbf{K}}\big|\varepsilon^k c(x,y)v_k(\tau,x,y)\big| &= O(\varepsilon^k),\\
\sup_{(t,x,y)\in[0,T]\times\mathbf{K}}\big|\varepsilon^k\mathcal{L}_2(x,y)v_k(\tau,x,y)\big| &= O(\varepsilon^k).
\end{aligned}$$

Thus,

$$\sup_{(t,x,y)\in[0,T]\times\mathbf{K}} |L_\varepsilon(t,x,y)e_{\varepsilon,k}(t,x,y)| = O(\varepsilon^k) \quad \text{for } k \le n+1.$$

The proof of the lemma is concluded. □

Next, we derive a lemma, which is an estimate on a nonhomogeneous Cauchy problem.

LEMMA 4.2. *Suppose that there is a function $\eta_\varepsilon(t,x,y)$ such that for $k \le n+1$,*

$$\sup_{(t,x,y)\in[0,T]\times\mathbf{K}} |\eta_\varepsilon(t,x,y)| = O(\varepsilon^k).$$

Then the solution of the Cauchy problem

$$L_\varepsilon(t,x,y)\zeta_\varepsilon(t,x,y) = \eta_\varepsilon(t,x,y), \qquad \zeta_\varepsilon(0,x,y) = 0, \tag{4.3}$$

satisfies $\sup_{(t,x,y)\in[0,T]\times\mathbf{K}} |\zeta_\varepsilon(t,x,y)| = O(\varepsilon^k)$.

Proof. Let $Z_\varepsilon^{(x,y)}(t) = (X_\varepsilon^x(t), Y_\varepsilon^y(t))$ be the diffusion process whose generator is given by $\mathcal{L}_1(x,y)/\varepsilon + \mathcal{L}_2(x,y)$. Using the known probabilistic representation for the solution of (4.3) (see, e.g., [10, Lemma 1], [8], [16]), we can obtain the upper bounds

$$\sup_{(t,x,y)\in[0,T]\times\mathbf{K}} |\zeta_\varepsilon(t,x,y)| \le Te^{T\widetilde{C}}, \quad \sup_{(t,x,y)\in[0,T]\times\mathbf{K}} |\eta_\varepsilon(t,x,y)| = O(\varepsilon^k),$$

where $\widetilde{C} = \sup_{(x,y)\in\mathbf{K}} |c(x,y)|$. The lemma thus follows. □

Now we are in a position to obtain the desired error bounds. The result is stated in the following theorem.

THEOREM 4.3. *Assume* (A) *holds. Then for the asymptotic expansions constructed in Theorem* 3.9, $\sup_{(t,x,y)\in[0,T]\times\mathbf{K}} |e_{\varepsilon,n}(t,x,y)| = O(\varepsilon^{n+1})$.

Proof. By virtue of Lemma 4.1, $\sup_{(t,x,y)\in[0,T]\times\mathbf{K}} |L_\varepsilon(t,x,y)e_{\varepsilon,n+1}(t,x,y)| = O(\varepsilon^{n+1})$. An application of Lemma 4.2 leads to $\sup_{(t,x,y)\in[0,T]\times\mathbf{K}} |e_{\varepsilon,n+1}(t,x,y)| = O(\varepsilon^{n+1})$. The smoothness of $u_{n+1}(\cdot)$ implies that

$$\sup_{(t,x,y)\in[0,T]\times\mathbf{K}} |\varepsilon^{n+1}u_{n+1}(t,x,y)| = O(\varepsilon^{n+1}).$$

By the exponential decay of $v_{n+1}(t/\varepsilon,x,y)$, it is bounded, and hence

$$\sup_{(t,x,y)\in[0,T]\times\mathbf{K}} |\varepsilon^{n+1}v_{n+1}(t/\varepsilon,x,y)| = O(\varepsilon^{n+1}).$$

Since

$$e_{\varepsilon,n+1}(t,x,y) = e_{\varepsilon,n}(t,x,y) + \varepsilon^{n+1}u_{n+1}(t,x,y) + \varepsilon^{n+1}v_{n+1}(t/\varepsilon,x,y), \tag{4.4}$$

we obtain, from (4.4),

$$\begin{aligned}
&\sup_{(t,x,y)\in[0,T]\times\mathbf{K}} |e_{\varepsilon,n}(t,x,y)| \\
&\quad = \sup_{(t,x,y)\in[0,T]\times\mathbf{K}} \left|\varepsilon^{n+1}u_{n+1}(t,x,y) + \varepsilon^{n+1}v_{n+1}(t/\varepsilon,x,y) - e_{\varepsilon,n+1}(t,x,y)\right| \\
&\quad = O(\varepsilon^{n+1}).
\end{aligned}$$

The theorem is proved. □

5. Interpretation, examples, and extensions.

5.1. Discussion. The asymptotic expansions obtained in this paper provide some new insight even for the leading term $u_0(\cdot)$. In the literature (see [11, 21, 26]), only asymptotic behavior of the slow component $Y_\varepsilon(t)$ was usually considered. By using our asymptotic expansions, results concerning the fast component also hold. We give some probabilistic interpretation of our theorem in what follows.

Suppose that $\varphi(x,y)$ is any sufficiently smooth function. Denote by $Y^y(t)$ the Markov diffusion process on $\mathbf{K}_d$ with the generator $\overline{\mathcal{L}}_2(y)$ defined in (3.18). Assume that $c(x,y) = f(x,y) \equiv 0$ in (2.4). Then it follows from Theorem 4.3 and the probabilistic interpretation of the solution of (2.4), for $t > c\varepsilon \ln(1/\varepsilon)$, that

$$\lim_{\varepsilon\to 0} E\varphi(X_\varepsilon^{x,y}(t), Y_\varepsilon^{x,y}(t)) = E^y\overline{\varphi}(Y^y(t)) = E\int_{\mathbf{K}_r} \varphi(x, Y^y(t))\mu(x, Y^y(t))dx. \tag{5.1}$$

Approximating $\varphi(x,y)$ by the product of the indicator functions $\chi^x(A)$ and $\chi^y(B)$, we conclude that

$$\lim_{\varepsilon\to 0} P(X_\varepsilon^{x,y}(t) \in A, Y_\varepsilon^{x,y}(t) \in B) = \int_A dx \int_B P(Y^y(t) \in dz)\mu(x,z). \tag{5.2}$$

In particular, for $B = \mathsf{Y}$, the entire space containing the range of $Y(\cdot)$, we have

$$P(X_\varepsilon^{x,y}(t) \in A) \to \int_A E\mu(x, Y^y(t))dx \quad \text{as } \varepsilon \to 0.$$

Thus the limit behavior of finite-dimensional distributions can be deduced from (5.1) and (5.2). In fact, using the Markov property of $(X_\varepsilon(\cdot), Y_\varepsilon(\cdot))$ and $Y(\cdot)$, for any t_1 and t_2 not depending on ε with $0 < t_1 < t_2$, from (5.2), we have

$$\begin{aligned}
&P(X_\varepsilon^{x,y}(t_1) \in A_1, X_\varepsilon^{x,y}(t_2) \in A_2)\\
&\quad = \int_{A_1}\int_{\mathsf{Y}} P(X_\varepsilon^{x,y}(t) \in dx_1, Y_\varepsilon^{x,y}(t_1) \in dy_1)P(X_\varepsilon^{x_1,y_1}(t_2 - t_1) \in A_2)\\
&\quad \to \int_{A_1}\int_{\mathsf{Y}} P(Y^y(t_1) \in dy_1)\mu(x_1,y_1)dx_1 \int_{A_2} E\mu(z, Y^{y_1}(t_2 - t_1))dz \quad \text{as } \varepsilon \to 0\\
&\quad = \int_{A_1}\int_{A_2} E\mu(x_1, Y^y(t_1))E\mu(z, Y^{Y^y(t_1)}(t_2 - t_1))dx_1dz\\
&\quad = \int_{A_1}\int_{A_2} E[\mu(x_1, Y^y(t_1))E\mu(z, Y^y(t_2))|Y^y(t_1)]dx_1dz\\
&\quad = E\int_{A_1}\int_{A_2} \mu(x_1, Y^y(t_1))\mu(z, Y^y(t_2))dx_1dz.
\end{aligned}$$

Completely analogously, one obtains the convergence of the finite-dimensional distributions. We record this in the following proposition.

PROPOSITION 5.1. *For any t_i, $i \le n$, independent of ε satisfying $0 < t_1 < t_2 < \cdots < t_n$,*

$$\begin{aligned}
&\lim_{\varepsilon\to 0} P\left(X_\varepsilon^{x,y}(t_1) \in A_1, \ldots, X_\varepsilon^{x,y}(t_n) \in A_n\right)\\
&\qquad = E\int_{A_1}\cdots\int_{A_n} \mu(x_1, Y^y(t_1))\mu(x_2, Y^y(t_2))\cdots\mu(x_n, Y^y(t_n))dx_1dx_2\cdots dx_n.
\end{aligned}$$

5.2. Examples.

Example 5.2. Let us consider the problem of (2.4) with

$$f(x,y)=0, \qquad \varphi(x,y)=1, \qquad c(x,y)=\iota s c_0(x,y),$$

where $\iota=\sqrt{-1}$. In this case, the solution of (2.4) is

$$u_\varepsilon(t,x,y)=E\exp\left(\iota s\int_0^t c_0(X_\varepsilon^{x,y}(\zeta),Y_\varepsilon^{x,y}(\zeta))d\zeta\right).$$

This is the well-known Feinman–Kac formula. As $\varepsilon\to 0$, using our asymptotic expansions,

$$u_\varepsilon(t,x,y)\to E\exp\left(\iota s\int_0^t \bar{c}_0(Y^y(\zeta))d\zeta\right).$$

Thus, we have proved the following assertion.

PROPOSITION 5.3. *As $\varepsilon\to 0$, the functional $\int_0^t c_0(X_\varepsilon^{x,y}(\zeta),Y_\varepsilon^{x,y}(\zeta))d\zeta$ converges in distribution to $\int_0^t\int_{\mathbf{K}_r} c_0(x,Y^y(\zeta))\mu(x,Y^y(\zeta))dxd\zeta$.*

Example 5.4. Consider a Markov process on a two-dimensional torus having a fast component and a slow component. The fast component is a Brownian motion on a circle of length 1, and the generator of the slow component is given by $b_2(x,y)\frac{\partial}{\partial y}+a_2(x,y)\frac{\partial^2}{\partial y^2}$. Then it follows from (5.1) that

$$E\varphi(X_\varepsilon(t),Y_\varepsilon(t))\to E\int_0^1\varphi(x,Y(t))dx=E\overline{\varphi(\cdot,Y(t))}. \tag{5.3}$$

Let us assume that $\overline{\varphi(\cdot,y)}=0$. A natural question arises: What is the main term in the asymptotic expansion of the solution of (2.4) in this case?

Consider the Cauchy problem

$$\frac{\partial u_\varepsilon}{\partial t}=\frac{1}{2\varepsilon}\frac{\partial^2 u_\varepsilon}{\partial x^2}+b_2(x,y)\frac{\partial u_\varepsilon}{\partial y}+a_2(x,y)\frac{\partial^2 u_\varepsilon}{\partial y^2}, \qquad u_\varepsilon(0,x,y)=\varphi(x,y), \tag{5.4}$$

on the surface of the two-dimensional torus, where $b_2(\cdot)$ and $a_2(\cdot)$ are periodic in x and y with period 1. Assuming $\varphi(x,y)$ to be sufficiently smooth and

$$\int_0^1\varphi(x,y)dx=0 \quad \text{for all } y\in[0,1] \tag{5.5}$$

yields that $\overline{\varphi}(\cdot,y)=0$. It follows from (3.16) and (3.17) that $u_0(t,y)=0$. So the first nonzero outer expansion term in the asymptotic expansion is $\varepsilon u_1(t,x,y)$. It is clear from (3.34) and (3.35) that $\widetilde{u}_1(t,x,y)=0$. So the main term in the asymptotic expansion is $\varepsilon U_1(t,y)$. Note that $v_0(\tau,x,y)$ is a periodic in x with period 1 solution of

$$\frac{\partial}{\partial\tau}v_0(\tau,x,y)=\frac{1}{2}\frac{\partial^2}{\partial x^2}v_0(\tau,x,y), \qquad v_0(0,x,y)=\varphi(x,y). \tag{5.6}$$

Using the Fourier expansions, we have

$$\varphi(x,y)=\sum_{k=1}^\infty\left(\widetilde{\phi}_k(y)\sin(2\pi kx)+\widehat{\phi}_k(y)\cos(2\pi kx)\right). \tag{5.7}$$

We can then write the solution of (5.6) as

$$v_0(\tau,x,y)=\sum_{k=1}^{\infty}(\widetilde{\phi}_k(y)\sin(2\pi kx)+\widehat{\phi}_k(y)\cos(2\pi kx))e^{-2\pi^2k^2\tau}. \tag{5.8}$$

It follows from (3.42) that

$$U_1(0,y)=\int_0^{\infty}ds\int_0^1\mathcal{L}_2(x,y)v_0(s,x,y)dx. \tag{5.9}$$

We introduce the following notation:

$$\begin{aligned}
b_{2,k}^{(1)}(y)&=\int_0^1 b_2(x,y)\sin(2\pi kx)dx,\\
b_{2,k}^{(2)}(y)&=\int_0^1 b_2(x,y)\cos(2\pi kx)dx,\\
a_{2,k}^{(1)}(y)&=\int_0^1 a_2(x,y)\sin(2\pi kx)dx,\\
a_{2,k}^{(2)}(y)&=\int_0^1 a_2(x,y)\cos(2\pi kx)dx.
\end{aligned} \tag{5.10}$$

Then we have, from (5.8),

$$\begin{aligned}
&\int_0^1\mathcal{L}_2(x,y)v_0(s,x,y)dx\\
&\quad=\sum_{k=1}^{\infty}e^{-2\pi^2k^2s}\left[b_{2,k}^{(1)}(y)\widetilde{\phi}_k'(y)+b_{2,k}^{(2)}(y)\widehat{\phi}_k'(y)+a_{2,k}^{(1)}(y)\widetilde{\phi}_k''(y)+a_{2,k}^{(2)}(y)\widehat{\phi}_k''(y)\right].
\end{aligned} \tag{5.11}$$

It follows from (5.9) and (5.11) that

$$\begin{aligned}
&U_1(0,y)\\
&\quad=\sum_{k=1}^{\infty}\frac{1}{2\pi^2k^2}\left[b_{2,k}^{(1)}(y)\widetilde{\phi}_k'(y)+b_{2,k}^{(2)}(y)\widehat{\phi}_k'(y)+a_{2,k}^{(1)}(y)\widetilde{\phi}_k''(y)+a_{2,k}^{(2)}(y)\widehat{\phi}_k''(y)\right],
\end{aligned} \tag{5.12}$$

where $\widetilde{\phi}_k'(y)$, $\widehat{\phi}_k'(y)$, $\widetilde{\phi}_k''(y)$, and $\widehat{\phi}_k''(y)$ are the first and the second order derivatives of $\widetilde{\phi}_k(y)$ and $\widehat{\phi}_k(y)$, respectively. The initial condition (5.12) can be written in a more elegant form with the help of the following lemma.

LEMMA 5.5. *Let $g(x)$ for $x\in[0,1]$ have a Fourier expansion*

$$g(x)=\sum_{k=1}^{\infty}\left(g_k^{(1)}\sin(2\pi kx)+g_k^{(2)}\cos(2\pi kx)\right). \tag{5.13}$$

Then

$$\begin{aligned}
&\frac{1}{2\pi^2}\sum_{k=1}^{\infty}\frac{1}{k^2}\left(g_k^{(1)}\sin(2\pi kx)+g_k^{(2)}\cos(2\pi kx)\right)\\
&\quad=\int_0^1(1-z)^2g(z)dz+(1-2x)\int_0^1 zg(z)dz-2\int_0^x g(z)(x-z)dz.
\end{aligned} \tag{5.14}$$

Proof. The proof uses integration by parts and is straightforward; we omit the details here. □

With $j = 1, 2$, $\varphi^{(j)}(x, y) = (\partial^j/\partial y^j)\varphi(x, y)$, we have, from (5.7),

$$\varphi^{(j)}(x, y) = \sum_{k=1}^{\infty} \left(\widetilde{\varphi}_k^{(j)}(y) \sin(2\pi kx) + \widehat{\varphi}_k^{(j)}(y) \cos(2\pi kx) \right). \tag{5.15}$$

Thus, from (5.14) and (5.15),

$$\begin{aligned} &\frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{1}{k^2} \left(\widetilde{\varphi}_k^{(j)}(y) \sin(2\pi kx) + \widehat{\varphi}_k^{(j)}(y) \cos(2\pi kx) \right) \\ &\quad = \int_0^1 (1-z)^2 \varphi^{(j)}(z, y) dz + (1-2x) \int_0^1 z\varphi^{(j)}(z, y) dz - 2 \int_0^x \varphi^{(j)}(z, y)(x-z) dz. \end{aligned}$$

Therefore, this equation and (5.12) lead to

$$\begin{aligned} U_1(0, y) &= 2 \int_0^1 (1-z)^2 dz \int_0^1 \mathcal{L}_2(x, y)\varphi(z, y) dx \\ &\quad + 2 \int_0^1 z dz \int_0^1 (1-2x)\mathcal{L}_2(x, y)\varphi(z, y) dx \\ &\quad - 4 \int_0^1 dz \int_z^1 \mathcal{L}_2(x, y)\varphi(z, y)(x-z) dx. \end{aligned} \tag{5.16}$$

What we have proven is the following proposition.

PROPOSITION 5.6. *For the solution of the initial value problem* (5.4) *with initial condition* $\varphi(x, y)$ *satisfying* (5.5) *and* $t > c\varepsilon \ln(1/\varepsilon)$ *for sufficiently large* c,

$$u_\varepsilon(t, x, y) = \varepsilon U_1(t, y) + O(\varepsilon^2),$$

where $U_1(t, y)$ *is the solution of*

$$\begin{cases} \dfrac{\partial}{\partial t} U_1(t, y) = \overline{\mathcal{L}}_2(y) U_1(t, y), \\ U_1(0, y) \quad \textit{satisfies } (5.16). \end{cases}$$

5.3. Extensions. The results obtained can be extended to nonhomogeneous models and the inclusion of mixed derivatives in the operator $\mathcal{L}_2(x, y)$. We mention them in what follows.

(a) Certain nonstationarity may be added. In particular, the generators $\mathcal{L}_1(x, y)$ and $\mathcal{L}_2(x, y)$ may depend on time t as well. That is, we may consider (2.4) with the replacement of $\mathcal{L}_1(x, y)$ and $\mathcal{L}_2(x, y)$ by $\mathcal{L}_1(t, x, y)$ and $\mathcal{L}_2(t, x, y)$. In this case, we assume that the coefficients of $\mathcal{L}_i(t, x, y)$ are $n + 2$ times continuously differentiable in t in addition to the conditions used in (A). One takes Taylor expansions of the generators about the point $t = 0$. That is, we will need to use

$$\mathcal{L}_i(\varepsilon\tau, x, y) = \sum_{k=0}^{n+1} \frac{(\varepsilon\tau)^k}{k!} \frac{\partial^k}{\partial t^k} \mathcal{L}_i(0, x, y) + O((\varepsilon\tau)^{n+2}) \quad \text{for } i = 1, 2.$$

Then we use this to find the initial layer terms (see the use of the Taylor expansions in the forward equations [13, 14]) and proceed as in the previous case.

(b) Mixed derivatives may be included in the operator. In lieu of (2.5), with $\mathcal{L}_1(x,y)$ and $\mathcal{L}_2(x,y)$ defined as before, one may consider

$$\begin{cases} \dfrac{\partial u_\varepsilon}{\partial t} = \left(\dfrac{1}{\varepsilon}\mathcal{L}_1(x,y) + \dfrac{1}{\sqrt{\varepsilon}}\mathcal{L}_3(x,y) + \mathcal{L}_2(x,y)\right) u_\varepsilon + c(x,y)u_\varepsilon + f(x,y), \\ u_\varepsilon(0,x,y) = \varphi(x,y), \end{cases}$$

where

$$\mathcal{L}_3(x,y) = \sum_{i=1}^{r}\sum_{j=1}^{d} a_{2,ij}(x,y)\frac{\partial^2}{\partial x_i \partial y_j}.$$

The essential ideas remain the same, but the notation will be more complex.

6. Appendix

Proof of Lemma 3.1. Let $u(t,x)$ be the solution of

$$\frac{\partial u}{\partial t} = \mathcal{L}(x)u, \qquad u(0,x) = \psi(x).$$

Define $v(t,x) = \int_0^t u(s,x)ds$. Then $v(t,x)$ satisfies

$$\begin{cases} \dfrac{\partial}{\partial t}v(t,x) = \mathcal{L}(x)v(t,x) + \psi(x), \\ v(0,x) = 0. \end{cases} \tag{6.1}$$

In view of the stochastic representation, since $u(s,x) = E\psi(X^x(s))$, by using (3.6) and (3.7),

$$|u(s,x)| = |u(s,x) - \langle \psi, \mu\rangle\,| \le c_1 \exp(-c_2 s).$$

As a result, the integral $\int_0^\infty u(s,x)ds$ makes sense. Define

$$v(x) = \lim_{t\to\infty} v(t,x).$$

Due to (3.7), $(\partial/\partial t)v(t,x) \to 0$ as $t \to \infty$.

We need only show that the derivatives of $v(\cdot)$ exist. In fact,

$$v(x) = \int_0^\infty E\psi(X^x(s))ds = \left(\int_0^1 + \int_1^\infty\right) E\psi(X^x(s))ds = v_1(x) + v_2(x),$$

where

$$v_1(x) = \int_0^1 ds \int P(x,s,y)\psi(y)dy, \qquad v_2(x) = \int_1^\infty ds \int P(x,s,y)\psi(y)dy.$$

For $v_1(x)$, it is easily obtained that

$$\frac{\partial}{\partial x}v_1(x) = \int_0^1 \int \frac{\partial}{\partial x}P(x,s,y)\psi(y)dy.$$

As for $v_2(x)$, by virtue of the Chapman–Kolmogorov equation,

$$v_2(x) = \int_1^\infty ds \int\int P(x,1,z)P(z,s-1,y)dz\psi(y)dy.$$

Now, since $(\partial/\partial x)P(x,1,z)$ is bounded,

$$\frac{\partial}{\partial x}v_2(x) = \int_1^\infty ds \int\int \frac{\partial}{\partial x}P(x,1,z)P(z,s-1,y)dz\psi(y)dy.$$

Similarly, it can be shown that $(\partial^2/\partial x^2)v(x)$ exists. Passing the limit as $t \to \infty$ in (6.1), the desired result then follows. □

Proof of Lemma 3.2. Let $W(x)$ be a solution of (3.9), and $x_0 \in \mathbf{K}_r$. Then $W(x)$ is a solution of the Dirichlet problem in the punched region $\Omega_{x_0} = \mathbf{K}_r \cap N_\delta^c(x_0)$:

$$\mathcal{L}(x)\widetilde{W}(x) = 0, \quad x \in \mathbf{K}_r \cap N_\delta^c(x_0), \qquad \widetilde{W}(x) = W(x), \quad x \in \partial N_\delta(x_0), \tag{6.2}$$

where $N_\delta(x_0) = \{x : |x-x_0| < \delta\}$, $\partial N_\delta(x_0)$ denotes its boundary, and $N_\delta^c(x_0)$ denotes the complement of $N_\delta(x_0)$. It follows from the maximum principle that

$$\min_{x\in\partial N_\delta(x_0)} W(x) \le \min_{x\in\mathbf{K}_r\cap N_\delta^c(x_0)} W(x) \le \max_{x\in\mathbf{K}_r\cap N_\delta^c(x_0)} W(x) \le \max_{x\in\partial N_\delta(x_0)} W(x).$$

Letting δ shrink to 0 and using continuity of $W(x)$ at x_0 yield the desired result. □

Proof of Lemma 3.4. The solution of (3.13) admits a probabilistic representation (see, e.g., [16])

$$V(\tau,x) = E\psi(X^x(\tau)) + \int_0^\tau EF(s, X^x(\tau-s))ds,$$

where $X^x(t)$ is the diffusion process associated with the generator $\mathcal{L}(x)$ satisfying $X^x(0) = x$. It follows from (3.7) that

$$\begin{aligned}\left|\int_0^{\tau/2} dsEF(s,X^x(\tau-s)) - \int_0^{\tau/2} ds\int_{\mathbf{K}_r} F(s,x)\mu(x)dx\right| &\le c_1\frac{\tau}{2}\exp(-c_2\tau/2)\\ &\le c_1\exp(-c_2\tau),\end{aligned}$$

where we have used the convention that c_i are generic positive constants, whose values may change for different appearances. Note also that

$$\begin{aligned}&\left|\int_0^\tau dsEF(s,X^x(\tau-s)) - \int_0^\infty\int_{\mathbf{K}_r} F(s,x)\mu(x)dx\right|\\ &\quad= \left|\int_0^{\tau/2} dsEF(s,X^x(\tau-s)) - \int_0^{\tau/2} ds\int_{\mathbf{K}_r} F(s,x)\mu(x)dx\right.\\ &\qquad\left. + \int_{\tau/2}^\tau dsEF(s,X^x(\tau-s)) - \int_{\tau/2}^\infty ds\int_{\mathbf{K}_r} F(s,x)\mu(x)dx\right|\\ &\quad\le c_1e^{-c_2\tau/2} + \left|\int_{\tau/2}^\infty ds\int_{\mathbf{K}_r} F(s,x)\mu(x)dx\right| + \int_{\tau/2}^\infty |EF(s,X^x(\tau-s))|ds.\end{aligned} \tag{6.3}$$

The assertion of the lemma follows from (3.14) and (6.3). The proof of the lemma is concluded. □

Acknowledgments. We thank Nick Krylov for bringing to our attention reference [1]. We are grateful to the reviewers and the editors for valuable and detailed comments and suggestions.

REFERENCES

[1] M. S. Agranovich, *Elliptic operators on closed manifolds*, in Current Problems in Mathematics: Fundamental Directions, Vol. 63, Akad. Nauk SSR, Moscow, 1990, pp. 5–129.

[2] N. N. Bogoliubov and Y. A. Mitropolskii, *Asymptotic Methods in the Theory of Nonlinear Oscillations*, Gordon and Breach, New York, 1961.

[3] R. Burridge, G. Papanicolaou, P. Sheng, and B. White, *Probing a random medium with a pulse*, SIAM J. Appl. Math., 49 (1989), pp. 582–607.

[4] J. L. Doob, *Stochastic Processes*, Wiley, New York, 1990.

[5] G. Folland, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1976.

[6] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[7] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, Berlin, 2001.

[8] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, 2nd ed., Kodansha, Tokyo, North–Holland, Amsterdam, 1989.

[9] A. M. Il'in, *Matching of Asymptotic Expansions of Solutions of Boundary Value Problems*, Transl. Math. Monogr. 102, AMS, Providence, RI, 1992.

[10] R. Z. Khasminskii, *Principle of averaging for parabolic and elliptic differential equations and for Markov processes with small diffusion*, Theory Probab. Appl., 8 (1963), pp. 1–21.

[11] R. Z. Khasminskii, *On an averaging principle for Ito stochastic differential equations*, Kybernetika, 4 (1968), pp. 260–279.

[12] R. Z. Khasminskii, *Stochastic Stability of Differential Equations*, Sijthoff and Noordhoff, Alphen aan den Rijn, The Netherlands, 1980.

[13] R. Z. Khasminskii and G. Yin, *Asymptotic series for singularly perturbed Kolmogorov–Fokker–Planck equations*, SIAM J. Appl. Math., 56 (1996), pp. 1766–1793.

[14] R. Z. Khasminskii and G. Yin, *On transition densities of singularly perturbed diffusions with fast and slow components*, SIAM J. Appl. Math., 56 (1996), pp. 1794–1819.

[15] R. Z. Khasminskii and G. Yin, *Asymptotic behavior of parabolic equations arising from null-recurrent diffusions*, J. Differential Equations, 161 (2000), pp. 154–173.

[16] N. V. Krylov, *Introduction to Diffusion Processes*, AMS, Providence, RI, 1994.

[17] N. V. Krylov, *Lectures on Elliptic and Parabolic Equations in Hölder Spaces*, AMS, Providence, RI, 1996.

[18] H. J. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhäuser Boston, Boston, MA, 1990.

[19] O. A. Ladyzhenskaia, V. A. Solonnikov, and N. N. Ural'tseva, *Linear and Quasi-Linear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.

[20] W. A. Massey and W. Whitt, *Uniform acceleration expansions for Markov chains with time-varying rates*, Ann. Appl. Probab., 8 (1998), pp. 1130–1155.

[21] G. C. Papanicolaou, *Some probabilistic problems and methods in singular perturbations*, Rocky Mountain J. Math., 6 (1976), pp. 653–674.

[22] G. C. Papanicolaou, *Introduction to the asymptotic analysis of stochastic equations*, in Modern Modeling of Continuum Phenomena, Lectures in Appl. Math. 16, AMS, Providence, RI, 1977, pp. 109–147.

[23] E. Pardoux and A. Yu. Veretennikov, *On the Poisson equation and diffusion approximation*, Ann. Probab., 29 (2001), pp. 1061–1085.

[24] S. P. Sethi and Q. Zhang, *Hierarchical Decision Making in Stochastic Manufacturing Systems*, Birkhäuser Boston, Boston, 1994.

[25] H. A. Simon and A. Ando, *Aggregation of variables in dynamic systems*, Econometrica, 29 (1961), pp. 111–138.

[26] A. V. Skorohod, *Asymptotic Methods of the Theory of Stochastic Differential Equations*, Transl. Math. Monogr. 78, AMS, Providence, RI, 1989.

[27] M. I. Vishik and L. A. Lyusternik, *Regular degeneration and boundary layer for linear differential equations containing a small parameter*, Uspekhi Mat. Nauk, 12 (1957), pp. 3–122.

[28] G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications: A Singular Perturbation Approach*, Springer, New York, 1998.

SIAM J. MATH. ANAL.
Vol. 35, No. 6, pp. 1561–1597

NONLINEAR STABILITY OF STRONG RAREFACTION WAVES FOR COMPRESSIBLE NAVIER–STOKES EQUATIONS*

KENJI NISHIHARA†, TONG YANG‡, AND HUIJIANG ZHAO§

Abstract. This paper is concerned with the time-asymptotic behavior toward strong rarefaction waves of solutions to one-dimensional compressible Navier–Stokes equations. Assume that the corresponding Riemann problem to the compressible Euler equations can be solved by rarefaction waves $(V^R, U^R, S^R)(t,x)$. If the initial data $(v_0, u_0, s_0)(x)$ to the nonisentropic compressible Navier–Stokes equations is a small perturbation of an approximate rarefaction wave constructed as in [S. Kawashima, A. Matsumura, and K. Nishihara, *Proc. Japan Acad. Ser. A*, 62 (1986), pp. 249–252], then we show that, for the general gas, the Cauchy problem admits a unique global smooth solution $(v,u,s)(t,x)$ which tends to $(V^R, U^R, S^R)(t,x)$ as t tends to infinity. A global stability result can also be established for the nonisentropic ideal polytropic gas, provided that the adiabatic exponent γ is close to 1. Furthermore, we show that for the isentropic compressible Navier–Stokes equations, the corresponding global stability result holds, provided that the resulting compressible Euler equations are strictly hyperbolic and both characteristical fields are genuinely nonlinear. Here, global stability means that the initial perturbation can be large. Since we do not require the strength of the rarefaction waves to be small, these results give the nonlinear stability of strong rarefaction waves for the one-dimensional compressible Navier–Stokes equations.

Key words. strong rarefaction waves, global stability, compressible Navier–Stokes equations

AMS subject classifications. 35L65, 35L60

DOI. 10.1137/S003614100342735X

1. Introduction and the main results.

Consider the one-dimensional compressible Navier–Stokes equations in the Lagrangian coordinates,

$$\begin{cases} v_t - u_x = 0, \\ u_t + p_x = \left(\mu \dfrac{u_x}{v}\right)_x, \\ \left(e + \dfrac{u^2}{2}\right)_t + (up)_x = \left(\kappa \dfrac{\theta_x}{v} + \mu \dfrac{u u_x}{v}\right)_x, \end{cases} \tag{1.1}$$

where the unknowns $v > 0, u$, $\theta > 0$, p, e, and s represent the specific volume, the velocity, the absolute temperature, the pressure, the internal energy, and the entropy of the gas, respectively. The coefficients of viscosity and heat-conductivity, μ and κ, are assumed to be positive constants. We assume, as is usual in thermodynamics,

*Received by the editors May 5, 2003; accepted for publication (in revised form) August 15, 2003; published electronically April 7, 2004.

http://www.siam.org/journals/sima/35-6/42735.html

†School of Political Science and Economics, Waseda University, Tokyo 169-8050, Japan (kenji@waseda.jp). This author was supported in part by the Grant-in-Aid for Scientific Research (C) (2) 13640223 of the Japan Society for the Promotion of Science.

‡Department of Mathematics, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong (matyang@math.cityu.edu.hk). This author was supported by Strategic Research Grant of City University of Hong Kong 7001439.

§Wuhan Institute of Physics and Mathematics, The Chinese Academy of Sciences, Wuhan 430071, People's Republic of China and School of Political Science and Economics, Waseda University, Tokyo 169-8050, Japan (hhjjzhao@hotmail.com). This author was supported in part by the JSPS Research Fellowship for Foreign Researchers, the National Natural Science Foundation of China under contracts 10001036 and 10329101, respectively, and the grant from the Chinese Academy of Sciences entitled "Yin Jin Guo Wai Jie Chu Ren Cai Ji Jin."

that by any given two of the five thermodynamical variables, v, p, e, θ, and s, the remaining three variables are expressed.

The second law of thermodynamics asserts that

$$\theta ds = de + pdv,$$

from which, if we choose (v,θ), (v,s), or (v,e) as independent variables and write $(p,e,s) = (p,e,s)(v,\theta)$, or $(p,e,\theta) = (\tilde{p},\tilde{e},\tilde{\theta})(v,s)$, or $(p,s,\theta) = (\hat{p},\hat{s},\hat{\theta})(v,e)$, respectively, then we can deduce that

$$\begin{cases} s_v(v,\theta) = p_\theta(v,\theta), \\ s_\theta(v,\theta) = \dfrac{e_\theta(v,\theta)}{\theta}, \\ e_v(v,\theta) = \theta p_\theta(v,\theta) - p(v,\theta), \end{cases} \tag{1.2}$$

$$\begin{cases} \tilde{e}_v(v,s) = -p(v,\theta), & \tilde{e}_s(v,s) = \theta, \\ \tilde{p}_v(v,s) = p_v(v,\theta) - \dfrac{\theta(p_\theta(v,\theta))^2}{e_\theta(v,\theta)}, & \tilde{p}_s(v,s) = \dfrac{\theta p_\theta(v,\theta)}{e_\theta(v,\theta)}, \\ \tilde{\theta}_v(v,s) = -\dfrac{\theta p_\theta(v,\theta)}{e_\theta(v,\theta)}, & \tilde{\theta}_s(v,s) = \dfrac{\theta}{e_\theta(v,\theta)}, \end{cases} \tag{1.3}$$

or

$$\begin{cases} \hat{s}_e(v,e) = \dfrac{1}{\theta}, & \hat{s}_v(v,e) = \dfrac{p(v,\theta)}{\theta}, \\ \hat{p}_e(v,e) = \dfrac{p_\theta(v,\theta)}{e_\theta(v,\theta)}, & \hat{p}_v(v,e) = \left(p_v(v,\theta) - \dfrac{\theta(p_\theta(v,\theta))^2}{e_\theta(v,\theta)}\right) + \dfrac{p(v,\theta)p_\theta(v,\theta)}{e_\theta(v,\theta)}, \\ \hat{\theta}_e(v,e) = \dfrac{1}{e_\theta(v,\theta)}, & \hat{\theta}_v(v,e) = \dfrac{p(v,\theta) - \theta p_\theta(v,\theta)}{e_\theta(v,\theta)}. \end{cases} \tag{1.4}$$

From (1.3) and (1.4), we get that

$$\tilde{p}_v(v,s) = \hat{p}_v(v,e) - p(v,\theta)\hat{p}_e(v,e). \tag{1.5}$$

In this paper, we are interested in showing that the strong expansion waves for (1.1) are nonlinear stable. For this, it is convenient to work with the equations for the entropy s and the absolute temperature θ, i.e.,

$$s_t = \kappa\left(\frac{\theta_x}{v\theta}\right)_x + \kappa\frac{\theta_x^2}{v\theta^2} + \mu\frac{u_x^2}{v\theta} \tag{1.6}$$

and

$$\theta_t + \frac{\theta p_\theta(v,\theta)}{e_\theta(v,\theta)}u_x = \frac{\kappa}{e_\theta(v,\theta)}\left(\frac{\theta_x}{v}\right)_x + \frac{\mu}{e_\theta(v,\theta)}\frac{u_x^2}{v}. \tag{1.7}$$

In fact, for smooth solutions, $(1.1)_1$, $(1.1)_2$, $(1.1)_3$ are equivalent to $(1.1)_1$, $(1.1)_2$, (1.6) or $(1.1)_1$, $(1.1)_2$, (1.7). In what follows, we will consider $(1.1)_1$, $(1.1)_2$, (1.6) with the initial data

$$(v,u,s)(t,x)|_{t=0} = (v_0,u_0,s_0)(x) \to (v_\pm,u_\pm,s_\pm) \quad \text{as } x \to \pm\infty. \tag{1.8}$$

Here $v_\pm > 0,\ u_\pm,\ s_\pm$ are constants. Since we will focus on the expansion waves to (1.1), we assume that $s_+ = s_- = \overline{s}$ in the rest of this paper.

For expansion waves, the right-hand side of (1.1) decays faster than each term on the left-hand side. Therefore, the compressible Navier–Stokes equations (1.1) may be approximated, time-asymptotically, by the compressible Euler equations

$$\begin{cases} v_t - u_x = 0, \\ u_t + \tilde{p}(v,s)_x = 0, \\ s_t = 0. \end{cases} \tag{1.9}$$

There are two families of expansion (rarefaction) waves for (1.9) which are solutions of the compressible Euler equations (1.9) with Riemann data $(v_0^R, u_0^R, s_0^R)(x)$ (cf. [1]):

$$(v,u,s)(t,x)|_{t=0} = (v_0^R, u_0^R, s_0^R)(x) = \begin{cases} (v_-, u_-, s_-), & x < 0, \\ (v_+, u_+, s_+), & x > 0. \end{cases} \tag{1.10}$$

For illustration, we consider only the 1-rarefaction wave $(V^R, U^R, S^R)(t,x)$, which is characterized by

$$\begin{cases} S^R(t,x) = \overline{s}, \\ U^R(t,x) - \displaystyle\int^{V^R(t,x)} \sqrt{-\tilde{p}_v(z,\overline{s})}\,dz = u_\pm - \int^{v_\pm} \sqrt{-\tilde{p}_v(z,\overline{s})}\,dz, \\ \lambda_{1x}(V^R(t,x), S^R(t,x)) > 0, \qquad \lambda_1(v,s) = -\sqrt{-\tilde{p}_v(v,s)}. \end{cases} \tag{1.11}$$

The case for the 3-rarefaction wave can be discussed similarly.

Before stating the main results, we first list the assumptions on the pressure function $p(v,\theta)$ and the internal energy $e(v,\theta)$ used throughout this paper:

$$(\mathrm{H}_1) \qquad p_v(v,\theta) = \frac{\partial p(v,\theta)}{\partial v} < 0, \qquad e_\theta(v,\theta) = \frac{\partial e(v,\theta)}{\partial \theta} > 0$$

and

$$(\mathrm{H}_2) \qquad \tilde{p}_{vv}(v,s) = \frac{\partial^2 \tilde{p}(v,s)}{\partial v^2} > 0 \quad \text{and } \tilde{p}(v,s) \text{ is convex with respect to } (v,s).$$

From (1.3) and (H_1), we can deduce that

$$\tilde{p}_v(v,s) = p_v(v,\theta) - \frac{\theta(p_\theta(v,\theta))^2}{e_\theta(v,\theta)} < 0, \tag{1.12}$$

$$\begin{cases} \tilde{e}_{ss}(v,s) = \dfrac{\theta}{e_\theta(v,\theta)} > 0, \\ \tilde{e}_{vs}(v,s) = \dfrac{\theta p_\theta(v,\theta)}{e_\theta(v,\theta)}, \\ \tilde{e}_{vv}(v,s) = -p_v(v,\theta) + \dfrac{\theta(p_\theta(v,\theta))^2}{e_\theta(v,\theta)} > 0, \end{cases}$$

and

$$\tilde{e}_{ss}(v,s)\tilde{e}_{vv}(v,s)-(\tilde{e}_{vs}(v,s))^2=-\frac{\theta p_v(v,\theta)}{e_\theta(v,\theta)}>0. \tag{1.13}$$

Equation (1.13) implies that $\tilde{e}(v,s)$ is convex with respect to v and s. Consequently, $\tilde{e}(v,s)+\frac{1}{2}u^2$ is a strictly convex function of (v,u,s). Now we can construct the following normalized entropy $\eta(v,u,s;V,U,S)$ around $(V,U,S)(t,x)$, which is the smooth approximation of the 1-rarefaction waves $(V^R,U^R,S^R)(t,x)$:

$$\begin{aligned}\eta(v,u,s;V,U,S)=&\left(e(v,\theta)+\frac{u^2}{2}\right)-\left(e(V,\Theta)+\frac{U^2}{2}\right)\\&-\{-p(V,\Theta)(v-V)+U(u-U)+\Theta(s-S)\}.\end{aligned} \tag{1.14}$$

Here we have used the fact that $\tilde{e}_v(v,s)=-p(v,\theta)$, $\tilde{e}_s(v,s)=\theta$. The approximate rarefaction waves $V(t,x)$, $U(t,x)$, $S(t,x)$, and $\Theta(t,x)$ are constructed as follows (cf. [20]).

Given a suitably small but fixed constant $\varepsilon>0$, let $w(t,x)$ be the unique global smooth solution to the Cauchy problem

$$\begin{cases} w_t+ww_x=0,\\ \\ w(t,x)|_{t=0}=w_0(x):=\dfrac{\lambda_1(v_-,\overline{s})+\lambda_1(v_+,\overline{s})}{2}+\dfrac{\lambda_1(v_+,\overline{s})-\lambda_1(v_-,\overline{s})}{2}\tanh(\varepsilon x);\end{cases} \tag{1.15}$$

then $V(t,x)$, $U(t,x)$, $S(t,x)$, and $\Theta(t,x)$ are defined by

$$\begin{cases} \lambda_1(V(t,x),\overline{s})=-\sqrt{-\tilde{p}_v(V(t,x),\overline{s})}=w(t,x),\\ \\ U(t,x)=u_\pm+\displaystyle\int_{v_\pm}^{V(t,x)}\sqrt{-\tilde{p}_v(z,\overline{s})}dz,\\ S(t,x)=\overline{s},\\ \\ \Theta(t,x)=\tilde{\theta}(V(t,x),\overline{s}).\end{cases} \tag{1.16}$$

Under the above preparation, for the general gas, our stability result on strong rarefaction waves $(V^R,U^R,S^R)(t,x)$ can be stated as in the following.

THEOREM 1.1 (local stability result for general gas). *Assume that* $(V^R,U^R,S^R)(t,x)$ *is the* 1*-rarefaction wave solution to the Riemann problem of the compressible Euler equations* (1.9), (1.10) *and that the initial data* $(v_0,u_0,s_0)(x)$ *of the compressible Navier–Stokes equations* $(1.1)_1$, $(1.1)_2$, (1.6) *satisfies* (1.8)*:*

$$\begin{cases} 0<2\underline{V}\le v_0(x),\ V(t,x)\le\dfrac{1}{2}\overline{V},\\ \\ 0<2\underline{\Theta}\le\theta_0(x),\ \Theta(t,x)\le\dfrac{1}{2}\overline{\Theta},\end{cases} \tag{1.17}$$

for all $(t,x)\in\mathbf{R}_+\times\mathbf{R}$ *and some positive constants* $\underline{V}$, $\overline{V}$, $\underline{\Theta}$, *and* $\overline{\Theta}$, *and*

$$N(0)=\|(v_0(x)-V(0,x),u_0(x)-U(0,x),s_0(x)-\overline{s})\|_{H^2(\mathbf{R})}$$

is sufficiently small. Then the Cauchy problem (1.1), (1.8) *admits a unique global smooth solution* $(v,u,s)(t,x)$ *satisfying*

$$\lim_{t\to+\infty}\sup_{x\in\mathbf{R}}\{|(v(t,x)-V^R(t,x),u(t,x)-U^R(t,x),s(t,x)-\overline{s})|\}=0. \tag{1.18}$$

Note that the essential meaning of nonlinear stability of rarefaction waves to the compressible Navier–Stokes equations (1.1), (1.8) in [12], [15], [20], [21], [22] is that if $(v_0,u_0,s_0)(x)$ is a (small or large) perturbation of $(V(0,x),U(0,x),\overline{s})$, the smooth approximation of the rarefaction wave solutions $(V^R(t,x),U^R(t,x),\overline{s})$, then the Cauchy problem of the compressible Navier–Stokes equations (1.1), (1.8) admits a unique global smooth solution $(v,u,s)(t,x)$ which tends time-asymptotically to $(V^R(t,x),U^R(t,x),\overline{s})$. In this sense, the result obtained in Theorem 1.1 does imply the nonlinear stability of strong rarefaction waves for the compressible Navier–Stokes equations. But, due to the assumption that the initial perturbation $(v_0(x)-V(0,x),u_0(x)-U(0,x),s_0(x)-\overline{s})$ should be small, the nonlinear stability result obtained in Theorem 1.1 is essentially local. Then a natural question of importance and interest is how to get the global stability result which is for large perturbation. Our second purpose is to devote to this problem and show that, for the ideal polytropic gas, such a global stability result indeed holds for γ near 1 without the weakness of the rarefaction waves. To state the result precisely, we recall that for the ideal polytropic gas, $(p,e)(v,\theta)$ have the following special constitutive relations:

$$p(v,\theta)=\frac{R\theta}{v}=Av^{-\gamma}\exp\left(\frac{\gamma-1}{R}s\right),\qquad e(v,\theta)=\frac{R\theta}{\gamma-1}, \tag{1.19}$$

where $R>0$ is the gas constant, $\gamma>1$ the adiabatic constant, and A a positive constant.

Our second result is stated as follows.

THEOREM 1.2 (global stability result for the ideal polytropic gas). *Assume that* $(V^R(t,x),U^R(t,x),\overline{s})$ *is the* 1*-rarefaction wave solution of the Riemann problem of the compressible Euler equations* (1.9), (1.10) *and that* $(p,e)(v,\theta)$ *satisfy the constitutive relations* (1.19). *Then for any* $(v_0(x)-V(0,x),u_0(x)-U(0,x),s_0(x)-\overline{s})\in H^2(\mathbf{R})$ *satisfying* (1.17) *and its* $H^1(\mathbf{R})$*-norm to be bounded by a constant independent of* $\frac{1}{\varepsilon}$, *the corresponding Cauchy problem* (1.1), (1.8) *admits a unique global smooth solution* $(v,u,s)(t,x)$ *satisfying* (1.18), *provided that* $\gamma-1$ *is sufficiently small.*

In the proof of Theorem 1.2, the assumption that γ is close to 1 is used for obtaining the a priori assumption $0<\underline{\Theta}<\theta(t,x)<\overline{\Theta}$ for $(t,x)\in[0,\infty]\times\mathbf{R}$ so that $\theta(t,x)-\Theta(t,x)$ is small. Hence, one can imagine that for the isentropic polytropic gas, such a smallness assumption can be removed, and this has been obtained by Matsumura and Nishihara in [21], [22] by cleverly introducing another type of smooth approximation of the rarefaction wave solution. That is, $w_0(x)$ in $(1.15)_2$ is replaced by

$$\begin{aligned}w(t,x)|_{t=0}=\underline{w}_0(x)=&\frac{\lambda_1(v_-,\overline{s})+\lambda_1(v_+,\overline{s})}{2}\\&+\frac{\lambda_1(v_+,\overline{s})-\lambda_1(v_-,\overline{s})}{2}K_q\int_0^{\varepsilon x}(1+y^2)^{-q}dy,\end{aligned} \tag{1.20}$$

where $K_q>0$ is a constant satisfying

$$K_q\int_0^{+\infty}(1+y^2)^{-q}dy=1 \tag{1.21}$$

for some suitably large constant $q > 0$.

Our third purpose is to show the global stability result on strong rarefaction waves for a p-system with viscosity with a general pressure $p = p(v)$. To state this result, we recall that the isentropic compressible Navier–Stokes equations in Lagrangian coordinates can be written as

$$\begin{cases} v_t - u_x = 0, \\ u_t + p(v)_x = \mu \left(\dfrac{u_x}{v} \right)_x, \end{cases} \tag{1.22}$$

with the initial data

$$(v, u)(t, x)|_{t=0} = (v_0, u_0)(x) \to (v_\pm, u_\pm) \quad \text{as } x \to \pm\infty. \tag{1.23}$$

Here $v_\pm > 0$ and $u_\pm$ are given constants so that the Riemann problem of the isentropic compressible Euler equations

$$\begin{cases} v_t - u_x = 0, \\ u_t + p(v)_x = 0, \end{cases} \tag{1.24}$$

with the Riemann data

$$(v, u)(t, x)|_{t=0} = (\overline{v}_0^R, \overline{u}_0^R)(x) = \begin{cases} (v_-, u_-), & x < 0, \\ (v_+, u_+), & x > 0, \end{cases} \tag{1.25}$$

is assumed to admit a unique 1-rarefaction wave solution $(\overline{V}^R, \overline{U}^R)(t, x)$.

We assume only that $p(v)$ is a positive smooth function for $v > 0$ and satisfies

$$p'(v) < 0, \quad p''(v) > 0 \quad \text{for } v > 0. \tag{1.26}$$

Under the above assumptions, we have the following theorem.

THEOREM 1.3 (global stability result for general isentropic gas). *Assume that the Riemann problem* (1.24), (1.25) *to the compressible Euler equations admits a unique* 1*-rarefaction wave solution* $(\overline{V}^R, \overline{U}^R)(t, x)$ *and that* $(\overline{V}, \overline{U})(t, x)$ *is a smooth approximation of the Riemann solution* $(\overline{V}^R, \overline{U}^R)(t, x)$ *constructed by*

$$\begin{cases} \overline{V}(t, x) = \lambda_1^{-1}(\overline{w}(t, x)), \quad \lambda_1(v) = -\sqrt{-p'(v)}, \\ \overline{U}(t, x) = u_\pm + \displaystyle\int_{v_\pm}^{\overline{V}(t,x)} \sqrt{-p'(s)}\, ds. \end{cases} \tag{1.27}$$

Here $\overline{w}(t, x)$ *is the unique smooth solution to the following Cauchy problem:*

$$\begin{cases} w_t + w w_x = 0, \\ w(t, x)|_{t=0} = \overline{w}_0(x) = \dfrac{\lambda_1(v_-) + \lambda_1(v_+)}{2} + \dfrac{\lambda_1(v_+) - \lambda_1(v_-)}{2} \tanh(\varepsilon x). \end{cases} \tag{1.28}$$

Then for any $p(v)$ *satisfying* (1.26) *and* $(v_0(x) - \overline{V}(0, x), u_0(x) - \overline{U}(0, x)) \in H^2(\mathbf{R})$ *satisfying* $0 < 2\underline{V} \le v_0(x), \overline{V}(t, x) \le \frac{1}{2}\overline{V}$ *for all* $(t, x) \in \mathbf{R}_+ \times \mathbf{R}$ *and some positive constants* $\underline{V}$, $\overline{V}$ *and with its* $H^1(\mathbf{R})$*-norm bounded by a constant independent of the*

quantity $\frac{1}{\varepsilon}$*, the Cauchy problem* (1.22), (1.23) *admits a unique global smooth solution* $(v,u)(t,x)$ *satisfying*

$$\lim_{t\to+\infty}\sup_{x\in\mathbf{R}}\left\{\left|\left(v-\overline{V}^R,u-\overline{U}^R\right)(t,x)\right|\right\}=0. \tag{1.29}$$

Remark 1.4. In [21] and [22], the assumption that $p(v)=v^{-\gamma}$ $(\gamma\geq 1)$ plays an essential role in the analysis, and it is worth pointing out that even by using their smooth approximation of the Riemann solutions, their arguments cannot be applied to the case when $p(v)$ satisfies only (1.26). However, we have assumed that the $H^1(\mathbf{R})$-norm of the initial perturbation is bounded by a constant independent of $\frac{1}{\varepsilon}$ with small fixed number $\varepsilon>0$. This implies that the data $(v_0,u_0)(x)$ for (1.23) is initially rather flat though $(v_0(x),\overline{V}(0,x),u_0(x)-\overline{U}(0,x))$ may be large. Therefore, we should seek the global solution and its behavior for any data $(v_0,u_0)(x)$ with $\|(v_0(x)-v_\pm,u_0(x)-u_\pm)\|_{H^1(\mathbf{R}_\pm)}$ bounded. This will be done under some additional assumptions on $p(v)$ in Theorem 1.5.

In Theorems 1.1, 1.2, and 1.3, we assume that the solutions to the corresponding Riemann problem of the compressible Euler equations consists of only one rarefaction wave. In fact, such a restriction can be removed by suitably modifying the arguments used in the proof of the theorems. To simplify the presentation, we use the isentropic compressible Navier–Stokes equations to explain this. Suppose that the solution $(\overline{V}^R,\overline{U}^R)(t,x)$ to the Riemann problem (1.24), (1.25) consists of one 1-rarefaction wave $(\overline{V}_1^R,\overline{U}_1^R)(t,x)$ and one 2-rarefaction wave $(\overline{V}_2^R,\overline{U}_2^R)(t,x)$. That is, there exists a unique constant state $(\overline{v},\overline{u})\in\mathbf{R}^2$ such that (v_-,u_-) and $(\overline{v},\overline{u})$ are connected by one 1-rarefaction wave $(\overline{V}_1^R,\overline{U}_1^R)(t,x)$, i.e., $(\overline{v},\overline{u})\in R_1(v_-,u_-)$, while $(\overline{v},\overline{u})$ and (v_+,u_+) are connected by one 2-rarefaction wave $(\overline{V}_2^R,\overline{U}_2^R)(t,x)$, i.e., $(v_+,u_+)\in R_2(\overline{v},\overline{u})$. Here

$$\begin{cases} R_1(v_-,u_-)=\left\{(v,u)\;\middle|\; u=u_-+\displaystyle\int_{v_-}^{v}\sqrt{-p'(s)}ds,\ u\geq u_-\right\},\\[2ex] R_2(\overline{v},\overline{u})=\left\{(v,u)\;\middle|\; u=\overline{u}-\displaystyle\int_{\overline{v}}^{v}\sqrt{-p'(s)}ds,\ u\geq \overline{u}\right\}. \end{cases} \tag{1.30}$$

Consequently,

$$\left(\overline{V}^R,\overline{U}^R\right)(t,x)=(\overline{V}_1^R(t,x)+\overline{V}_2^R(t,x)-\overline{v},\overline{U}_1^R(t,x)+\overline{U}_2^R(t,x)-\overline{u}). \tag{1.31}$$

Let $\overline{w}_i(t,x)$ $(i=1,2)$ be the unique global smooth solution to the following Cauchy problem:

$$\begin{cases} \overline{w}_{it}+\overline{w}_i\overline{w}_{ix}=0,\\[2ex] \overline{w}_i(t,x)|_{t=0}=\overline{w}_{i0}(x)=\dfrac{\overline{w}_{i-}+\overline{w}_{i+}}{2}+\dfrac{\overline{w}_{i+}-\overline{w}_{i-}}{2}\tanh(\varepsilon x),\quad i=1,2; \end{cases} \tag{1.32}$$

then, as in [20], the smooth approximate solution $(\overline{V},\overline{U})(t,x)$ of $(\overline{V}^R,\overline{U}^R)(t,x)$ is constructed as follows:

$$(\overline{V},\overline{U})(t,x)=(\overline{V}_1(t,x)+\overline{V}_2(t,x)-\overline{v},\overline{U}_1(t,x)+\overline{U}_2(t,x)-\overline{u}), \tag{1.33}$$

where $(\overline{V}_1, \overline{U}_1)(t,x)$ (resp., $(\overline{V}_2, \overline{U}_2)(t,x)$) is defined by

(1.34)
$$\begin{cases} \lambda_1(\overline{V}_1(t,x)) = \overline{w}_1(t,x) & (\text{resp., } \lambda_2(\overline{V}_2(t,x)) = \overline{w}_2(t,x)), \\ \overline{U}_1 = u_- + \displaystyle\int_{v_-}^{\overline{V}_1(t,x)} \sqrt{-p'(s)}ds & \left(\text{resp., } \overline{U}_2(t,x) = \overline{u} - \displaystyle\int_{\overline{v}}^{\overline{V}_2(t,x)} \sqrt{-p'(s)}ds\right) \end{cases}$$

and $\overline{w}_1(t,x)$ (resp., $\overline{w}_2(t,x)$) is the solution of (1.32) with $\overline{w}_{1-} = \lambda_1(v_-)$ and $\overline{w}_{1+} = \lambda_1(\overline{v})$ (resp., $\overline{w}_{2-} = \lambda_2(\overline{v})$ and $\overline{w}_{2+} = \lambda_2(v_+)$).

It is easy to deduce that the smooth functions $(\overline{V}, \overline{U})(t,x)$ satisfy the system

$$\begin{cases} \overline{V}_t - \overline{U}_x = 0, \\ \overline{U}_t + p(\overline{V})_x = g(\overline{V})_x, \end{cases} \tag{1.35}$$

where $g(\overline{V}) = p(\overline{V}) - p(\overline{V}_1) - p(\overline{V}_2) + p(\overline{v})$. Hence, we need only to control $g(\overline{V}(t,x))_x$ suitably in this case. Notice that from the properties on the smooth approximation of the rarefaction wave solution stated in section 2 (cf. [19]), we have only

$$\int_0^t \|g(\overline{V}(\tau))_x\|_{L^p(\mathbf{R})} d\tau \le O(1)\varepsilon^{-\frac{1}{p}}. \tag{1.36}$$

From this observation, together with the fact that, in deducing our main results, we need the smallness of ε, a quantity introduced in the construction of the smooth approximation to the rarefaction wave solutions, to close the energy estimates, it seems hopeless to use our method to deal with the nonlinear stability of the superposition of rarefaction waves of different families.

We note, however, that $g(\overline{V}(t,x))_x$ satisfies the following estimate (cf. [20]): There exist constants $C > 0$, $\alpha > 0$ such that for $t \ge 0$, $x \in \mathbf{R}$,

$$|g(\overline{V}(t,x))_x| \le C\varepsilon \exp(-\alpha\varepsilon(|x| + t)). \tag{1.37}$$

From (1.37), we can see that, as in [14] for the study of nonlinear stability of travelling wave solutions to dissipative hyperbolic systems of conservation laws, if we give the smooth approximation $\overline{V}(t,x)$ a shift, that is, if we let $\overline{V}'(t,x) = \overline{V}(t + t_0, x)$ with $t_0 > 0$ being a suitably chosen fixed constant, then we have for $\overline{V}'(t,x)$ that

$$\int_0^t \|g(\overline{V}'(\tau))_x\|_{L^p(\mathbf{R})} \le O(1)\varepsilon^{-\frac{1}{p}} \exp(-\alpha\varepsilon t_0). \tag{1.38}$$

If we let, for example, $t_0 = \varepsilon^{-2}$, then the right-hand side of (1.38) is controlled by $O(1)\varepsilon^{-\frac{1}{p}} \exp(-\frac{\alpha}{\varepsilon})$, which can be as small as we want if we choose $\varepsilon > 0$ sufficiently small. Consequently, our method can indeed be applied directly to deal with the nonlinear stability of the superposition of rarefaction waves of different families, provided that we approximate the rarefaction wave solutions by $\overline{V}'(t,x)$. (Note that in this case the initial data $(v_0, u_0)(x)$ of the compressible Navier–Stokes equations (1.24) is a perturbation of $(\overline{V}, \overline{U})(t_0, x)$.)

In Theorems 1.2 and 1.3, we assume that the H^1-norm of the initial perturbation is bounded by a constant independent of $\frac{1}{\varepsilon}$, which is excluded under the additional

assumption

$$\begin{cases} p(v) \geq C_1^{-1} v^{-1}, \quad C_1 p(v) \geq v|p'(v)| = -vp'(v) \geq C_1^{-1} & (0 < v \leq 1), \\ -p'(v) \geq C_1^{-1} v^{-C_1} \quad (v \geq 1) \end{cases} \tag{1.39}$$

for arbitrarily fixed constant $C_1 > 2$. Note that (1.39) derives

$$\begin{cases} C_1^{-1} v^{-1} \leq p(v) \leq p(1) v^{-C_1} \quad (0 < v \leq 1), \\ p(v) \geq p(\infty) + \dfrac{v^{1-C_1}}{C_1(C_1 - 1)} \quad (v \geq 1). \end{cases} \tag{1.40}$$

Hence, though (1.40) is not a sufficient condition for (1.39), the assumption (1.39), roughly speaking, seems to be reasonable, including the typical pressure model $p(v) = v^{-\gamma}$ ($\gamma \geq 1$). Then we have the final theorem.

THEOREM 1.5. *Assume that $p(v)$ satisfies* (1.26) *and* (1.39) *and that the solution $(\overline{V}^R, \overline{U}^R)(t,x)$ to the Riemann problem* (1.24), (1.25) *is given by* (1.31). *Let $(\overline{\overline{V}}, \overline{\overline{U}})(t,x)$ be a smooth approximation of the Riemann solution $(\overline{V}^R, \overline{U}^R)(t,x)$ constructed by* (1.33)–(1.34) *with $\overline{w}_{i0}(x)$ in* (1.32) *being replaced by*

$$\frac{\overline{w}_{i-} + \overline{w}_{i+}}{2} + \frac{\overline{w}_{i+} - \overline{w}_{i-}}{2} K_q \int_0^{\varepsilon x} (1 + y^2)^{-q} dy$$

for $q > \frac{3}{2}$ and K_q satisfying (1.21).

Then for any $(v_0(x) - \overline{\overline{V}}(0,x), u_0(x) - \overline{\overline{U}}(0,x)) \in H^1(\mathbf{R})$ satisfying $0 < 2\underline{V} \leq v_0(x)$, $\overline{\overline{V}}(t,x) \leq \frac{1}{2}\overline{V}$ for all $(t,x) \in \mathbf{R}_+ \times \mathbf{R}$ and some positive constants $\underline{V}$, $\overline{V}$, the Cauchy problem (1.22), (1.23) *admits a unique global smooth solution $(v,u)(t,x)$ satisfying* (1.29)

Before concluding this section, we point out that the large time behavior of solutions to the compressible Navier–Stokes equations (1.1), (1.8) has been studied by many people; cf. [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24] and the references cited therein. When the initial data $(v_0, u_0, s_0)(x)$ is a small perturbation of a nonvacuum constant state, i.e., $v_- = v_+ > 0, u_- = u_+, s_- = s_+$, quite perfect results have been obtained; cf. [10] and [17]. In the case when the far fields of the initial data are different, i.e., $(v_-, u_-, s_-) \neq (v_+, u_+, s_+)$, many interesting results have been obtained: When the solutions to the corresponding Riemann problem consist of only shock waves, the nonlinear stability of travelling wave solutions has been established by [11], [14], [19], etc. However, when the solutions to the corresponding Riemann problem consist of only rarefaction waves, the corresponding nonlinear stability results are obtained by [12], [15], [21], and [22].

This paper is arranged as follows. We will give some properties of the smooth approximation of the rarefaction wave solutions in section 2. The proof of Theorems 1.1, 1.2, 1.3, and 1.5 are given in sections 3, 4, 5, and 6, respectively.

Throughout the rest of this paper, C or $O(1)$ will be used to denote a generic positive constant independent of t and x, and $C_i(\cdot,\cdot)$ ($i \in \mathbf{Z}_+$) stands for some generic constants depending only on the quantities listed in the parentheses. For two functions $f(x)$ and $g(x)$, $f(x) \sim g(x)$ as $x \to a$ means that there exists a positive constant $C > 0$ such that $C^{-1} f(x) \leq g(x) \leq C f(x)$ in the neighborhood of a. $H^l(\mathbf{R})$ ($l \geq 0$) denotes the usual Sobolev space with norm $\|\cdot\|_l$, and $\|\cdot\|_0 = \|\cdot\|$ will be used to denote

the usual L^2-norm. For a vector $a = (a_1, a_2, \dots, a_n)$, $|a| = (\sum_{j=1}^n a_j^2)^{\frac{1}{2}}$. Finally, for $1 \le p \le +\infty$, $f(x) \in L^p(\mathbf{R}, \mathbf{R}^n)$, $|f|_p = (\int_{\mathbf{R}} |f(x)|^p dx)^{\frac{1}{p}}$. It is easy to see that $|\cdot|_2 = \|\cdot\|$.

2. Properties of smooth approximate solution of the Riemann problem. In the same situation as in [12], we start with the Riemann problem for the typical Burgers equation:

$$\begin{cases} w_t^R + w^R w_x^R = 0, \\ w^R(0,x) = w_0^R(x) = \begin{cases} w_- \equiv \lambda_1(v_-, \bar{s}) < 0, & x < 0, \\ w_+ \equiv \lambda_1(v_+, \bar{s}) < 0, & x > 0, \end{cases} \end{cases} \tag{2.1}$$

with $w_- < w_+ < 0$. As is well known, (2.1) has a continuous weak solution of the form $w^R(\frac{x}{t})$ given by

$$w^R(\xi) = \begin{cases} w_-, & \xi \le w_-, \\ \xi, & w_- \le \xi \le w_+, \\ w_+, & \xi \ge w_+. \end{cases} \tag{2.2}$$

The main idea in [12] is to approximate $w^R(\frac{x}{t})$ by the solution $w(t,x)$ of the Cauchy problem (1.15). Since $w_0(x)$ is strictly increasing, we have the following lemma (cf. [12]).

LEMMA 2.1. *If $w_- < w_+$, then the Cauchy problem* (1.15) *has a unique global smooth solution $w(t,x)$ satisfying the following:*

(i) $w_- < w(t,x) < w_+ < 0$, $w_x(t,x) > 0$ *for all* $(t,x) \in \mathbf{R}_+ \times \mathbf{R}$.

(ii) *For any p $(1 \le p \le \infty)$, there exists a constant $C(p)$, depending only on p, such that*

$$\begin{cases} |w_x(t)|_p \le C(p) \min\{\tilde{w}\varepsilon^{1-\frac{1}{p}}, \tilde{w}^{\frac{1}{p}} t^{-1+\frac{1}{p}}\}, \\ |w_{xx}(t)|_p \le C(p) \min\{\tilde{w}\varepsilon^{2-\frac{1}{p}}, \varepsilon^{1-\frac{1}{p}} t^{-1}\}, \\ |w_{xxx}(t)|_p \le C(p) \min\{\tilde{w}\varepsilon^{3-\frac{1}{p}}, \varepsilon^{2-\frac{1}{p}} t^{-1}\}. \end{cases}$$

(iii) $\lim_{t\to+\infty} \sup_{x\in\mathbf{R}} |w(t,x) - w^R(\frac{x}{t})| = 0$.

Here $\tilde{w} = w_+ - w_-$.

Having obtained $w(t,x)$, we define $V(t,x)$, $U(t,x)$, $S(t,x)$, and $\Theta(t,x)$ according to (1.16). Since $\tilde{p}(v,s)$ satisfies (H_1) and (H_2), one can deduce that $V(t,x)$, $U(t,x)$, $S(t,x)$, and $\Theta(t,x)$ are globally (both with respect to t and x) well defined and smooth, and it is not difficult to check that $V(t,x)$, $U(t,x)$, $S(t,x)$, and $\Theta(t,x)$ satisfy

$$\begin{cases} V_t - U_x = 0, \\ U_t + p(V,\Theta)_x = 0, \\ \left(e(V,\Theta) + \dfrac{U^2}{2}\right)_t + (Up(V,\Theta))_x = 0, \\ \Theta_t + \dfrac{\Theta p_\theta(V,\Theta)}{e_\theta(V,\Theta)} U_x = 0, \\ S_t(V,\Theta) = 0, \end{cases} \tag{2.3}$$

and, due to Lemma 2.1, $V(t,x)$, $U(t,x)$, $S(t,x)$, and $\Theta(t,x)$ have the following properties.

LEMMA 2.2. *The smooth functions $V(t,x)$, $U(t,x)$, $S(t,x)$, and $\Theta(t,x)$ constructed above have the following properties by denoting $\delta = |v_- - v_+| + |u_- - u_+|$:*

(i) $V_t(t,x) = U_x(t,x) > 0$ *for all* $x \in \mathbf{R}$, $t \geq 0$.

(ii) *For any p $(1 \leq p \leq +\infty)$, there exists a constant $C(p)$, depending only on p, such that*

$$\begin{cases} |(V_x, U_x, \Theta_x)(t)|_p \leq C(p) \min\{\delta\varepsilon^{1-\frac{1}{p}}, \delta^{\frac{1}{p}} t^{-1+\frac{1}{p}}\}, \\ |(V_{xx}, U_{xx}, \Theta_{xx})(t)|_p \leq C(p) \min\{\delta\varepsilon^{2-\frac{1}{p}}, \varepsilon^{1-\frac{1}{p}} t^{-1}\}, \\ |(V_{xxx}, U_{xxx}, \Theta_{xxx})(t)|_p \leq C(p) \min\{\delta\varepsilon^{3-\frac{1}{p}}, \varepsilon^{2-\frac{1}{p}} t^{-1}\}. \end{cases}$$

(iii) $\lim_{t\to+\infty} \sup_{x\in\mathbf{R}} |(V,U,S,\Theta)(t,x) - (V^R, U^R, S^R, \Theta^R)(t,x)| = 0$.

(iv) $|(V_t, U_t, \Theta_t)(t,x)| \leq O(1)|(V_x, U_x, \Theta_x)(t,x)|$.

Similar estimates hold for the global smooth functions $(\overline{V}, \overline{U})(t,x)$ defined by (1.27). Moreover, $(\overline{\overline{V}}, \overline{\overline{U}})(t,x)$ constructed in Theorem 1.5 have the following properties.

LEMMA 2.3. *The smooth functions $(\overline{\overline{V}}, \overline{\overline{U}})(t,x)$ constructed in Theorem* 1.5 *satisfy the following:*

(i) $\overline{\overline{V}}_t(t,x) = \overline{\overline{U}}_x(t,x) > 0$.

(ii) *For any $p \in [1, +\infty]$, there exists a positive constant $C(p,q)$ such that*

$$\begin{cases} \left|\left(\overline{\overline{V}}_x, \overline{\overline{U}}_x\right)(t)\right|_p \leq C(p,q) \min\{\delta\varepsilon^{1-\frac{1}{p}}, \delta^{\frac{1}{p}}(1+t)^{-1+\frac{1}{p}}\}, \\ \left|\left(\overline{\overline{V}}_{xx}, \overline{\overline{U}}_{xx}\right)(t)\right|_p \leq C(p,q) \min\{\delta^{-\frac{p-1}{2pq}} \varepsilon^{(1-\frac{1}{2q})(1-\frac{1}{p})}(1+t)^{-1-\frac{p-1}{2pq}}, \delta^{\frac{1}{p}}(1+t)^{-2+\frac{1}{p}}\}, \\ \left|g\left(\overline{\overline{V}}\right)_x(t)\right|_p \leq C(p,q)\varepsilon^{1-\frac{1}{p}}\delta^2\{(1+(\varepsilon\lambda_2(\overline{v})t)^2)^{-\frac{q}{3}} + (1+(\varepsilon\lambda_1(\overline{v})t)^2)^{-\frac{q}{3}}\}. \end{cases}$$

Especially

$$\begin{cases} \displaystyle\int_0^\infty \left|\left(\overline{\overline{V}}_{xx}, \overline{\overline{U}}_{xx}\right)(t)\right|_p dt \leq C(p,q)\delta^{-\frac{p-1}{2pq}} & \text{for } p > 1, \\ \displaystyle\int_0^\infty \left|g\left(\overline{\overline{V}}\right)_x(t)\right|_p dt \leq C(p,q)\varepsilon^{-\frac{1}{p}}\delta^2 & \text{for } p \geq 1,\ q > \dfrac{3}{2}. \end{cases}$$

(iii) $\lim_{t\to+\infty} \sup_{x\in\mathbf{R}} |(\overline{\overline{V}}, \overline{\overline{U}})(t,x) - (\overline{V}^R, \overline{U}^R)(t,x)| = 0$.

(iv) $|(\overline{\overline{V}}_t, \overline{\overline{U}}_t)(t,x)| \leq O(1)|(\overline{\overline{V}}_x, \overline{\overline{U}}_x)(t,x)|$.

3. The proof of Theorem 1.1. This section is devoted to proving Theorem 1.1. To this end, setting

(3.1)
$$(\varphi, \psi, \phi, \xi)(t,x) = (v(t,x) - V(t,x), u(t,x) - U(t,x), \theta(t,x) - \Theta(t,x), s(t,x) - \overline{s}),$$

we can deduce that $(\varphi,\psi,\phi,\xi)(t,x)$ solves

(3.2)
$$\begin{cases} \varphi_t-\psi_x=0,\\ \psi_t+[p(v,\theta)-p(V,\Theta)]_x=\mu\left(\dfrac{u_x}{v}\right)_x,\\ \phi_t+\dfrac{\theta p_\theta(v,\theta)}{e_\theta(v,\theta)}\psi_x+\left(\dfrac{\theta p_\theta(v,\theta)}{e_\theta(v,\theta)}-\dfrac{\Theta p_\theta(V,\Theta)}{e_\theta(V,\Theta)}\right)U_x=\dfrac{1}{e_\theta(v,\theta)}\left\{\kappa\left(\dfrac{\theta_x}{v}\right)_x+\mu\dfrac{u_x^2}{v}\right\},\\ \xi_t=\kappa\left(\dfrac{\theta_x}{v\theta}\right)_x+\kappa\dfrac{\theta_x^2}{v\theta^2}+\mu\dfrac{u_x^2}{v\theta}, \end{cases}$$

with initial data

$$\begin{aligned}(\varphi,\psi,\phi,\xi)(t,x)|_{t=0}&=(\varphi_0,\psi_0,\phi_0,\xi_0)(x)\\ &=(v_0(x)-V(0,x),u_0(x)-U(0,x),\theta_0(x)-\Theta(0,x),s_0(x)-\overline{s}).\end{aligned}\tag{3.3}$$

For convenience of presentation, in what follows we will choose (v,θ) as independent variables and for some fixed $T>0$, we define the solution space of (3.2), (3.3) by

$$X(0,T):=\left\{(\varphi,\psi,\phi)(t,x)\;\middle|\;\begin{array}{l}(\varphi,\psi,\phi)(t,x)\in C^0(0,T;H^2(\mathbf{R}))\\ (\psi_x,\phi_x)(t,x)\in L^2(0,T;H^2(\mathbf{R}))\end{array}\right\}.\tag{3.4}$$

Under the assumptions listed in Theorem 1.1, we can get the following local existence result (cf. [15], [21]).

LEMMA 3.1. *Under the assumptions stated in Theorem* 1.1*, the Cauchy problem* (3.2), (3.3) *admits a unique smooth solution* $(\varphi(t,x),\psi(t,x),\phi(t,x))\in X(0,t_1)$ *for some sufficiently small* $t_1>0$*, and* $(\varphi(t,x),\psi(t,x),\phi(t,x))$ *satisfies*

$$\begin{cases}0<\underline{V}\le\varphi(t,x)+V(t,x)\le\overline{V},\\ 0<\underline{\Theta}\le\phi(t,x)+\Theta(t,x)\le\overline{\Theta}\end{cases}\tag{3.5}$$

and

$$\sup_{[0,t_1]}(\|(\varphi,\psi,\phi)(t)\|_2)\le 2\|(\varphi_0,\psi_0,\phi_0)\|_2.\tag{3.6}$$

To extend the local solution obtained in Lemma 3.1 globally, we need only to get a priori estimates. For this purpose, suppose that $(\varphi,\psi,\phi)(t,x)$ obtained in Lemma 3.1 has been extended to the time $t=T>t_1$, i.e., $(\varphi,\psi,\phi)(t,x)\in X(0,T)$, and satisfies

$$N(t):=\sup_{0\le\tau\le t}\{\|(\varphi,\psi,\phi)(\tau)\|_2\}\le\eta,\quad 0\le t\le T,\tag{3.7}$$

for some positive constant $\eta>0$. Based on the a priori estimates (3.7), if we can show that $(\varphi,\psi,\phi)(t,x)$ satisfies

$$\begin{aligned}&\|(\varphi,\psi,\phi)(t)\|_2^2+\int_0^t\{\|\sqrt{V_t(\tau)}(\varphi,\phi)(\tau)\|^2+\|(\psi_x,\phi_x)(\tau)\|_2^2\}d\tau\\ &\quad\le C(\eta)\{\|(\varphi_0,\psi_0,\phi_0)\|_2^2+\varepsilon^{\frac{1}{4}}\},\end{aligned}\tag{3.8}$$

then, by choosing the initial perturbation $N(0)$ and ε sufficiently small, we can deduce that

$$N(t) < \eta, \quad 0 \le t \le T. \tag{3.9}$$

This implies that the a priori assumption (3.7) is reasonable and, consequently, Theorem 1.1 follows immediately by the standard continuity argument.

Now, to complete the proof of Theorem 1.1, we need only to show that (3.8) is true, provided that $\eta > 0$ is chosen to be sufficiently small.

In fact, from (3.7) and (1.17), we have from Sobolev's inequality and by choosing $\eta > 0$ sufficiently small that for $0 \le t \le T$, $x \in \mathbf{R}$, $v(t,x)$ and $\theta(t,x)$ satisfy

$$\begin{cases} 0 < \underline{V} \le v(t,x) = \varphi(t,x) + V(t,x) \le \overline{V}, \\ 0 < \underline{\Theta} \le \theta(t,x) = \phi(t,x) + \Theta(t,x) \le \overline{\Theta} \end{cases} \tag{3.10}$$

and

$$\sup_{0\le\tau\le t,\ x\in\mathbf{R}} \left| \frac{\partial^i}{\partial x^i}(\varphi,\psi,\phi)(\tau,x) \right| \le O(1)N(t), \quad i = 0, 1. \tag{3.11}$$

Due to

(3.12)

$$\begin{aligned} &\eta_t(v,u,\theta;V,U,\Theta) + \{(p(v,\theta) - p(V,\Theta))\psi\}_x + \left\{ \mu\Theta\frac{\psi_x^2}{v\theta} + \kappa\Theta\frac{\phi_x^2}{v\theta^2} \right\} \\ &\quad + \{\tilde{p}(v,s) - \tilde{p}(V,\overline{s}) - \tilde{p}_v(V,\overline{s})\varphi - \tilde{p}_s(V,\overline{s})\xi\}U_x \\ &= \left(\mu\frac{\psi\psi_x}{v} + \kappa\frac{\phi\phi_x}{v\theta} \right)_x + \left(-\mu\frac{U_x\psi\varphi_x}{v^2} + 2\mu\frac{U_x\phi\psi_x}{v\theta} - \kappa\frac{\Theta_x\phi\varphi_x}{v^2\theta} + \kappa\frac{\Theta_x\phi\phi_x}{v\theta^2} \right) \\ &\quad + \left(\mu\frac{U_{xx}\psi}{v} + \kappa\frac{\Theta_{xx}\phi}{v\theta} \right) + \left(-\mu\frac{V_xU_x\psi}{v^2} + \mu\frac{U_x^2\phi}{v\theta} - \kappa\frac{V_x\Theta_x\phi}{v^2\theta} \right), \end{aligned}$$

we have by integrating (3.12) with respect to t and x over $[0,t]\times\mathbf{R}$ that

$$\begin{aligned} &\|(\varphi,\psi,\phi)(t)\|^2 + \int_0^t \{\|\sqrt{V_t(\tau)}(\varphi,\phi)(\tau)\|^2 + \|(\psi_x,\phi_x)(\tau)\|^2\}d\tau \\ &\qquad \le C(\eta)\left\{ \|(\varphi_0,\psi_0,\phi_0)\|^2 + \sum_{j=1}^{3} R_j \right\}, \end{aligned} \tag{3.13}$$

where

$$\begin{cases} R_1 = \displaystyle\int_0^t \int_{\mathbf{R}} (|\psi U_{xx}| + |\phi\Theta_{xx}|)(\tau,x)dxd\tau, \\ R_2 = \displaystyle\int_0^t \int_{\mathbf{R}} (|U_x\psi\varphi_x| + |U_x\psi_x\phi| + |\Theta_x\phi\psi_x| + |\Theta_x\phi\phi_x|)(\tau,x)dxd\tau, \\ R_3 = \displaystyle\int_0^t \int_{\mathbf{R}} (|V_xU_x\psi| + |U_x^2\phi| + |V_x\Theta_x\phi|)(\tau,x)dxd\tau. \end{cases}$$

Here we have used the assumption that $\tilde{p}(v,s)$ is a convex function of v and s, the fact that $|(\varphi,\psi,\xi)|^2$ is equivalent to $|(\varphi,\psi,\phi)|^2$, and (3.7), (3.10), and (3.11).

R_j $(j = 1, 2, 3)$ can be estimated as follows:

$$
\begin{aligned}
R_1 &\le O(1)\int_0^t \|(\psi,\phi)(\tau)\|^{\frac12}\|(\psi_x,\phi_x)(\tau)\|^{\frac12}|U_{xx}(\tau)|_1 d\tau \\
&\le O(1)\left\{N(t)\int_0^t \|(\psi_x,\phi_x)(\tau)\|^2 d\tau + \int_0^t |U_{xx}(\tau)|_1^{\frac43} d\tau\right\} \qquad (3.14)\\
&\le O(1)\left\{N(t)\int_0^t \|(\psi_x,\phi_x)(\tau)\|^2 d\tau + \varepsilon^{\frac14}\right\},
\end{aligned}
$$

$$
\begin{aligned}
R_3 &\le O(1)\int_0^t \|(\psi,\phi)(\tau)\|^{\frac12}\|(\psi_x,\phi_x)(\tau)\|^{\frac12}\|U_x(\tau)\|^2 d\tau \\
&\le O(1)\left\{N(t)\int_0^t \|(\psi_x,\phi_x)(\tau)\|^2 d\tau + \int_0^t \|U_x(\tau)\|^{\frac83} d\tau\right\} \qquad (3.15)\\
&\le O(1)\left\{N(t)\int_0^t \|(\psi_x,\phi_x)(\tau)\|^2 d\tau + \varepsilon^{\frac14}\right\},
\end{aligned}
$$

and

$$
\begin{aligned}
R_2 &\le O(1)\int_0^t \|(\varphi,\psi,\phi)(\tau)\|^{\frac12}\|(\varphi_x,\psi_x,\phi_x)(\tau)\|^{\frac32}\|U_x(\tau)\| d\tau \\
&\le O(1)\left\{N(t)^{\frac23}\int_0^t \|(\varphi_x,\psi_x,\phi_x)(\tau)\|^2 d\tau + \int_0^t \|U_x(\tau)\|^4 d\tau\right\} \qquad (3.16)\\
&\le O(1)\left\{N(t)^{\frac23}\int_0^t \|(\varphi_x,\psi_x,\phi_x)(\tau)\|^2 d\tau + \varepsilon^{\frac14}\right\}.
\end{aligned}
$$

Here we have used the following inequality:

$$
\begin{aligned}
\int_0^t \left|\frac{\partial^i}{\partial x^i}U(\tau)\right|_p^{a+b} d\tau &\le \sup_{[0,t]}\left(\left|\frac{\partial^i}{\partial x^i}U(\tau)\right|_p^a\right)\int_0^t \left|\frac{\partial^i}{\partial x^i}U(\tau)\right|_p^b d\tau \\
&\le O(1)\varepsilon^{\left(i-\frac1p\right)a}\int_0^t \left|\frac{\partial^i}{\partial x^i}U(\tau)\right|_p^b d\tau.
\end{aligned}
$$

Substituting (3.14)–(3.16) into (3.13), we have by using the fact that $N(t)$ is sufficiently small that

$$
\begin{aligned}
&\|(\varphi,\psi,\phi)(t)\|^2 + \int_0^t \{\|\sqrt{V_t(\tau)}(\varphi,\phi)(\tau)\|^2 + \|(\psi_x,\phi_x)(\tau)\|^2\} d\tau \\
&\le C(\eta)\left\{\|(\varphi_0,\psi_0,\phi_0)\|^2 + \varepsilon^{\frac14} + N(t)^{\frac23}\int_0^t \|\varphi_x(\tau)\|^2 d\tau\right\}. \qquad (3.17)
\end{aligned}
$$

Now we turn to dealing with the term $\int_0^t \|\varphi_x(\tau)\|^2 d\tau$. To do so, we have from

(3.1) and (3.2) that

$$\begin{aligned}
&\left\{\frac{\mu}{2}\left(\frac{\varphi_x}{v}\right)^2-\frac{\varphi_x}{v}\psi\right\}_t-p_v(v,\theta)\frac{\varphi_x^2}{v}-\left(\frac{\psi_x^2}{v}+\frac{p_\theta(v,\theta)\varphi_x\phi_x}{v}\right)+\left(\frac{\psi\psi_x}{v}\right)_x\\
&\quad=\left\{V_x\left[p_v(v,\theta)-p_v(V,\Theta)\right]\frac{\varphi_x}{v}+\Theta_x\left[p_\theta(v,\theta)-p_\theta(V,\Theta)\right]\frac{\varphi_x}{v}\right.\\
&\qquad\left.+\frac{U_x\psi\varphi_x}{v^2}-\frac{V_x\psi\psi_x}{v^2}\right\}\\
&\qquad+\mu\frac{V_x\psi_x\varphi_x}{v^3}-\mu\frac{U_{xx}\varphi_x}{v^2}+\mu\frac{V_xU_x\varphi_x}{v^3}.
\end{aligned}\tag{3.18}$$

Integrating (3.18) with respect to t and x over $[0,t]\times\mathbf{R}$, we have from (3.10), (3.11), and $p_v(v,\theta)<0$ that

$$\begin{aligned}
&\|\varphi_x(t)\|^2+\int_0^t\|\varphi_x(\tau)\|^2d\tau\\
&\quad\le C(\eta)\left\{\|(\varphi_{0x},\psi_0)\|^2+\|\psi(t)\|^2\right.\\
&\qquad\left.+\int_0^t(\|(\psi_x,\phi_x)(\tau)\|^2+\|\sqrt{V_t(\tau)}(\varphi,\phi)(\tau)\|^2)d\tau+\sum_{j=4}^5R_j\right\}.
\end{aligned}\tag{3.19}$$

Here

$$\begin{cases}
R_4=\displaystyle\int_0^t\int_{\mathbf{R}}(|\psi U_x\varphi_x|+|V_x\psi\psi_x|)(\tau,x)dxd\tau,\\
R_5=\displaystyle\int_0^t\int_{\mathbf{R}}(|V_x\psi_x\varphi_x|+|U_{xx}\varphi_x|+|U_x^2\varphi_x|)(\tau,x)dxd\tau.
\end{cases}$$

Since

$$\begin{aligned}
R_4&\le O(1)\int_0^t\|\psi(\tau)\|^{\frac12}\|\psi_x(\tau)\|^{\frac12}\|U_x(\tau)\|\|(\varphi_x,\psi_x)(\tau)\|d\tau\\
&\le O(1)\left\{N(t)^{\frac23}\int_0^t\|(\varphi_x,\psi_x)(\tau)\|^2d\tau+\int_0^t\|U_x(\tau)\|^4d\tau\right\}\\
&\le O(1)\left\{N(t)^{\frac23}\int_0^t\|(\varphi_x,\psi_x)(\tau)\|^2d\tau+\varepsilon^{\frac14}\right\}
\end{aligned}\tag{3.20}$$

and

$$\begin{aligned}
R_5&\le\frac12\int_0^t\|\varphi_x(\tau)\|^2d\tau+O(1)\left\{\int_0^t\|\psi_x(\tau)\|^2d\tau\right.\\
&\qquad\left.+\int_0^t\left(|U_x(\tau)|_4^4+|U_{xx}(\tau)|_1^2\right)d\tau\right\}\\
&\le\frac12\int_0^t\|\varphi_x(\tau)\|^2d\tau+O(1)\left\{\int_0^t\|\psi_x(\tau)\|^2d\tau+\varepsilon^{\frac14}\right\},
\end{aligned}\tag{3.21}$$

we get from (3.19)–(3.21) and the fact that $N(t)$ can be chosen sufficiently small that

$$\begin{aligned}
&\|\varphi_x(t)\|^2+\int_0^t\|\varphi_x(\tau)\|^2d\tau\\
&\le C(\eta)\left\{\|(\varphi_{0x},\psi_0)\|^2+\|\psi(t)\|^2\right.\\
&\qquad\left.+\int_0^t(\|(\psi_x,\phi_x)(\tau)\|^2+\|\sqrt{V_t(\tau)}(\varphi,\phi)(\tau)\|^2)d\tau+\varepsilon^{\frac14}\right\}.
\end{aligned}\tag{3.22}$$

Multiplying (3.17) by a suitably large positive constant λ and adding the resultant inequality to (3.22), we have from the fact that $N(t)$ is sufficiently small that

$$\begin{aligned}
&\|(\varphi,\psi,\phi,\varphi_x)(t)\|^2+\int_0^t\{\|\sqrt{V_t(\tau)}(\varphi,\phi)(\tau)\|^2+\|(\varphi_x,\psi_x,\phi_x)(\tau)\|^2\}d\tau\\
&\le C(\eta)\{\|(\varphi_0,\psi_0,\phi_0,\varphi_{0x})\|^2+\varepsilon^{\frac14}\}.
\end{aligned}\tag{3.23}$$

Equation (3.23) is the so-called basic energy estimate. Having obtained this, the other higher order energy estimates can easily be obtained by exploiting the same argument. For completeness, we show only how to estimate $\|(\psi_x,\phi_x)(t)\|^2$ in the following.

We first estimate $\|\psi_x(t)\|^2$. To this end, multiplying $(3.2)_2$ by $-\psi_{xx}(t,x)$, we have

$$\begin{aligned}
&\left(\frac{\psi_x^2}{2}\right)_t+\mu\frac{\psi_{xx}^2}{v}-(\psi_t\psi_x)_x\\
&=(p_v(v,\theta)\varphi_x+p_\theta(v,\theta)\phi_x)\psi_{xx}+\mu\frac{\varphi_x\psi_x\psi_{xx}}{v^2}\\
&\quad+\{V_x[p_v(v,\theta)-p_v(V,\Theta)]\psi_{xx}+\Theta_x[p_\theta(v,\theta)-p_\theta(V,\Theta)]\psi_{xx}\}\\
&\quad+\left(\mu\frac{V_x\psi_x\psi_{xx}}{v^2}+\mu\frac{U_x\varphi_x\psi_{xx}}{v^2}\right)-\mu\frac{U_{xx}\psi_{xx}}{v}+\mu\frac{V_xU_x\psi_{xx}}{v^2}.
\end{aligned}$$

Integrating the above identity with respect to t and x over $[0,t]\times\mathbf{R}$, we have from (3.7), (3.10), and (3.11) that

$$\begin{aligned}
&\|\psi_x(t)\|^2+\int_0^t\|\psi_{xx}(\tau)\|^2d\tau\\
&\le C(\eta)\left\{\|\psi_{0x}\|^2+\int_0^t\|(\varphi_x,\psi_x,\phi_x)(\tau)\|^2d\tau+\int_0^t\|\sqrt{V_t(\tau)}(\varphi,\phi)(\tau)\|^2d\tau\right.\\
&\qquad+\int_0^t\int_{\mathbf{R}}[|\varphi_x\psi_x\psi_{xx}|+|V_x\psi_x\psi_{xx}|+|U_x\varphi_x\psi_{xx}|\\
&\qquad\left.+|\psi_{xx}|(|U_{xx}|+|U_x^2|)](\tau,x)dxd\tau\right\}.
\end{aligned}\tag{3.24}$$

Due to

$$\begin{aligned}
&\int_0^t \int_{\mathbf{R}} |(\varphi_x \psi_x \psi_{xx})(\tau, x)| dx d\tau \\
&\qquad \le O(1) \int_0^t \|\varphi_x(\tau)\| \|\psi_x(\tau)\|^{\frac{1}{2}} \|\psi_{xx}(\tau)\|^{\frac{3}{2}} d\tau \qquad (3.25)\\
&\qquad \le \frac{1}{4} \int_0^t \|\psi_{xx}(\tau)\|^2 d\tau + O(1) N(t)^4 \int_0^t \|\psi_x(\tau)\|^2 d\tau,
\end{aligned}$$

$$\begin{aligned}
&\int_0^t \int_{\mathbf{R}} |\psi_{xx}(\tau, x)| (|V_x \psi_x| + |U_x \varphi_x|)(\tau, x) dx d\tau \\
&\qquad \le \frac{1}{4} \int_0^t \|\psi_{xx}(\tau)\|^2 d\tau + O(1) \int_0^t \|(\varphi_x, \psi_x)(\tau)\|^2 d\tau, \qquad (3.26)
\end{aligned}$$

$$\begin{aligned}
&\int_0^t \int_{\mathbf{R}} |\psi_{xx}(\tau, x)| (|U_{xx}| + |U_x^2|)(\tau, x) dx d\tau \\
&\qquad \le \frac{1}{4} \int_0^t \|\psi_{xx}(\tau)\|^2 d\tau + O(1) \int_0^t \left(\|U_{xx}(\tau)\|^2 + |U_x(\tau)|_4^4 \right) d\tau \qquad (3.27)\\
&\qquad \le \frac{1}{4} \int_0^t \|\psi_{xx}(\tau)\|^2 d\tau + O(1) \varepsilon^{\frac{1}{4}},
\end{aligned}$$

inserting (3.25)–(3.27) into (3.24), we deduce from (3.23) that

$$\|\psi_x(t)\|^2 + \int_0^t \|\psi_{xx}(\tau)\|^2 d\tau \le C(\eta) \{ \|(\varphi_0, \psi_0, \phi_0)\|_1^2 + \varepsilon^{\frac{1}{4}} \}. \qquad (3.28)$$

Finally, we estimate $\|\phi_x(t)\|^2$. To do so, we multiply $(3.2)_3$ by $-\phi_{xx}(t, x)$ to get

$$\begin{aligned}
&\left(\frac{\phi_x^2}{2} \right)_t + \frac{\kappa}{v e_\theta(v, \theta)} \phi_{xx}^2 - (\phi_x \phi_t)_x \\
&\quad = \frac{\theta p_\theta(v, \theta)}{e_\theta(v, \theta)} \psi_x \phi_{xx} \\
&\quad + \left\{ \frac{\kappa}{e_\theta(v, \theta)} \frac{\varphi_x \phi_x}{v^2} - \frac{\mu}{e_\theta(v, \theta)} \frac{\psi_x^2}{v} + U_x \left[\frac{\theta p_\theta(v, \theta)}{e_\theta(v, \theta)} - \frac{\Theta p_\theta(V, \Theta)}{e_\theta(V, \Theta)} \right] \right\} \phi_{xx} \qquad (3.29)\\
&\quad + \left\{ \frac{\kappa}{e_\theta(v, \theta)} \left(\frac{V_x \phi_x \phi_{xx}}{v^2} + \frac{\Theta_x \varphi_x \phi_{xx}}{v^2} \right) - \frac{2\mu}{e_\theta(v, \theta)} \frac{V_x \psi_x \phi_{xx}}{v} \right\} \\
&\quad - \frac{\kappa}{e_\theta(v, \theta)} \frac{\Theta_{xx} \phi_{xx}}{v} + \left\{ \frac{\kappa}{e_\theta(v, \theta)} \frac{V_x \Theta_x \phi_{xx}}{v^2} + \frac{\mu}{e_\theta(v, \theta)} \frac{U_x^2 \phi_{xx}}{v} \right\}.
\end{aligned}$$

Integrating (3.29) with respect to t and x over $[0, t] \times \mathbf{R}$, we get from (3.7), (3.10),

and (3.11) that

$$
\begin{aligned}
&\|\phi_x(t)\|^2+\int_0^t\|\phi_{xx}(\tau)\|^2d\tau\\
&\quad\le C(\eta)\Big\{\|\phi_{0x}\|^2+\int_0^t\|\psi_x(\tau)\|^2d\tau\\
&\qquad+\int_0^t\int_{\mathbf{R}}[|\phi_{xx}|(|\varphi_x\phi_x|+|\psi_x^2|)+|U_x\phi_{xx}|(|\phi|+|\varphi|)\\
&\qquad\qquad+|\phi_{xx}|(|V_x\phi_x|+|\Theta_x\varphi_x|+|U_x\psi_x|)\\
&\qquad\qquad+|\phi_{xx}|(|U_{xx}|+|U_x|^2)](\tau,x)dxd\tau\Big\}.
\end{aligned}
\tag{3.30}
$$

Since

$$
\begin{aligned}
&\int_0^t\int_{\mathbf{R}}|(\varphi_x\phi_x\phi_{xx})(\tau,x)|dxd\tau\\
&\quad\le O(1)\int_0^t\|\varphi_x(\tau)\|\|\phi_x(\tau)\|^{\frac12}\|\phi_{xx}(\tau)\|^{\frac32}d\tau\\
&\quad\le\frac18\int_0^t\|\phi_{xx}(\tau)\|^2d\tau+O(1)N(t)^4\int_0^t\|\phi_x(\tau)\|^2d\tau,
\end{aligned}
\tag{3.31}
$$

$$
\begin{aligned}
&\int_0^t\int_{\mathbf{R}}|(\psi_x^2\phi_{xx})(\tau,x)|dxd\tau\\
&\quad\le O(1)\int_0^t\|\psi_x(\tau)\|^{\frac32}\|\psi_{xx}(\tau)\|^{\frac12}\|\phi_{xx}(\tau)\|d\tau\\
&\quad\le\frac18\int_0^t\|\phi_{xx}(\tau)\|^2d\tau+O(1)N(t)^{\frac32}\int_0^t\left(\|\psi_x(\tau)\|^2+\|\psi_{xx}(\tau)\|^2\right)d\tau,
\end{aligned}
\tag{3.32}
$$

$$
\begin{aligned}
&\int_0^t\int_{\mathbf{R}}|(U_x\phi_{xx})(\tau,x)|(|\varphi|+|\phi|)(\tau,x)dxd\tau\\
&\quad\le\frac18\int_0^t\|\phi_{xx}(\tau)\|^2d\tau+O(1)\int_0^t\|\sqrt{V_t(\tau)}(\varphi,\phi)(\tau)\|^2d\tau,
\end{aligned}
\tag{3.33}
$$

$$
\begin{aligned}
&\int_0^t\int_{\mathbf{R}}|\phi_{xx}(\tau,x)|(|V_x\phi_x|+|\Theta_x\varphi_x|+|U_x\psi_x|)(\tau,x)dxd\tau\\
&\quad\le\frac18\int_0^t\|\phi_{xx}(\tau)\|^2d\tau+O(1)\int_0^t\|(\varphi_x,\psi_x,\phi_x)(\tau)\|^2d\tau,
\end{aligned}
\tag{3.34}
$$

$$
\begin{aligned}
&\int_0^t\int_{\mathbf{R}}|\phi_{xx}(\tau,x)|(|U_{xx}|+|U_x^2|)(\tau,x)dxd\tau\\
&\quad\le\frac18\int_0^t\|\phi_{xx}(\tau)\|^2d\tau+O(1)\int_0^t\left(\|U_{xx}(\tau)\|^2+|U_x(\tau)|_4^4\right)d\tau\\
&\quad\le\int_0^t\|\phi_{xx}(\tau)\|^2d\tau+O(1)\varepsilon^{\frac14},
\end{aligned}
\tag{3.35}
$$

we have from (3.30)–(3.35), (3.23), and (3.28) that

$$\|\phi_x(t)\|^2 + \int_0^t \|\phi_{xx}(\tau)\|^2 d\tau \le C(\eta)\{\|(\varphi_0,\psi_0,\phi_0)\|_1^2 + \varepsilon^{\frac{1}{4}}\}. \tag{3.36}$$

Putting (3.23), (3.28), and (3.36) together, we get that

$$\begin{aligned}&\|(\varphi,\psi,\phi)(t)\|_1^2 + \int_0^t \{\|\sqrt{V_t(\tau)}(\varphi(\tau),\phi(\tau))\|^2 + \|(\varphi_x,\psi_x,\phi_x)(\tau)\|_1^2\}d\tau\\ &\quad\le C(\eta)\{\|(\varphi_0,\psi_0,\phi_0)\|_1^2 + \varepsilon^{\frac{1}{4}}\}.\end{aligned} \tag{3.37}$$

By repeating the same argument, we can deduce that

$$\begin{aligned}&\|(\varphi_{xx},\psi_{xx},\phi_{xx})(t)\|^2 + \int_0^t \|(\psi_{xxx},\phi_{xxx},\varphi_{xx})(\tau)\|^2 d\tau\\ &\quad\le C(\eta)\{\|(\varphi_0,\psi_0,\phi_0)\|_2^2 + \varepsilon^{\frac{1}{4}}\},\end{aligned} \tag{3.38}$$

and (3.8) follows immediately from (3.37) and (3.38). This completes the proof of Theorem 1.1.

4. The proof of Theorem 1.2. Note that for the ideal polytropic gas, $p(v,\theta)$ and $e(v,\theta)$ satisfy the special constitutive relations (1.19). From (1.19) and the assumptions listed in Theorem 1.2, we can get

$$\|\phi_0\|_2 \le C(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V})(\gamma-1)\|(\varphi_0,\xi_0)\|_2. \tag{4.1}$$

Moreover, the entropy $\eta(v,u,\theta;V,U,\Theta)$ defined by (1.14) takes the form

$$\eta(v,u,\theta;V,U,\Theta) = R\Theta\Phi\left(\frac{v}{V}\right) + \frac{1}{2}(u-U)^2 + \frac{R\Theta}{\gamma-1}\Phi\left(\frac{\theta}{\Theta}\right) \tag{4.2}$$

with

$$\Phi(s) = s - \ln s - 1. \tag{4.3}$$

Since we want to get a global stability result, the techniques used in section 3 no longer apply. To overcome this difficulty, our main idea is the following: we look for solution $(\varphi(t,x),\psi(t,x),\phi(t,x))$ of (3.1), (3.2) in the solution space

$$X(0,\infty) := \bigcup_{0<\underline{m}<\overline{m}<\infty,0<M<\infty} \ \bigcap_{0\le t_1<t_2<\infty} X_{\underline{m},\overline{m},M}(t_1,t_2;\underline{\Theta},\overline{\Theta}),$$

where for t_1,t_2 $(0\le t_1<t_2<\infty)$ and $\underline{m}$, $\overline{m}$, M $(0<\underline{m}<\overline{m}<\infty,\ 0<M<\infty)$,

$$X_{\underline{m},\overline{m},M}(t_1,t_2;\underline{\Theta},\overline{\Theta}) = \left\{ (\varphi,\psi,\phi)(t,x) \;\middle|\; \begin{array}{c} (\varphi,\psi,\phi)(t,x)\in C^0([t_1,t_2];H^2(\mathbf{R}))\\ (\psi_x,\phi_x)(t,x)\in L^2(t_1,t_2;H^2(\mathbf{R}))\\ 0<\underline{\Theta}\le\phi(t,x)+\Theta(t,x)\le\overline{\Theta}\\ 0<\underline{m}\le\varphi(t,x)+V(t,x)\le\overline{m}\\ \sup\limits_{[t_1,t_2]}\{\|(\varphi,\psi,\phi)(t,x)\|_2\}\le M \end{array} \right\}. \tag{4.4}$$

By the local existence result established in [15] and [21] and from the assumptions listed in Theorem 1.2, we know that the Cauchy problem (3.1), (3.2) admits a unique smooth solution $(\varphi,\psi,\phi)(t,x)\in X_{\underline{V},\overline{V},M}(0,t_0:\underline{\Theta},\overline{\Theta})$ for some sufficiently small positive constant $t_0>0$ with $M=2\|(\varphi_0,\psi_0,\phi_0)\|_2$. Now suppose that such a solution has been extended to the time step $t=T$ with $(\varphi,\psi,\phi)(t,x)\in X(0,T)$ for some $T>0$ and satisfies the following a priori estimates: For each $(t,x)\in[0,T]\times\mathbf{R}$,

$$\begin{cases} 0<m_1\le v(t,x)=\varphi(t,x)+V(t,x)\le M_1,\\ 0<\underline{\Theta}\le\theta(t,x)=\phi(t,x)+\Theta(t,x)\le\overline{\Theta}. \end{cases} \tag{4.5}$$

Based on the a priori assumption (4.5), if we can show that there exists a positive constant $C_2(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V})$ which depends only on $\underline{\Theta}$, $\overline{\Theta}$, $\underline{V}$, $\overline{V}$, the initial data $(\varphi_0(x),\psi_0(x),\phi_0(x))$, and the system but is independent of m_1,M_1 such that for $(t,x)\in[0,T]\times\mathbf{R}$,

$$\begin{cases} 0<(C_2(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}))^{-1}\le v(t,x)=\varphi(t,x)+V(t,x)\le C_2(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}),\\ 0<\underline{\Theta}<\theta(t,x)=\phi(t,x)+\Theta(t,x)<\overline{\Theta}, \end{cases} \tag{4.6}$$

then, by combining the local existence result with the continuity argument, we can prove Theorem 1.2 easily. Thus, in the rest of this section, we pay our attention only to deducing (4.6) based on the a priori assumptions (4.5).

First, from (3.12), (1.19), (4.2), and (4.3), we have

$$\begin{aligned} &\int_{\mathbf{R}}\left\{R\Theta\Phi\left(\frac{v}{V}\right)+\frac{1}{2}\psi^2+\frac{R\Theta}{\gamma-1}\Phi\left(\frac{\theta}{\Theta}\right)\right\}(t,x)dx\\ &\quad+\int_0^t\int_{\mathbf{R}}\left\{\mu\Theta\frac{\psi_x^2}{v\theta}+\kappa\Theta\frac{\phi_x^2}{v\theta^2}\right\}(\tau,x)dxd\tau\\ &\quad+\int_0^t\int_{\mathbf{R}}\left\{\tilde{p}_v(v,s)-\tilde{p}_v(V,\overline{s})\varphi-\tilde{p}_s(V,\overline{s})\xi\right\}(\tau,x)U_x(\tau,x)dxd\tau\\ &=\int_{\mathbf{R}}\left\{R\Theta\Phi\left(\frac{v}{V}\right)+\frac{1}{2}\psi^2+\frac{R\Theta}{\gamma-1}\Phi\left(\frac{\theta}{\Theta}\right)\right\}(0,x)dx+\sum_{j=6}^{8}R_j. \end{aligned} \tag{4.7}$$

Here

$$\begin{cases} R_6=\displaystyle\int_0^t\int_{\mathbf{R}}\left\{-\mu\frac{U_x\psi\varphi_x}{v^2}+2\mu\frac{U_x\psi_x\phi}{v\theta}-\kappa\frac{\Theta_x\phi\varphi_x}{v^2\theta}+\kappa\frac{\Theta_x\phi\phi_x}{v\theta^2}\right\}(\tau,x)dxd\tau,\\ R_7=\displaystyle\int_0^t\int_{\mathbf{R}}\left(\mu\frac{U_{xx}\psi}{v}+\kappa\frac{\Theta_{xx}\phi}{v\theta}\right)(\tau,x)dxd\tau,\\ R_8=\displaystyle\int_0^t\int_{\mathbf{R}}\left(-\mu\frac{V_xU_x\psi}{v^2}+\mu\frac{U_x^2\phi}{v\theta}+\frac{V_x\Theta_x\phi}{v^2\theta}\right)(\tau,x)dxd\tau. \end{cases}$$

From the assumptions listed in Theorem 1.2 and the a priori assumptions (4.5), we have from (4.7) that

$$\begin{aligned} &\int_{\mathbf{R}}\left\{\Phi\left(\frac{v}{V}\right)+\psi^2+\frac{\phi^2}{\gamma-1}\right\}(t,x)dx+\int_0^t\int_{\mathbf{R}}\left\{\frac{\psi_x^2}{v}+\frac{\phi_x^2}{v}\right\}(\tau,x)dxd\tau\\ &\quad\le C\left(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}\right)\left\|\left(\varphi_0,\psi_0,\frac{\phi_0}{\sqrt{\gamma-1}}\right)\right\|+\sum_{j=6}^{8}R_j', \end{aligned} \tag{4.8}$$

where $R_j' = C(\underline{\Theta}, \overline{\Theta}, \underline{V}, \overline{V})R_j$. Note that such a $C(\underline{\Theta}, \overline{\Theta}, \underline{V}, \overline{V})$ is independent of m_1 and M_1.

Now we estimate R_j' $(j = 6, 7, 8)$ term by term. First, from the a priori estimates $(4.5)_2$, we have for each given $\alpha > 0$ that

$$
\begin{aligned}
R_6' &\le \int_0^t \int_{\mathbf{R}} \left(\alpha \frac{\varphi_x^2}{v^3} + \frac{\psi_x^2 + \phi_x^2}{4v} \right)(\tau, x) dx d\tau \\
&\quad + C\left(\underline{\Theta}, \overline{\Theta}, \underline{V}, \overline{V}\right) \int_0^t \int_{\mathbf{R}} \left(\frac{\psi^2}{v} + \frac{\phi^2}{v} \right)(\tau, x) |U_x^2(\tau, x)| dx d\tau \\
&\le \int_0^t \int_{\mathbf{R}} \left(\alpha \frac{\varphi_x^2}{v^3} + \frac{\psi_x^2 + \phi_x^2}{4v} \right)(\tau, x) dx d\tau \\
&\quad + C\left(\underline{\Theta}, \overline{\Theta}, \underline{V}, \overline{V}\right) C(m_1, M_1) \int_0^t \left\| \left(\psi, \frac{\phi}{\sqrt{\gamma - 1}} \right)(\tau) \right\|^2 \|U_x(\tau)\|^4 d\tau \\
&\le \int_0^t \int_{\mathbf{R}} \left(\alpha \frac{\varphi_x^2}{v^3} + \frac{\psi_x^2 + \phi_x^2}{4v} \right)(\tau, x) dx d\tau \\
&\quad + C\left(\underline{\Theta}, \overline{\Theta}, \underline{V}, \overline{V}\right) C(m_1, M_1) \varepsilon^{\frac{1}{2}} \int_0^t (1 + \tau)^{-\frac{3}{2}} \left\| \left(\psi, \frac{\phi}{\sqrt{\gamma - 1}} \right)(\tau) \right\|^2 d\tau.
\end{aligned} \tag{4.9}
$$

Similarly, we get

$$
\begin{aligned}
R_7' &\le \frac{1}{4} \int_0^t \int_{\mathbf{R}} \left(\frac{\psi_x^2 + \phi_x^2}{v} \right)(\tau, x) dx d\tau \\
&\quad + C\left(\underline{\Theta}, \overline{\Theta}, \underline{V}, \overline{V}\right) \left\{ 1 + C(m_1, M_1) \varepsilon^{\frac{1}{2}} \int_0^t (1 + \tau)^{-\frac{9}{7}} \left\| \left(\psi, \frac{\phi}{\sqrt{\gamma - 1}} \right)(\tau) \right\|^2 d\tau \right\}
\end{aligned} \tag{4.10}
$$

and

$$
R_8' \le C(\underline{\Theta}, \overline{\Theta}, \underline{V}, \overline{V}) \left\{ 1 + C(m_1, M_1) \varepsilon^{\frac{1}{2}} \int_0^t (1 + \tau)^{-\frac{21}{16}} \left\| \left(\psi, \frac{\phi}{\sqrt{\gamma - 1}} \right)(\tau) \right\|^2 d\tau \right\}. \tag{4.11}
$$

Inserting (4.9)–(4.11) into (4.8), we deduce that

$$
\begin{aligned}
&\int_{\mathbf{R}} \left\{ \Phi\left(\frac{v}{V} \right) + \psi^2 + \frac{\phi^2}{\gamma - 1} \right\}(t, x) dx + \int_0^t \int_{\mathbf{R}} \left\{ \frac{\psi_x^2}{v} + \frac{\phi_x^2}{v} \right\}(\tau, x) dx d\tau \\
&\quad \le C\left(\underline{\Theta}, \overline{\Theta}, \underline{V}, \overline{V}\right) \left\{ 1 + \left\| \left(\varphi_0, \psi_0, \frac{\phi_0}{\sqrt{\gamma - 1}} \right) \right\| \right. \\
&\qquad\qquad + \alpha \int_0^t \int_{\mathbf{R}} \left(\frac{\varphi_x^2}{v^3} \right)(\tau, x) dx d\tau \\
&\qquad\qquad \left. + C(m_1, M_1) \varepsilon^{\frac{1}{2}} \int_0^t (1 + \tau)^{-\frac{9}{7}} \left\| \left(\psi, \frac{\phi}{\sqrt{\gamma - 1}} \right)(\tau) \right\|^2 d\tau \right\}.
\end{aligned} \tag{4.12}
$$

Now we turn to controlling the term $\int_0^t \int_{\mathbf{R}} (\frac{\varphi_x^2}{v^3})(\tau, x) dx d\tau$. To this end, since for the ideal polytropic gas, $p_v(v, \theta) = -\frac{R\theta}{v^2}$, $p_\theta(v, \theta) = \frac{R}{v}$, we have from (3.18) and $(4.5)_2$

that

(4.13)
$$\begin{aligned}
&\int_{\mathbf{R}} \left(\frac{\varphi_x^2}{v^2}\right)(t,x)dx + \int_0^t \int_{\mathbf{R}} \left(\frac{\varphi_x^2}{v^3}\right)(\tau,x)dxd\tau \\
&\quad \le C(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}) \left\{ \|(\varphi_{0x},\psi_0)\|^2 + \|\psi(t)\|^2 + \int_0^t \int_{\mathbf{R}} \left(\frac{\psi_x^2+\phi_x^2}{v}\right)(\tau,x)dxd\tau \right\} \\
&\qquad + \sum_{j=9}^{10} R_j.
\end{aligned}$$

Here

$$\begin{cases}
R_9 = C\left(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}\right) \left| \displaystyle\int_0^t \int_{\mathbf{R}} \left\{ V_x \left[-\frac{R\theta}{v^2} + \frac{R\Theta}{V^2} \right] \frac{\varphi_x}{v} + \Theta_x \left[\frac{R}{v} - \frac{R}{V} \right] \frac{\varphi_x}{v} \right. \right. \\
\qquad\qquad \left. \left. + \dfrac{U_x\psi\varphi_x}{v^2} - \dfrac{V_x\psi\psi_x}{v^2} \right\}(\tau,x)dxd\tau \right|, \\
R_{10} = C\left(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}\right) \left| \displaystyle\int_0^t \int_{\mathbf{R}} \left(-\mu\frac{U_{xx}\varphi_x}{v^2} + \mu\frac{V_x\varphi_x\psi_x}{v^3} + \mu\frac{V_xU_x\varphi_x}{v^2} \right)(\tau,x)dxd\tau \right|.
\end{cases}$$

Now we estimate R_9 and R_{10} term by term. First, for R_9, we get by the Cauchy–Schwarz inequality and (4.5) that

$$\begin{aligned}
R_9 &\le C\left(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}\right) \int_0^t \int_{\mathbf{R}} \left\{ \frac{|V_x\varphi_x|}{v^3}(|\varphi|^2+|\varphi|+|\phi|) \right. \\
&\qquad \left. + \frac{|\varphi_x(\Theta_x\varphi+U_x\psi)|}{v^2} + \frac{|V_x\psi\psi_x|}{v^3} \right\}(\tau,x)dxd\tau \\
&\le \frac{1}{3}\int_0^t \int_{\mathbf{R}} \left(\frac{\varphi_x^2}{v^3} + \frac{\psi_x^2}{v}\right)(\tau,x)dxd\tau \\
&\quad + C\left(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}\right) \int_0^t \int_{\mathbf{R}} \left\{ \frac{V_x^2}{v^3}(|\varphi|^4+|\varphi|^2+|\phi|^2) \right. \\
&\qquad \left. + \frac{\Theta_x^2\varphi^2+U_x^2\psi^2}{v} + \frac{V_x^2\psi^2}{v^5} \right\}(\tau,x)dxd\tau \qquad (4.14) \\
&\le \frac{1}{3}\int_0^t \int_{\mathbf{R}} \left(\frac{\varphi_x^2}{v^3} + \frac{\psi_x^2}{v}\right)(\tau,x)dxd\tau \\
&\quad + C\left(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}\right) C(m_1,M_1) \int_0^t |U_x(\tau)|_\infty^2 \left\| \left(\sqrt{\Phi\left(\frac{v}{V}\right)}, \psi, \frac{\phi}{\sqrt{\gamma-1}} \right)(\tau) \right\|^2 d\tau \\
&\le \frac{1}{3}\int_0^t \int_{\mathbf{R}} \left(\frac{\varphi_x^2}{v^3} + \frac{\psi_x^2}{v}\right)(\tau,x)dxd\tau \\
&\quad + C\left(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}\right) C(m_1,M_1)\varepsilon^{\frac{1}{2}} \int_0^t (1+\tau)^{-\frac{3}{2}} \left\| \left(\sqrt{\Phi\left(\frac{v}{V}\right)}, \psi, \frac{\phi}{\sqrt{\gamma-1}} \right)(\tau) \right\|^2 d\tau.
\end{aligned}$$

As to R_{10}, we have

$$
\begin{aligned}
R_{10} &\le \frac{1}{3}\int_0^t\int_{\mathbf{R}}\left(\frac{\varphi_x^2}{v^3}+\frac{\psi_x^2}{v}\right)(\tau,x)dxd\tau\\
&\quad + C\left(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}\right)\int_0^t\int_{\mathbf{R}}\left\{\frac{V_x^2\varphi_x^2}{v^5}+\frac{U_{xx}^2}{v}+\frac{U_x^4}{v^3}\right\}(\tau,x)dxd\tau\\
&\le \frac{1}{3}\int_0^t\int_{\mathbf{R}}\left(\frac{\varphi_x^2}{v^3}+\frac{\psi_x^2}{v}\right)(\tau,x)dxd\tau\\
&\quad + C\left(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}\right)C(m_1,M_1)\varepsilon^{\frac{1}{2}}\left\{\int_0^t(1+\tau)^{-\frac{3}{2}}\left\|\left(\frac{\varphi_x}{v}\right)(\tau)\right\|^2 d\tau+1\right\}.
\end{aligned}
\tag{4.15}
$$

Substituting (4.14) and (4.15) into (4.13), we deduce that

$$
\begin{aligned}
&\int_{\mathbf{R}}\left(\frac{\varphi_x^2}{v^2}\right)(t,x)dx+\int_0^t\int_{\mathbf{R}}\left(\frac{\varphi_x^2}{v^3}\right)(\tau,x)dxd\tau\\
&\quad\le C\left(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}\right)\left\{1+\|(\varphi_{0x},\psi_0)\|^2+\|\psi(t)\|^2+\int_0^t\int_{\mathbf{R}}\left(\frac{\psi_x^2+\phi_x^2}{v}\right)(\tau,x)dxd\tau\right\}\\
&\qquad + C\left(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}\right)C(m_1,M_1)\varepsilon^{\frac{1}{2}}\left\{\int_0^t(1+\tau)^{-\frac{3}{2}}\left\|\left(\sqrt{\Phi\left(\frac{v}{V}\right)},\psi,\right.\right.\right.\\
&\qquad\qquad\left.\left.\left.\frac{\phi}{\sqrt{\gamma-1}},\frac{\varphi_x}{v}\right)(\tau)\right\|^2 d\tau+1\right\}.
\end{aligned}
\tag{4.16}
$$

Multiplying (4.12) by a suitably large positive constant λ and adding the result to (4.16), we have by choosing $\alpha>0$ sufficiently small that

$$
\begin{aligned}
&\left\|\left(\sqrt{\Phi\left(\frac{v}{V}\right)},\psi,\frac{\phi}{\sqrt{\gamma-1}},\frac{\varphi_x}{v}\right)(t)\right\|^2+\int_0^t\left\|\left(\frac{\varphi_x}{v^{\frac{3}{2}}},\frac{\psi_x}{\sqrt{v}},\frac{\phi_x}{\sqrt{v}}\right)(\tau)\right\|^2 d\tau\\
&\quad\le C\left(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}\right)\left\{1+\left\|\left(\varphi_0,\psi_0,\frac{\phi_0}{\sqrt{\gamma-1}}\right)\right\|_1^2\right\}\\
&\qquad + C\left(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}\right)C(m_1,M_1)\varepsilon^{\frac{1}{2}}\left\{\int_0^t(1+\tau)^{-\frac{9}{7}}\left\|\left(\sqrt{\Phi\left(\frac{v}{V}\right)},\psi,\right.\right.\right.\\
&\qquad\qquad\left.\left.\left.\frac{\phi}{\sqrt{\gamma-1}},\frac{\varphi_x}{v}\right)(\tau)\right\|^2 d\tau+1\right\}.
\end{aligned}
\tag{4.17}
$$

To use the method of Kanel' (cf. [9]) to deduce a lower bound and an upper bound for $v(t,x)$, we need to estimate $\|(\frac{\tilde{v}_x}{\tilde{v}})(t)\|^2$, where $\tilde{v}=\frac{v}{V}$. In fact, since

$$
\frac{\tilde{v}_x}{\tilde{v}}=\frac{\varphi_x}{v}-\left(\frac{V_x}{v}-\frac{V_x}{V}\right),
$$

we have from $(4.5)_1$ that

$$
\begin{aligned}
\left\|\left(\frac{\tilde{v}_x}{\tilde{v}}\right)(t)\right\|^2 &\le 2\left\|\left(\frac{\varphi_x}{v}\right)(t)\right\|^2+C\left(\underline{\Theta},\overline{\Theta}\right)C(m_1,M_1)\|V_x(t)\|^2\\
&\le 2\left\|\left(\frac{\varphi_x}{v}\right)(t)\right\|^2+C\left(\underline{\Theta},\overline{\Theta}\right)C(m_1,M_1)\varepsilon.
\end{aligned}
\tag{4.18}
$$

Combining (4.17) with (4.18), we can deduce that there exists a positive constant $C_3(m_1, M_1)$ depending only on m_1, M_1 such that

(4.19)
$$\begin{aligned}
&\left\|\left(\sqrt{\Phi\left(\frac{v}{V}\right)}, \psi, \frac{\phi}{\sqrt{\gamma-1}}, \frac{\varphi_x}{v}, \frac{\tilde{v}_x}{\tilde{v}}\right)(t)\right\|^2 + \int_0^t \left\|\left(\frac{\varphi_x}{v^{\frac{3}{2}}}, \frac{\psi_x}{\sqrt{v}}, \frac{\phi_x}{\sqrt{v}}\right)(\tau)\right\|^2 d\tau \\
&\quad \le C\left(\underline{\Theta}, \overline{\Theta}, \underline{V}, \overline{V}\right)\left\{1 + \left\|\left(\varphi_0, \psi_0, \frac{\phi_0}{\sqrt{\gamma-1}}\right)\right\|_1^2\right\} \\
&\qquad + C\left(\underline{\Theta}, \overline{\Theta}, \underline{V}, \overline{V}\right) C_3(m_1, M_1)\varepsilon^{\frac{1}{2}}\left\{\int_0^t (1+\tau)^{-\frac{9}{7}}\left\|\left(\sqrt{\Phi\left(\frac{v}{V}\right)}, \psi, \right.\right.\right. \\
&\qquad\qquad \left.\left.\left.\frac{\phi}{\sqrt{\gamma-1}}, \frac{\varphi_x}{v}\right)(\tau)\right\|^2 d\tau + 1\right\}.
\end{aligned}$$

By choosing $\varepsilon < 1$ sufficiently small such that

$$C_3(m_1, M_1)\varepsilon^{\frac{1}{2}} < 1, \tag{4.20}$$

we have from Gronwall's inequality and (4.19) that

$$\begin{aligned}
&\left\|\left(\sqrt{\Phi\left(\frac{v}{V}\right)}, \psi, \frac{\phi}{\sqrt{\gamma-1}}, \frac{\varphi_x}{v}, \frac{\tilde{v}_x}{\tilde{v}}\right)(t)\right\|^2 + \int_0^t \left\|\left(\frac{\varphi_x}{v^{\frac{3}{2}}}, \frac{\psi_x}{\sqrt{v}}, \frac{\phi_x}{\sqrt{v}}\right)(\tau)\right\|^2 d\tau \\
&\quad \le C\left(\underline{\Theta}, \overline{\Theta}, \underline{V}, \overline{V}\right)\left\{1 + \left\|\left(\varphi_0, \psi_0, \frac{\phi_0}{\sqrt{\gamma-1}}\right)\right\|_1^2\right\}.
\end{aligned} \tag{4.21}$$

It is worth pointing out that the most important thing is that the constant on the right-hand side of (4.21) does not depend on m_1 and M_1.

Now we use the method of Kanel' to deduce the desired bounds on $v(t, x)$. To this end, let

$$\Psi(\tilde{v}) = \int_1^{\tilde{v}} \frac{\sqrt{\Phi(\eta)}}{\eta} d\eta, \quad \Phi(\eta) = \eta - \ln\eta - 1. \tag{4.22}$$

Since

$$\Psi(\tilde{v}) \to \begin{cases} -\infty & \text{as } \tilde{v} \to 0_+, \\ +\infty & \text{as } \tilde{v} \to +\infty \end{cases} \tag{4.23}$$

and

$$|\Psi(\tilde{v}(t,x))| = \left|\int_{-\infty}^x \frac{\partial}{\partial y}\Psi(\tilde{v}(t,y))\, dy\right| \le \frac{1}{2}\int_{\mathbf{R}}\left(\Phi\left(\frac{v}{V}\right) + \left(\frac{\tilde{v}_x}{\tilde{v}}\right)^2\right)(t,x)dx, \tag{4.24}$$

$(4.6)_1$ follows from (4.21)–(4.24).

Having obtained $(4.6)_1$, we deduce from (4.21) that

$$\begin{aligned}
&\left\|\left(\varphi, \psi, \frac{\phi}{\sqrt{\gamma-1}}, \varphi_x\right)(t)\right\|^2 + \int_0^t \left\|\left(\varphi_x, \psi_x, \phi_x\right)(\tau)\right\|^2 d\tau \\
&\quad \le C\left(\underline{\Theta}, \overline{\Theta}, \underline{V}, \overline{V}\right)\left\{1 + \left\|\left(\varphi_0, \psi_0, \frac{\phi_0}{\sqrt{\gamma-1}}\right)\right\|_1^2\right\}.
\end{aligned} \tag{4.25}$$

With $(4.6)_1$ and (4.25) in hand, similar to the argument used in section 2, we can get from the a priori assumption $(4.5)_2$ that

$$\begin{aligned}\left\|\left(\varphi_x,\psi_x,\frac{\phi_x}{\sqrt{\gamma-1}}\right)(t)\right\|^2+\int_0^t\left\|\left(\psi_{xx},\phi_{xx}\right)(\tau)\right\|^2d\tau\\ \le C\left(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}\right)\left\{1+\left\|\left(\varphi_0,\psi_0,\frac{\phi_0}{\sqrt{\gamma-1}}\right)\right\|_1^2\right\}.\end{aligned}\tag{4.26}$$

From (4.25) and (4.26), we get from (4.1) that

$$\begin{aligned}\|\phi(t)\|^2+\|\phi_x(t)\|^2&\le(\gamma-1)C\left(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}\right)\left\{1+\left\|\left(\varphi_0,\psi_0,\frac{\phi_0}{\sqrt{\gamma-1}}\right)\right\|_1^2\right\}\\ &\le(\gamma-1)C(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V})\{1+\|(\varphi_0,\psi_0,\xi_0)\|_1^2\}.\end{aligned}\tag{4.27}$$

Consequently,

$$\begin{aligned}\sup_{[0,T]}|\phi(t)|_\infty&\le\sup_{[0,T]}\{\|\phi(t)\|^{\frac12}\|\phi_x(t)\|^{\frac12}\}\\ &\le(\gamma-1)C(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V})\{1+\|(\varphi_0,\psi_0,\xi_0)\|_1^2\}.\end{aligned}\tag{4.28}$$

By choosing $\gamma-1$ sufficiently small such that

$$(\gamma-1)C(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V})\{1+\|(\varphi_0,\psi_0,\xi_0)\|_1^2\}<\min\left\{\frac{\overline{\Theta}}{2},\underline{\Theta}\right\},$$

we have for each $(t,x)\in[0,T]\times\mathbf{R}$ that

$$\theta(t,x)=\Theta(t,x)+\phi(t,x)\le\frac{\overline{\Theta}}{2}+(\gamma-1)C(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V})\{1+\|(\varphi_0,\psi_0,\xi_0)\|_1^2\}<\overline{\Theta}$$

and

$$\theta(t,x)=\Theta(t,x)+\phi(t,x)\ge2\underline{\Theta}-(\gamma-1)C(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V})\{1+\|(\varphi_0,\psi_0,\xi_0)\|_1^2\}>\underline{\Theta}.$$

This proves $(4.6)_2$. Note that the results of the above analysis also indicate that if $\|(\varphi_0(x),\psi_0(x),\xi_0(x))\|_1$ is bounded by a constant independent of $\frac{1}{\varepsilon}$, then the constant $C_2(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V})$ in (4.6) can also be chosen independent of $\frac{1}{\varepsilon}$. Thus from (4.20) and the continuity argument, to prove Theorem 1.2, we need only to take the fixed positive constant ε such that

$$0<\varepsilon<\min\{1,(C_3(\underline{V},\overline{V}))^{-2},(C_3((C_2(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V}))^{-1},C_2(\underline{\Theta},\overline{\Theta},\underline{V},\overline{V})))^{-2}\}.$$

This completes the proof of Theorem 1.2.

5. The proof of Theorem 1.3. In this section, we use the main idea used in proving Theorem 1.2 to prove Theorem 1.3. To do so, let $(\varphi,\psi)=(v-\overline{V},u-\overline{U})$; it is easy to check that (φ,ψ) solves

$$\begin{cases}\varphi_t-\psi_x=0,\\ \psi_t+[p(\varphi+\overline{V})-p(\overline{V})]_x-\mu\left(\dfrac{u_x}{v}-\dfrac{\overline{U}_x}{\overline{V}}\right)_x=\mu\left(\dfrac{\overline{U}_x}{\overline{V}}\right)_x,\end{cases}\tag{5.1}$$

with initial data

$$(\varphi,\psi)(t,x)|_{t=0}=(\varphi_0,\psi_0)(x)=(v_0(x)-\overline{V}(0,x),u_0(x)-\overline{U}(0,x)).\tag{5.2}$$

Similar to the proof of Theorem 1.2, all that we need to do is to show that, under the a priori assumption

$$0<m_1\le v(t,x)=\varphi(t,x)+\overline{V}(t,x)\le M_1\quad\forall(t,x)\in[0,T]\times\mathbf{R}\tag{5.3}$$

for some $T>0$, one can indeed deduce that there exists a positive constant $C_4(\underline{V},\overline{V})>0$ which depends only on the initial data $(\varphi_0(x),\psi_0(x))$ and the system but is independent of m_1 and M_1 such that

$$\begin{aligned}0<\left(C_4\left(\underline{V},\overline{V}\right)\right)^{-1}&\le v(t,x)=\varphi(t,x)+\overline{V}(t,x)\\&\le C_4\left(\underline{V},\overline{V}\right)\quad\forall(t,x)\in[0,T]\times\mathbf{R}.\end{aligned}\tag{5.4}$$

To prove (5.4), we first perform some energy estimates. First, multiplying $(5.1)_1$ by $[p(\overline{V})-p(\overline{V}+\varphi)]$, $(5.1)_2$ by ψ, adding the resultant two identities, and integrating it with respect to t and x over $[0,t]\times\mathbf{R}$, we have from the assumptions listed in Theorem 1.3 and some integrations by parts that

$$\begin{aligned}&\left\|\left(\sqrt{\Phi(v,\overline{V})},\psi\right)(t)\right\|^2+\int_0^t\left\|\left(\frac{\psi_x}{\sqrt{v}}\right)(\tau)\right\|^2d\tau\\&\le C\left(\underline{V},\overline{V}\right)\left\{\|(\varphi_0,\psi_0)\|^2\right.\\&\qquad\left.+\int_0^t\int_{\mathbf{R}}\left(\left|\frac{\overline{V}_x\psi_x\varphi}{v}\right|+|\psi|(|\overline{U}_{xx}|+|\overline{U}_x^2|)\right)(\tau,x)dxd\tau\right\}.\end{aligned}\tag{5.5}$$

Here $\Phi(v,\overline{V})=p(\overline{V})\varphi-\int_{\overline{V}}^{v}p(s)ds$.

By exploiting the same argument to estimate $R_j'(j=6,7,8)$, we can get from the a priori assumption (5.3) that

$$
\begin{aligned}
&C\left(\underline{V},\overline{V}\right)\int_0^t\int_{\mathbf{R}}\left(\left|\frac{\overline{V}_x\psi_x\varphi}{v}\right|+|\psi|(|\overline{U}_{xx}|+|\overline{U}_x^2|)\right)(\tau,x)dxd\tau\\
&\le\frac{1}{2}\int_0^t\left\|\left(\frac{\psi_x}{\sqrt{v}}\right)(\tau)\right\|^2d\tau\\
&\quad+C\left(\underline{V},\overline{V}\right)C(m_1,M_1)\varepsilon^{\frac{1}{2}}\left\{1+\int_0^t(1+\tau)^{-\frac{7}{6}}\left\|\left(\sqrt{\Phi(v,\overline{V})},\psi\right)(\tau)\right\|^2d\tau\right\}.
\end{aligned}\tag{5.6}
$$

Combining (5.6) with (5.5), we deduce that

$$
\begin{aligned}
&\left\|\left(\sqrt{\Phi(v,\overline{V})},\psi\right)(t)\right\|^2+\int_0^t\left\|\left(\frac{\psi_x}{\sqrt{v}}\right)(\tau)\right\|^2d\tau\\
&\le C(\underline{V},\overline{V})\{1+\|(\varphi_0,\psi_0)\|^2\}\\
&\quad+C\left(\underline{V},\overline{V}\right)C(m_1,M_1)\varepsilon^{\frac{1}{2}}\left\{1+\int_0^t(1+\tau)^{-\frac{7}{6}}\left\|\left(\sqrt{\Phi(v,\overline{V})},\psi\right)(\tau)\right\|^2d\tau\right\}.
\end{aligned}\tag{5.7}
$$

Second, similar to (3.18), we have from (5.1) that

$$
\begin{aligned}
&\left[\frac{\mu}{2}\left(\frac{\varphi_x}{v}\right)^2-\frac{\varphi_x}{v}\psi\right]_t-\frac{\psi_x^2}{v}-p'\left(\varphi+\overline{V}\right)\frac{\varphi_x^2}{v}\\
&=[p'(\varphi+\overline{V})-p'(\overline{V})]\frac{\overline{V}_x\varphi_x}{v}-\mu\frac{\overline{U}_{xx}\varphi_x}{v^2}+\mu\frac{\overline{V}_x\varphi_x\psi_x}{v^3}+\mu\frac{\overline{V}_x\overline{U}_x\varphi_x}{v^3}\\
&\quad-\left(\frac{\psi\psi_x}{v}\right)_x-\frac{\psi\psi_x\overline{V}_x-\psi\varphi_x\overline{U}_x}{v^2}.
\end{aligned}\tag{5.8}
$$

Integrating (5.8) with respect to t and x over $[0,t]\times\mathbf{R}$, we have by some integrations by parts that

$$
\begin{aligned}
&\int_{\mathbf{R}}\left\{\frac{\mu}{2}\left(\frac{\varphi_x}{v}\right)^2-\frac{\varphi_x}{v}\psi\right\}(t,x)dx-\int_0^t\int_{\mathbf{R}}\left(p'\left(\overline{V}+\varphi\right)\frac{\varphi_x^2}{v}\right)(\tau,x)dxd\tau\\
&=\int_{\mathbf{R}}\left\{\frac{\mu}{2}\left(\frac{\varphi_{0x}}{v_0}\right)^2-\frac{\varphi_{0x}}{v_0}\psi_0\right\}(x)dx+\int_0^t\int_{\mathbf{R}}\left(\frac{\psi_x^2}{v}\right)(\tau,x)dxd\tau\\
&\quad+\int_0^t\int_{\mathbf{R}}\left(\left[p'\left(\varphi+\overline{V}\right)-p'\left(\overline{V}\right)\right]\frac{\overline{V}_x\varphi_x}{v}\right)(\tau,x)dxd\tau\\
&\quad-\mu\int_0^t\int_{\mathbf{R}}\left(\frac{\overline{U}_{xx}\varphi_x}{v^2}\right)(\tau,x)dxd\tau+\mu\int_0^t\int_{\mathbf{R}}\left(\frac{\overline{V}_x\varphi_x\psi_x}{v^3}\right)(\tau,x)dxd\tau\\
&\quad+\mu\int_0^t\int_{\mathbf{R}}\left(\frac{\overline{V}_x\overline{U}_x\varphi_x}{v^3}\right)(\tau,x)dxd\tau-\int_0^t\int_{\mathbf{R}}\left(\frac{\overline{V}_x\psi\psi_x-\overline{U}_x\varphi_x\psi}{v^2}\right)(\tau,x)dxd\tau.
\end{aligned}\tag{5.9}
$$

Noticing the a priori assumption (5.3) on $v(t,x)$, we have from the Cauchy–Schwarz inequality that

$$
\begin{aligned}
&\int_0^t \int_{\mathbf{R}} \left([p'(\varphi+\overline{V}) - p'(\overline{V})] \frac{\overline{V}_x \varphi_x}{v} \right)(\tau,x) dx d\tau \\
&\le C(m_1,M_1) \int_0^t |\overline{V}_x(\tau)|_\infty \|\varphi(\tau)\| \|\varphi_x(\tau)\| d\tau \\
&\le C(m_1,M_1) \int_0^t |\overline{V}_x(\tau)|_\infty \left\| \sqrt{\Phi(v,\overline{V})}(\tau) \right\| \left\| \left(\sqrt{\frac{-p'(\overline{V}+\varphi)}{v}} \varphi_x \right)(\tau) \right\| d\tau \qquad (5.10) \\
&\le \frac{1}{6} \int_0^t \left\| \left(\sqrt{\frac{-p'(\overline{V}+\varphi)}{v}} \varphi_x \right)(\tau) \right\|^2 d\tau \\
&\quad + C(m_1,M_1) \varepsilon^{\frac{1}{2}} \int_0^t (1+\tau)^{-\frac{3}{2}} \left\| \sqrt{\Phi(v,\overline{V})}(\tau) \right\|^2 d\tau, \\
&\mu \int_0^t \int_{\mathbf{R}} \left(\frac{\overline{U}_{xx} \varphi_x}{v^2} \right)(\tau,x) dx d\tau \\
&\le C(m_1,M_1) \int_0^t \left\| \left(\sqrt{\frac{-p'(\overline{V}+\varphi)}{v}} \varphi_x \right)(\tau) \right\| \|\overline{U}_{xx}(\tau)\| d\tau \qquad (5.11) \\
&\le \frac{1}{6} \int_0^t \left\| \left(\sqrt{\frac{-p'(\overline{V}+\varphi)}{v}} \varphi_x \right)(\tau) \right\|^2 d\tau + C(m_1,M_1) \varepsilon^{\frac{1}{2}}, \\
&\mu \int_0^t \int_{\mathbf{R}} \left(\frac{\overline{V}_x \psi_x \varphi_x}{v^3} \right)(\tau,x) dx d\tau \\
&\le C(m_1,M_1) \int_0^t |\overline{V}_x(\tau)|_\infty \left\| \left(\frac{\psi_x}{\sqrt{v}} \right)(\tau) \right\| \left\| \left(\frac{\varphi_x}{v} \right)(\tau) \right\| d\tau \qquad (5.12) \\
&\le \int_0^t \left\| \left(\frac{\psi_x}{v} \right)(\tau) \right\|^2 d\tau + C(m_1,M_1) \varepsilon^{\frac{1}{2}} \int_0^t (1+\tau)^{-\frac{3}{2}} \left\| \left(\frac{\varphi_x}{v} \right)(\tau) \right\|^2 d\tau,
\end{aligned}
$$

$$
\begin{aligned}
&\mu \int_0^t \int_{\mathbf{R}} \left(\frac{\overline{V}_x \overline{U}_x \varphi_x}{v^3} \right)(\tau,x) dx d\tau \\
&\le C(m_1,M_1) \int_0^t \left\| \left(\sqrt{\frac{-p'(\overline{V}+\varphi)}{v}} \varphi_x \right)(\tau) \right\| |\overline{V}_x(\tau)|_4^2 d\tau \qquad (5.13) \\
&\le \frac{1}{6} \int_0^t \left\| \left(\sqrt{\frac{-p'(\overline{V}+\varphi)}{v}} \varphi_x \right)(\tau) \right\|^2 d\tau + C(m_1,M_1) \varepsilon^{\frac{1}{2}},
\end{aligned}
$$

and

$$
\begin{aligned}
&\mu\int_0^t\int_{\mathbf{R}}\left(\frac{\overline{V}_x\psi_x\psi-\overline{U}_x\psi\varphi_x}{v^2}\right)(\tau,x)dxd\tau\\
&\quad\le C(m_1,M_1)\int_0^t|\overline{V}_x(\tau)|_\infty(\|\psi_x(\tau)\|\|\psi(\tau)\|+\|\varphi_x(\tau)\|\|\psi(\tau)\|)d\tau\\
&\quad\le C(m_1,M_1)\int_0^t|\overline{V}_x(\tau)|_\infty\left(\left\|\left(\frac{\psi_x}{\sqrt{v}}\right)(\tau)\right\|\|\psi(\tau)\|\right.\\
&\qquad\qquad\left.+\left\|\left(\sqrt{\frac{-p'\left(\overline{V}+\varphi\right)}{v}}\varphi_x\right)(\tau)\right\|\|\psi(\tau)\|\right)d\tau\\
&\quad\le\frac{1}{6}\int_0^t\left\|\left(\sqrt{\frac{-p'\left(\overline{V}+\varphi\right)}{v}}\varphi_x\right)(\tau)\right\|^2d\tau+\int_0^t\int_{\mathbf{R}}\left(\frac{\psi_x^2}{v}\right)(\tau)dxd\tau\\
&\qquad+C\left(m_1,M_1\right)\varepsilon^{\frac{1}{2}}\int_0^t(1+\tau)^{-\frac{3}{2}}\|\psi(\tau)\|^2d\tau.
\end{aligned}
\tag{5.14}
$$

Substituting (5.10)–(5.14) into (5.9), we arrive at

$$
\begin{aligned}
&\int_{\mathbf{R}}\left\{\frac{\mu}{2}\left(\frac{\varphi_x}{v}\right)^2-\frac{\varphi_x}{v}\psi\right\}(t,x)dx-\int_0^t\int_{\mathbf{R}}\left(p'\left(\overline{V}+\varphi\right)\frac{\varphi_x^2}{v}\right)(\tau,x)dxd\tau\\
&\quad\le C\left(\underline{V},\overline{V}\right)\left\{1+\|(\varphi_{0x},\psi_0)\|^2+\int_0^t\int_{\mathbf{R}}\left(\frac{\psi_x^2}{v}\right)(\tau,x)dxd\tau\right\}\\
&\qquad+C(m_1,M_1)\varepsilon^{\frac{1}{2}}\left\{1+\int_0^t(1+\tau)^{-\frac{3}{2}}\left\|\left(\sqrt{\Phi(v,\overline{V})},\psi,\frac{\varphi_x}{v}\right)(\tau)\right\|^2d\tau\right\}.
\end{aligned}
\tag{5.15}
$$

Based on (5.7) and (5.15), we conclude that there exists a positive constant $C_5(m_1,M_1)$ such that

$$
\begin{aligned}
&\left\|\left(\sqrt{\Phi(v,\overline{V})},\psi,\frac{\varphi_x}{v}\right)(t)\right\|^2+\int_0^t\left\|\left(\frac{\psi_x}{\sqrt{v}},\sqrt{\frac{-p'(\overline{V}+\varphi)}{v}}\varphi_x\right)(\tau)\right\|^2d\tau\\
&\quad\le C(\underline{V},\overline{V})(1+\|(\varphi_0,\psi_0)\|_1^2)\\
&\qquad+C_5(m_1,M_1)\varepsilon^{\frac{1}{2}}\left\{1+\int_0^t(1+\tau)^{-\frac{7}{6}}\left\|\left(\sqrt{\Phi(v,\overline{V})},\psi,\frac{\varphi_x}{v}\right)(\tau)\right\|^2d\tau\right\}.
\end{aligned}
\tag{5.16}
$$

Furthermore, similar to the proof of (4.18), we have from (5.3) that there exists a constant $C_6(m_1,M_1)>0$ such that

$$
\left\|\left(\frac{\tilde{v}_x}{\tilde{v}}\right)(t)\right\|^2\le 2\left\|\left(\frac{\varphi_x}{v}\right)(t)\right\|^2+C(\underline{V},\overline{V})C_6(m_1,M_1)\varepsilon^{\frac{1}{2}}.
\tag{5.17}
$$

Here $\tilde{v}=\frac{v}{V}$.

From (5.16) and (5.17), if we choose ε sufficiently small such that

$$
0<\varepsilon<\min\{1,(C_5(m_1,M_1))^{-2},(C_6(m_1,M_1))^{-2}\},
\tag{5.18}
$$

then we can get from (5.16)–(5.18) and the Gronwall inequality that

$$\begin{aligned}&\left\|\left(\sqrt{\Phi(v,\overline{V})},\psi,\frac{\varphi_x}{v},\frac{\tilde{v}_x}{\tilde{v}}\right)(t)\right\|^2+\int_0^t\left\|\left(\frac{\psi_x}{\sqrt{v}},\sqrt{\frac{-p'(\overline{V}+\varphi)}{v}}\varphi_x\right)(\tau)\right\|^2 d\tau\\ &\qquad\le C(\underline{V},\overline{V})(1+\|(\varphi_0,\psi_0)\|_1^2).\end{aligned}\tag{5.19}$$

Here note also that the right-hand side of (5.19) is independent of m_1 and M_1.

Now we try to use (5.19) and the method of Kanel' to deduce a lower and an upper bound for $v(t,x)$. For this purpose, we first give the following two assertions.

Assertion A. There exists a positive constant $C_7>0$ such that

$$\lim_{v\to 0_+}\frac{\Phi(v,\overline{V})(1+\tilde{v})}{(1-\tilde{v})^2}\ge C_7.\tag{5.20}$$

Assertion B. There exists a positive constant $C_8>0$ such that

$$\lim_{v\to+\infty}\frac{\Phi(v,\overline{V})(1+\tilde{v})}{(1-\tilde{v})^2}\ge C_8.\tag{5.21}$$

To go directly to the proof of (5.4), we postpone the proof of the above two assertions to the end of this section.

From (5.20) and (5.21), we can conclude that there exists a suitably large fixed positive constant M such that

$$\begin{cases}\Phi(v,\overline{V})\ge\dfrac{C_7}{2}\dfrac{(1-\tilde{v})^2}{1+\tilde{v}} & \text{for } 0<v\le\dfrac{1}{M},\\ \Phi(v,\overline{V})\ge\dfrac{C_8}{2}\dfrac{(1-\tilde{v})^2}{1+\tilde{v}} & \text{for } M\le v<\infty.\end{cases}\tag{5.22}$$

Thus if we set $C_9:=\min_{[\frac{1}{M},M]}\{\frac{\Phi(v,\overline{V})(1+\tilde{v})}{(1-\tilde{v})^2}\}>0$ and let $C_{10}:=\min\{\frac{C_7}{2},\frac{C_8}{2},C_9\}>0$, we get

$$\Phi(v,\overline{V})\ge C_{10}\frac{(1-\tilde{v})^2}{1+\tilde{v}}:=\overline{\Phi}(\tilde{v}).\tag{5.23}$$

Here

$$\overline{\Phi}(\tilde{v})=\frac{(1-\tilde{v})^2}{1+\tilde{v}}\sim\begin{cases}1 & \text{as } \tilde{v}\to 0_+,\\ \tilde{v} & \text{as } \tilde{v}\to+\infty.\end{cases}\tag{5.24}$$

Combining (5.19) with (5.23), we deduce that

$$\int_{\mathbf{R}}\overline{\Phi}(\tilde{v}(t,x))dx\le C(\underline{V},\overline{V})\left\{1+\|(\varphi_0,\psi_0)\|_1^2\right\}.\tag{5.25}$$

Having obtained (5.25), similar to the proof of $(4.6)_2$, we set

$$\overline{\Psi}(\tilde{v})=\int_1^{\tilde{v}}\frac{\sqrt{\overline{\Phi}(\eta)}}{\eta}d\eta,$$

and we get from (5.24) that

$$\overline{\Psi}(\tilde{v}) \to \begin{cases} -\infty & \text{as } \tilde{v} \to 0_+, \\ +\infty & \text{as } \tilde{v} \to +\infty. \end{cases}$$

The above observation, together with the fact that

$$\left|\overline{\Psi}(\tilde{v}(t,x))\right| = \left|\int_{-\infty}^{x} \frac{\partial \overline{\Psi}(\tilde{v}(t,y))}{\partial y} dy\right| \le \left\|\sqrt{\overline{\Phi}(\tilde{v}(t))}\right\| \left\|\begin{pmatrix} \tilde{v}_x \\ \tilde{v} \end{pmatrix}(t)\right\| < \infty,$$

proves (5.4). Having obtained (5.4), the proof of Theorem 1.3 is completely similar to that of Theorem 1.2, and the details are omitted.

Now we turn to proving Assertion A and Assertion B. First, we prove Assertion A. For this purpose, we need only to consider the following two cases.

Case A_1. $p(v)$ is integrable at $v = 0$, i.e., $\int_0^{\overline{V}} p(s)ds < \infty$. In such a case, we have

$$\lim_{v \to 0_+} \frac{\Phi(v, \overline{V})(1+\tilde{v})}{(1-\tilde{v})^2} = -\overline{V} p\left(\overline{V}\right) + \int_0^{\overline{V}} p(s)ds. \tag{5.26}$$

Since $0 < v_- \le \overline{V}(t,x) \le v_+$ (here we have used the fact that for the 1-rarefaction wave, $\overline{V}_x(t,x) > 0$, but such an assumption is not essential) and $\frac{d}{dv}(-vp(v) + \int_0^v p(s)ds) = -vp'(v) > 0$, we can easily deduce that (5.20) holds with $C_7 = -v_- p(v_-) + \int_0^{v_-} p(s)ds$.

Case A_2. $p(v)$ is not integrable at $v = 0$, i.e., $\int_0^{\overline{V}} p(s)ds = \infty$. In such a case, (5.20) holds trivially.

This proves Assertion A.

As to Assertion B, since $p'(v) < 0$, and $p(v) > 0$ for each $v > 0$, we can conclude that there exists a constant $p_\infty \ge 0$ such that

$$\lim_{v \to +\infty} p(v) = p_\infty, \quad p_\infty < p(v_+) \le p(\overline{V}(t,x)), \tag{5.27}$$

for all $(t,x) \in [0,T] \times \mathbf{R}$.

Equation (5.27) implies that there exists a suitably chosen fixed constant $N > v_+ > 0$ such that for $v \ge N > v_+ > 0$,

$$\int_N^v p(s)ds < \int_N^v \left(p_\infty + \frac{p(v_+) - p_\infty}{2}\right) ds = \frac{p(v_+) + p_\infty}{2}(v - N).$$

Consequently, for $v \geq N > v_+ > 0$,

$$\begin{aligned}\frac{\Phi\left(v,\overline{V}\right)(1+\tilde{v})}{(1-\tilde{v})^2} &= \frac{1+\tilde{v}}{(1-\tilde{v})^2}\left(p\left(\overline{V}\right)\left(v-\overline{V}\right)-\int_{\overline{V}}^{v}p(s)ds\right)\\ &\geq \frac{1+\tilde{v}}{(1-\tilde{v})^2}\left(\left(p\left(\overline{V}\right)-\frac{p(v_+)+p_\infty}{2}\right)v-\overline{V}p\left(\overline{V}\right)\right.\\ &\qquad\left.+\frac{p(v_+)+p_\infty}{2}N-\int_{\overline{V}}^{N}p(s)ds\right)\\ &> \frac{1+\tilde{v}}{(1-\tilde{v})^2}\left(\frac{p(v_+)-p_\infty}{2}v-\overline{V}p\left(\overline{V}\right)\right.\\ &\qquad\left.+\frac{p(v_+)+p_\infty}{2}N-\int_{\overline{V}}^{N}p(s)ds\right).\end{aligned}\tag{5.28}$$

From (5.28) and (5.27), we can immediately deduce (5.21). This proves Assertion B.

6. The proof of Theorem 1.5. This section is devoted to proving Theorem 1.5. For this purpose, letting $(\varphi,\psi)=(v-\overline{\overline{V}},u-\overline{\overline{U}})$, we deduce that (φ,ψ) solves

$$\begin{cases}\varphi_t-\psi_x=0,\\ \psi_t+\left[p\left(\varphi+\overline{\overline{V}}\right)-p\left(\overline{\overline{V}}\right)\right]_x-\mu\left(\dfrac{u_x}{v}-\dfrac{\overline{\overline{U}}_x}{\overline{\overline{V}}}\right)_x=G_x,\end{cases}\tag{6.1}$$

with initial data

$$(\varphi,\psi)(t,x)|_{t=0}=(\varphi_0,\psi_0)(x)=\left(v_0(x)-\overline{\overline{V}}(0,x),u_0(x)-\overline{\overline{U}}(0,x)\right).\tag{6.2}$$

Here

$$G_x(t,x)=\mu\left(\frac{\overline{\overline{U}}}{\overline{\overline{V}}}\right)_x(t,x)-g\left(\overline{\overline{V}}\right)_x(t,x).\tag{6.3}$$

Multiplying $(6.1)_2$ by ψ, $(6.1)_1$ by $[p(\overline{\overline{V}})-p(v)]$, adding the results, and integrating it with respect to t and x over $[0,t)\times\mathbf{R}$ we deduce

$$\begin{aligned}&\frac{1}{2}\|\psi(t)\|^2+\int_{\mathbf{R}}\Phi\left(v,\overline{\overline{V}}\right)(t)dx\\ &\quad+\int_0^t\int_{\mathbf{R}}\left\{\mu\frac{\psi_x^2}{v}-\mu\frac{\overline{\overline{V}}_t\psi_x\varphi}{v\overline{\overline{V}}}+\left(p(v)-p\left(\overline{\overline{V}}\right)-p'\left(\overline{\overline{V}}\right)\varphi\right)\overline{\overline{V}}_t\right\}dxd\tau\\ &=\frac{1}{2}\|\psi_0\|^2+\int_{\mathbf{R}}\Phi\left(v_0,\overline{\overline{V}}(0,x)\right)dx+\int_0^t\int_{\mathbf{R}}G_x(\tau)\psi(\tau)dxd\tau.\end{aligned}\tag{6.4}$$

If we put $p(v)-p(\overline{\overline{V}})-p'(\overline{\overline{V}})\varphi=f(v,\overline{\overline{V}})\varphi^2$ and regard the three terms in the integrand in the third term of (6.4) as a quadratic equation of $\sqrt{\mu}\frac{\psi_x}{\sqrt{v}}$ and $\sqrt{f(v,\overline{\overline{V}})\overline{\overline{V}}_t}\varphi$,

then the discriminant is

$$D = \mu \frac{\overline{\overline{V}}_t}{\overline{\overline{V}}^2 v f\left(v, \overline{\overline{V}}\right)} - 4.$$

Noting that from (1.26) we can deduce that

$$\frac{1}{v f\left(v, \overline{\overline{V}}\right)} = \frac{\left(v - \overline{\overline{V}}\right)^2}{v\left(p(v) - p\left(\overline{\overline{V}}\right) - p'\left(\overline{\overline{V}}\right)\left(v - \overline{\overline{V}}\right)\right)}$$

is bounded as $v \to +\infty$ and the assumption (1.39) indicates that $\frac{1}{v f(v,\overline{\overline{V}})} \le C_1 \overline{\overline{V}}^2$ as $v \to 0_+$, we have that there exists a positive constant $C_{11} > 0$ such that

$$\frac{1}{v f\left(v, \overline{\overline{V}}\right)} \le C_{11} \overline{\overline{V}}^2. \tag{6.5}$$

On the other hand, we get from Lemma 2.3 that

$$\left|\overline{\overline{V}}_t(t,x)\right| \le C_{12}(q)\delta\varepsilon \tag{6.6}$$

for some positive constant $C_{12}(q) > 0$.

Thus if we choose $\varepsilon > 0$ as

$$\varepsilon = \frac{2}{\mu C_{11} C_{12}(q)\delta}, \quad \text{i.e., } \varepsilon = \varepsilon(\delta), \tag{6.7}$$

we have $D \le -2 < 0$, and, consequently, there exists a constant $C_{13} > 0$ such that

$$\begin{aligned} &\int_0^t \int_{\mathbf{R}} \left\{ \mu \frac{\psi_x^2}{v} - \mu \frac{\overline{\overline{V}}_t \psi_x \varphi}{v \overline{\overline{V}}} + \left(p(v) - p\left(\overline{\overline{V}}\right) - p'\left(\overline{\overline{V}}\right)\varphi\right) \overline{\overline{V}}_t \right\} dx d\tau \\ &\quad \ge C_{13} \int_0^t \int_{\mathbf{R}} \left\{ \frac{\psi_x^2}{v} + \left| \frac{\overline{\overline{V}}_t \psi_x \varphi}{v} \right| + \left(p(v) - p\left(\overline{\overline{V}}\right) - p'\left(\overline{\overline{V}}\right)\varphi\right) \overline{\overline{V}}_t \right\} dx d\tau. \end{aligned} \tag{6.8}$$

Furthermore, from Lemma 2.3 we deduce that

$$\begin{aligned} \int_0^t \int_{\mathbf{R}} G_x(\tau)\psi(\tau) dx d\tau &\le \int_0^t \|G_x(\tau)\| \|\psi(\tau)\| d\tau \\ &\le \int_0^t \|G_x(\tau)\| d\tau + \int_0^t \|G_x(\tau)\| \|\psi(\tau)\|^2 d\tau \\ &\le O(1)\left(1 + \int_0^t \|G_x(\tau)\| \|\psi(\tau)\|^2 d\tau\right). \end{aligned} \tag{6.9}$$

Inserting (6.8) and (6.9) into (6.4), we have from Lemma 2.3 and the Gronwall inequality that

$$\begin{aligned} &\frac{1}{2}\|\psi(t)\|^2 + \int_{\mathbf{R}} \Phi\left(v, \overline{\overline{V}}\right)(t) dx \\ &\quad + \int_0^t \int_{\mathbf{R}} \left\{ \frac{\psi_x^2}{v} + \left| \frac{\overline{\overline{V}}_t \psi_x \varphi}{v} \right| + \left(p(v) - p\left(\overline{\overline{V}}\right) - p'\left(\overline{\overline{V}}\right)\varphi\right) \overline{\overline{V}}_t \right\} dx d\tau \\ &\quad \le O(1)(1 + \|(\varphi_0, \psi_0)\|^2). \end{aligned} \tag{6.10}$$

Next we rewrite $(6.1)_2$ to the form of $\tilde{v} = \frac{v}{\overline{\overline{V}}}$:

$$\left(\mu\frac{\tilde{v}_x}{\tilde{v}} - \psi\right)_t + \overline{\overline{V}}|p'(v)|\tilde{v}_x = \left(vp'(v) - \overline{\overline{V}}p'\left(\overline{\overline{V}}\right)\right)\frac{\overline{\overline{V}}_x}{\overline{\overline{V}}} - G_x. \tag{6.11}$$

Multiplying (6.11) by $\frac{\tilde{v}_x}{\tilde{v}}$ and integrating the results with respect to t and x over $[0,t)\times\mathbf{R}$ gives

$$\begin{aligned}&\int_{\mathbf{R}}\left(\frac{\mu}{2}\left(\frac{\tilde{v}_x}{\tilde{v}}\right)^2 - \psi\frac{\tilde{v}_x}{\tilde{v}}\right)(\tau)dx\Big|_{\tau=0}^{\tau=t} + \int_0^t\int_{\mathbf{R}}\frac{\overline{\overline{V}}^2|p'(v)|\tilde{v}_x^2}{v}dxd\tau\\ &\quad= \int_0^t\int_{\mathbf{R}}\frac{\overline{\overline{V}}_x}{\overline{\overline{V}}}\left(vp'(v) - \overline{\overline{V}}p'\left(\overline{\overline{V}}\right)\right)\frac{\tilde{v}_x}{\tilde{v}}dxd\tau - \int_0^t\int_{\mathbf{R}}G_x\frac{\tilde{v}_x}{\tilde{v}}dxd\tau.\end{aligned} \tag{6.12}$$

We now estimate the two terms in the right-hand side of (6.12) term by term. First, similar to the proof of (6.9), we get that

$$-\int_0^t\int_{\mathbf{R}}G_x\frac{\tilde{v}_x}{\tilde{v}}dxd\tau \le O(1)\left(1 + \int_0^t\|G_x(\tau)\|\left\|\left(\frac{\tilde{v}_x}{\tilde{v}}\right)(\tau)\right\|^2 d\tau\right). \tag{6.13}$$

To deal with the other term, we decompose it into two parts:

$$\begin{aligned}&\int_0^t\int_{\mathbf{R}}\frac{\overline{\overline{V}}_x}{\overline{\overline{V}}}\left(vp'(v) - \overline{\overline{V}}p'\left(\overline{\overline{V}}\right)\right)\frac{\tilde{v}_x}{\tilde{v}}dxd\tau\\ &\quad= \left(\int\int_{\Omega_1} + \int\int_{\Omega_2}\right)\frac{\overline{\overline{V}}_x}{\overline{\overline{V}}}\left(vp'(v) - \overline{\overline{V}}p'\left(\overline{\overline{V}}\right)\right)\frac{\tilde{v}_x}{\tilde{v}}dxd\tau\\ &\quad= R_{11} + R_{12},\end{aligned} \tag{6.14}$$

where $\Omega_1 = \{(\tau,x) \mid 0 < v(\tau,x) \le 1, 0 \le \tau \le t\}$, $\Omega_2 = \{(\tau,x) \mid v(\tau,x) \ge 1, 0 \le \tau \le t\}$.

To bound R_{11} and R_{12}, we note first from the assumption (1.39) that

$$\begin{aligned}R_{11} &\le \frac{1}{4}\int_0^t\int_{\mathbf{R}}\frac{\overline{\overline{V}}^2|p'(v)|}{v}\tilde{v}_x^2dxd\tau + O(1)\int\int_{\Omega_1}\overline{\overline{V}}_x^2\frac{\left(vp'(v) - \overline{\overline{V}}p'\left(\overline{\overline{V}}\right)\right)^2}{vp'(v)}dxd\tau\\ &\le \frac{1}{4}\int_0^t\int_{\mathbf{R}}\frac{\overline{\overline{V}}^2|p'(v)|}{v}\tilde{v}_x^2dxd\tau + O(1)\int\int_{\Omega_1}\overline{\overline{V}}_x^2v|p'(v)|dxd\tau\\ &\le \frac{1}{4}\int_0^t\int_{\mathbf{R}}\frac{\overline{\overline{V}}^2|p'(v)|}{v}\tilde{v}_x^2dxd\tau + O(1)\int\int_{\Omega_1}\overline{\overline{V}}_x^2p(v)dxd\tau\\ &\le \frac{1}{4}\int_0^t\int_{\mathbf{R}}\frac{\overline{\overline{V}}^2|p'(v)|}{v}\tilde{v}_x^2dxd\tau\\ &\quad + O(1)\int_0^t\int_{\mathbf{R}}\overline{\overline{V}}_t\left(p(v) - p\left(\overline{\overline{V}}\right) - p'\left(\overline{\overline{V}}\right)\varphi\right)dxd\tau.\end{aligned} \tag{6.15}$$

As to R_{12}, we note first from (1.26) that for $v \ge 1$,

$$v|p'(v)| \le p(1) - p'(1), \tag{6.16}$$

which follows from

$$\begin{aligned} p(v) &= p(1) + \int_1^v p'(s)ds \le p(1) - \int_1^v (-p'(s))ds \\ &\le p(1) - (-p'(v))(v-1) = p(1) - p'(v) + p'(v)v \\ &\le p(1) - p'(1) + vp'(v). \end{aligned}$$

Hence for $\sigma \gg 1$, we have from (6.16) that

$$\begin{aligned} R_{12} &\le O(1)\int\int_{\Omega_2} \left|\overline{\overline{V}}_x\right| \frac{\tilde{v}_x}{\tilde{v}} dxd\tau \\ &\le \int\int_{\Omega_2} \left|\overline{\overline{V}}_x\right| \tilde{v} dxd\tau + O(1)\int\int_{\Omega_2} \left|\overline{\overline{V}}_x\right| \frac{\tilde{v}_x^2}{\tilde{v}^3} dxd\tau \\ &\le O(1)\int_0^t \int_{\mathbf{R}} \overline{\overline{V}}_t \left(p(v) - p\left(\overline{\overline{V}}\right) - p'\left(\overline{\overline{V}}\right)\varphi\right) dxd\tau \\ &\quad + O(1)\int\int_{\Omega_2} \left(\frac{|p'(v)|}{v}\tilde{v}_x^2\right)^{\frac{1}{\sigma}} \left|\overline{\overline{V}}_x\right| (v|p'(v)|)^{-\frac{1}{\sigma}} \frac{1}{v} \left(\frac{\tilde{v}_x}{\tilde{v}}\right)^{\frac{2(\sigma-1)}{\sigma}} dxd\tau \qquad (6.17) \\ &\le O(1)\int_0^t \int_{\mathbf{R}} \overline{\overline{V}}_t \left(p(v) - p\left(\overline{\overline{V}}\right) - p'\left(\overline{\overline{V}}\right)\varphi\right) dxd\tau \\ &\quad + \frac{1}{4}\int_0^t \int_{\mathbf{R}} \overline{\overline{V}}^2 \frac{|p'(v)|}{v} \tilde{v}_x^2 dxd\tau \\ &\quad + O(1)\int\int_{\Omega_2} \left|\overline{\overline{V}}_x\right|^{\frac{\sigma}{\sigma-1}} (v^{\sigma+1}|p'(v)|)^{-\frac{1}{\sigma-1}} \left(\frac{\tilde{v}_x}{\tilde{v}}\right)^2 dxd\tau. \end{aligned}$$

Now, taking $\sigma = C_1 - 1$, we have from (6.17) and the assumption (1.39) that

$$\begin{aligned} R_{12} &\le O(1)\int_0^t \int_{\mathbf{R}} \overline{\overline{V}}_t \left(p(v) - p\left(\overline{\overline{V}}\right) - p'\left(\overline{\overline{V}}\right)\varphi\right) dxd\tau \\ &\quad + \frac{1}{4}\int_0^t \int_{\mathbf{R}} \overline{\overline{V}}^2 \frac{|p'(v)|}{v} \tilde{v}_x^2 dxd\tau \qquad (6.18) \\ &\quad + O(1)\int_0^t (1+\tau)^{-\frac{C_1-1}{C_1-2}} \left\| \left(\frac{\tilde{v}_x}{\tilde{v}}\right)^2 (\tau) \right\|^2 d\tau. \end{aligned}$$

Substituting (6.15) and (6.18) into (6.14), we deduce that

$$\begin{aligned} &\left| \int_0^t \int_{\mathbf{R}} \frac{\overline{\overline{V}}_x}{\overline{\overline{V}}} \left(vp'(v) - \overline{\overline{V}} p'\left(\overline{\overline{V}}\right)\right) \frac{\tilde{v}_x}{\tilde{v}} dxd\tau \right| \\ &\quad \le O(1)\int_0^t \int_{\mathbf{R}} \overline{\overline{V}}_t \left(p(v) - p\left(\overline{\overline{V}}\right) - p'\left(\overline{\overline{V}}\right)\varphi\right) dxd\tau \\ &\quad\quad + \frac{1}{2}\int_0^t \int_{\mathbf{R}} \overline{\overline{V}}^2 \frac{|p'(v)|}{v} \tilde{v}_x^2 dxd\tau \qquad (6.19) \\ &\quad\quad + O(1)\int_0^t (1+\tau)^{-\frac{C_1-1}{C_1-2}} \left\| \left(\frac{\tilde{v}_x}{\tilde{v}}\right)^2 (\tau) \right\|^2 d\tau. \end{aligned}$$

Combining (6.19) and (6.13) with (6.12), we get that

$$\begin{aligned}
&\int_{\mathbf{R}} \left(\frac{\mu}{2} \left(\frac{\tilde{v}_x}{\tilde{v}} \right)^2 - \psi \frac{\tilde{v}_x}{\tilde{v}} \right)(t)dx + \int_0^t \int_{\mathbf{R}} \frac{\overline{\overline{V}}^2 |p'(v)| \tilde{v}_x^2}{v} dxd\tau \\
&\quad \le O(1) \left(1 + \|(\varphi_0, \psi_0)\|_1^2 \right) \\
&\quad\quad + O(1) \int_0^t \left(\|G_x(\tau)\| + (1+\tau)^{-\frac{C_1-1}{C_1-2}} \right) \left\| \left(\frac{\tilde{v}_x}{\tilde{v}} \right)^2 (\tau) \right\|^2 d\tau \\
&\quad\quad + O(1) \int_0^t \int_{\mathbf{R}} \overline{\overline{V}}_t \left(p(v) - p\left(\overline{\overline{V}} \right) - p'\left(\overline{\overline{V}} \right) \varphi \right) dxd\tau.
\end{aligned} \tag{6.20}$$

Equations (6.10) and (6.20), together with the Gronwall inequality, yield

$$\begin{aligned}
&\|\psi(t)\|^2 + \int_{\mathbf{R}} \Phi\left(v, \overline{\overline{V}} \right)(t)dx + \left\| \left(\frac{\tilde{v}_x}{\tilde{v}} \right)(t) \right\|^2 \\
&\quad + \int_0^t \int_{\mathbf{R}} \left\{ \frac{\psi_x^2}{v} + \left| \frac{\overline{\overline{V}}_t \psi_x \varphi}{v} \right| + \left(p(v) - p\left(\overline{\overline{V}} \right) - p'\left(\overline{\overline{V}} \right) \varphi \right) \overline{\overline{V}}_t \right\} dxd\tau \\
&\quad \le O(1)(1 + \|(\varphi_0, \psi_0)\|_1^2).
\end{aligned} \tag{6.21}$$

Having obtained (6.21), by repeating the argument used in the proof of Theorem 1.3, we can deduce that there exists a positive constant C_{14} such that

$$C_{14}^{-1} \le v(t,x) \le C_{14}, \quad t \ge 0,\ x \in \mathbf{R}. \tag{6.22}$$

With (6.22) in hand, by employing the method used in [21], we can easily deduce that the results stated in Theorem 1.5 are true. This completes the proof of Theorem 1.5.

REFERENCES

[1] R. Courant and K. O. Friedrichs, *Supersonic Flows and Shock Waves*, Wiley-Interscience, New York, 1948.

[2] Y. Hattori and K. Nishihara, *A note on the stability of the rarefaction wave of the Burgers equation*, Japan J. Indust. Appl. Math., 8 (1991), pp. 85–96.

[3] D. Hoff, *Global solutions of the equations of one-dimensional, compressible flow with large data and forces, and with differing end states*, Z. Angew. Math. Phys., 49 (1998), pp. 774–785.

[4] H. Hokari and A. Matsumura, *Asymptotic toward one-dimensional rarefaction wave for the solution of two-dimensional compressible Euler equations with an artificial viscosity*, Asymptot. Anal., 15 (1997), pp. 283–298.

[5] F.-M. Huang, A. Matsumura, and X.-D. Shi, *On the stability of contact discontinuity for compressible Navier–Stokes equations with free boundary*, Osaka J. Math., to appear.

[6] A. M. Il'in and O. A. Oleinik, *Behavior of the solution of the Cauchy problem for certain quasilinear equations for unbounded increase of time*, Amer. Math. Soc. Transl. Ser. 2, 42 (1964), pp. 19–23.

[7] K. Ito, *Asymptotic decay toward the planar rarefaction waves of solutions for viscous conservation laws in several space dimensions*, Math. Models Methods Appl. Sci., 6 (1996), pp. 315–338.

[8] S. Jiang and P. Zhang, *Global weak solutions to the Navier–Stokes equations for a 1D viscous polytropic ideal gas*, Quart. Appl. Math., 61 (2003), pp. 435–449.

[9] Y. Kanel', *On a model system of equations of one-dimensional gas motion*, (in Russian), Differencial'nya Uravnenija, 4 (1968), pp. 374–380.

[10] S. Kawashima, *Systems of a Hyperbolic-Parabolic Composite, with Applications to the Equations of Magnetohydrodynamics*, Ph.D. Thesis, Kyoto University, Kyoto, Japan, 1985.
[11] S. Kawashima and A. Matsumura, *Asymptotic stability of travelling wave solutions of systems for one-dimensional gas motion*, Comm. Math. Phys., 101 (1985), pp. 97–127.
[12] S. Kawashima, A. Matsumura, and K. Nishihara, *Asymptotic behaviour of solutions for the equations of a viscous heat-conductive gas*, Proc. Japan Acad. Ser. A, 62 (1986), pp. 249–252.
[13] O. A. Ladyzenskaja, V. A. Solonikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.
[14] T.-P. Liu, *Shock waves for compressible Navier–Stokes equations are stable*, Comm. Pure Appl. Math., 39 (1986), pp. 565–594.
[15] T.-P. Liu and Z.-P. Xin, *Nonlinear stability of rarefaction waves for compressible Navier–Stokes equations*, Comm. Math. Phys., 118 (1988), pp. 451–465.
[16] T.-P. Liu and Z.-P. Xin, *Pointwise decay to contact discontinuities for systems of viscous conservation laws*, Asian J. Math., 1 (1997), pp. 34–84.
[17] T.-P. Liu and Y.-N. Zeng, *Large time behavior of solutions for general quasilinear hyperbolic-parabolic systems of conservation laws*, Mem. Amer. Math. Soc., 125 (1997), pp. 1–120.
[18] A. Matsumura and T. Nishida, *The initial value problem for the equations of motion of viscous and heat-conductive gases*, J. Math. Kyoto Univ., 26 (1980), pp. 67–104.
[19] A. Matsumura and K. Nishihara, *On the stability of travelling wave solutions of a one-dimensional model system for compressible viscous gas*, Japan J. Appl. Math., 2 (1985), pp. 17–25.
[20] A. Matsumura and K. Nishihara, *Asymptotic toward the rarefaction waves of the solutions of a one-dimensional model system for compressible viscous gas*, Japan J. Appl. Math., 3 (1986), pp. 1–13.
[21] A. Matsumura and K. Nishihara, *Global stability of the rarefaction waves of a one-dimensional model system for compressible viscous gas*, Comm. Math. Phys., 144 (1992), pp. 325–335.
[22] A. Matsumura and K. Nishihara, *Global asymptotics toward the rarefaction wave for solutions of viscous p-system with boundary effect*, Quart. Appl. Math., 58 (2000), pp. 69–83.
[23] G. Whitham, *Linear and Nonlinear Waves*, Wiley-Interscience, New York, 1974.
[24] Z.-P. Xin, *Asymptotic stability of rarefaction waves for* 2×2 *viscous hyperbolic conservation laws*, J. Differential Equations, 73 (1988), pp. 45–77.
[25] Z.-P. Xin, *Zero dissipation limit to rarefaction waves for the one-dimensional Navier–Stokes equations of compressible isentropic gases*, Comm. Pure Appl. Math., 46 (1993), pp. 621–665.

SIAM J. MATH. ANAL.
Vol. 35, No. 6, pp. 1598–1616

ASYMPTOTIC APPROXIMATION OF THE SOLUTION OF THE LAPLACE EQUATION IN A DOMAIN WITH HIGHLY OSCILLATING BOUNDARY*

Y. AMIRAT†, O. BODART†, U. DE MAIO‡, AND A. GAUDIELLO§

Abstract. We study the asymptotic behavior of the solution of the Laplace equation in a domain, a part of whose boundary is highly oscillating. The motivation comes from the study of a longitudinal flow in an infinite horizontal domain bounded at the bottom by a wall and at the top by a rugose wall. The latter is a plane covered with periodic asperities whose size depends on a small parameter, $\varepsilon > 0$. The assumption of sharp asperities is made; that is, the height of the asperities is fixed. Using a boundary layer corrector, we derive and analyze a nonoscillating approximation of the solution at order $\mathcal{O}(\varepsilon^{3/2})$ for the H^1-norm.

Key words. asymptotic expansions, oscillating boundary

AMS subject classifications. 35J25, 35B40

DOI. 10.1137/S0036141003414877

1. Introduction.

Boundary-value problems involving oscillating boundaries or interfaces frequently arise when modelling problems in industrial applications, such as flows over rough walls, electromagnetic waves in a region containing a rough interface, and elastic bodies containing a rough interface. The mathematical analysis of these problems consists of studying the large scale behavior of the solution. The goal is to determine effective boundary conditions or to construct accurate and numerically implementable asymptotic approximations. The main difficulty comes from the presence of boundary layers near the rough region, whose effects on correctors or error estimates have to be taken into account.

In the present paper we consider a boundary-value problem for the Laplace equation, arising from the study of a laminar flow over a rough wall. The problem arises also in the study of the heat transmission in winglets.

Let us consider a viscous fluid in an infinite horizontal domain limited at the bottom by a wall $\mathcal{P}$ and at the top by a rough wall $\mathcal{R}_\varepsilon$. We assume that $\mathcal{P}$ moves at a constant horizontal velocity $\gamma = (0, g, 0)$, $g \in \mathbf{R}$, and that $\mathcal{R}_\varepsilon$ is at rest. The latter is assumed to consist of a plane wall covered with periodic asperities whose size depends on a small parameter, $\varepsilon > 0$, and with a fixed height. Let $0 < a_1 < b_1 < l_1$, and let η_ε be the εl_1-periodic function defined on $(0, \varepsilon l_1)$ by

$$\eta_\varepsilon(x_1) = \begin{cases} l_3' & \text{if } x_1 \in (\varepsilon a_1, \varepsilon b_1), \\ l_3 & \text{if } x_1 \in (0, \varepsilon l_1) \backslash (\varepsilon a_1, \varepsilon b_1) \end{cases}$$

*Received by the editors February 11, 2003; accepted for publication (in revised form) August 1, 2003; published electronically April 7, 2004. This work is partially supported by the CNR Project "Agenzia 2000" and is part of the European Research Training Network "Homogenization and Multiple Scales" (HMS 2000), contract HPRN-2000-00109.

http://www.siam.org/journals/sima/35-6/41487.html

†Laboratoire de Mathématiques Appliquées, CNRS (UMR 6620), Université Blaise Pascal (Clermont-Ferrand 2), 63177 Aubière cedex, France (amirat@math.univ-bpclermont.fr, obodart@math.univ-bpclermont.fr).

‡Dipartimento di Matematica e Applicazioni "R. Caccioppoli," Università di Napoli "Federico II," Complesso Monte S. Angelo, via Cintia, 80126 Napoli, Italy (udemaio@unina.it).

§Dipartimento di Automazione, Elettromagnetismo, Ingegneria dell'Informazione e Matematica Industriale, Università di Cassino, via G. Di Biasio 43, 03043 Cassino (FR), Italy (gaudiell@unina.it).

with $l_3' > l_3 > 0$. The domain of the flow is

$$\mathcal{O}_\varepsilon = \left\{(x', x_3) \in \mathbf{R}^3 :\ x' \in \mathbf{R}^2,\ b(x_1) < x_3 < \eta_\varepsilon(x_1)\right\},$$

where $x' = (x_1, x_2)$ and b is a smooth and l_1-periodic function on $\mathbf{R}$ such that $b(x_1) < l_3$ in $\mathbf{R}$. The domain $\mathcal{O}_\varepsilon$ is bounded at the bottom by

$$\mathcal{P} = \left\{(x', x_3) \in \mathbf{R}^3 :\ x' \in \mathbf{R}^2,\ x_3 = b(x_1)\right\}$$

and at the top by $\mathcal{R}_\varepsilon = \partial\mathcal{O}_\varepsilon \backslash \mathcal{P}$, where $\partial\mathcal{O}_\varepsilon$ denotes the boundary of $\mathcal{O}_\varepsilon$. The profile of the asperities is then comb shaped. Throughout the paper, we will assume that $1/\varepsilon \in \mathbf{N}$ so that η_ε is also periodic with period l_1. Thus, $\mathcal{O}_\varepsilon$ can be viewed as generated by periodic translations of the bounded domain

$$\left\{x \in \mathbf{R}^3 :\ 0 < x_1 < l_1,\ b(x_1) < x_3 < \eta_\varepsilon(x_1)\right\}.$$

We consider the longitudinal flow described by the velocity field v_ε in the form $v_\varepsilon = (0, u_\varepsilon, 0)$ (and the pressure $p_\varepsilon = 0$), where u_ε satisfies the Laplace equation

$$\begin{cases} \Delta u_\varepsilon = 0 \ \text{ in } \Omega_\varepsilon, \\ u_\varepsilon = 0 \ \text{ on } R_\varepsilon, \\ u_\varepsilon = g \ \text{ on } P, \\ u_\varepsilon \ l_1\text{-periodic with respect to } x_1, \ \text{ for a.e. } x_3 \in (b(0), l_3), \end{cases} \tag{1.1}$$

where Ω_ε is the bidimensional section (see Figure 2.1)

$$\Omega_\varepsilon = \left\{x = (x_1, x_3) \in \mathbf{R}^2 : 0 < x_1 < l_1,\ b(x_1) < x_3 < \eta_\varepsilon(x_1)\right\},$$

and

$$\begin{aligned} P &= \left\{x = (x_1, x_3) \in \mathbf{R}^2 : 0 < x_1 < l_1,\ x_3 = b(x_1)\right\}, \\ L &= \left\{x = (x_1, x_3) \in \mathbf{R}^2 : x_1 = 0,\ b(0) < x_3 < \eta(0)\right\}, \\ &\quad \cup \left\{x = (x_1, x_3) \in \mathbf{R}^2 : x_1 = l_1,\ b(l_1) < x_3 < \eta(l_1)\right\}, \end{aligned}$$

and $R_\varepsilon = \partial\Omega_\varepsilon \backslash (P \cup L)$, $\partial\Omega_\varepsilon$ being the boundary of Ω_ε.

The aim is to study the asymptotic behavior, as ε goes to 0, of the solution u_ε of (1.1).

Notice that problem (1.1) is also linked with the study of the heat transmission in winglets. In this case, u_ε represents the temperature in the medium, and the problem is to determine the asymptotic behavior of the temperature in the case where the number of winglets to cool down a hot part are increasing and they are getting increasingly near to each other.

The main difficulty arising in our study is due to the fact that the amplitude of the oscillations of the boundary is large. The case where $b = 0$ is studied in [4]. The assumption $b = 0$ allows us to consider solutions u_ε of problem (1.1) that are εl_1-periodic and then to construct an approximation of u_ε, up to an exponentially small error, by a nonoscillating explicit function. Here we consider the situation where the function b is not constant; hence the corresponding solution u_ε is not εl_1-periodic with respect to x_1.

Problems involving rough boundaries, in the case where the frequency and the amplitude of the oscillations of the boundary are of the same order ε, have been addressed

by many authors. In [1], an approximation at order $\mathcal{O}(\varepsilon^{\frac{3}{2}})$ for the H^1-norm is derived and analyzed for the Laplace equation, using a domain decomposition argument. In [2], Achdou, Pironneau, and Valentin consider a laminar flow over a rough wall with periodic roughness elements. Using asymptotic expansions and corresponding boundary layer correctors, the authors derive first and second order effective boundary conditions. In [3], Allaire and Amar give a nonoscillating approximation at order $\mathcal{O}(\varepsilon^{\frac{3}{2}})$ for the H^1-norm for the Laplace equation. In [22], Jäger and Mikelić consider the Laplace equation on a bounded domain consisting of a porous medium, a nonperforated domain, and an interface between them. Using boundary layers describing the interaction between the two media, the authors derive asymptotic approximations and establish L^2 estimates. In [5], for a flow governed by the Navier–Stokes equations in a domain $\mathcal{O}_\varepsilon$ corresponding to the case where $b = 0$ and the frequency and the amplitude of the oscillations of the boundary are of the same order ε, it is proved that, outside a neighborhood of the rugose zone, the flow behaves asymptotically as a Couette flow, up to an exponentially small error. The Laplace equation in a domain with very rapidly oscillating locally periodic boundary, the amplitude of the oscillations being ε and the frequency ε^α ($\alpha > 1$), is considered by Chechkin, Friedman, and Piatniski in [12]. In this paper, the authors analyze a first order approximation in the H^1-norm. Asymptotic limits of boundary-value problems in oscillating domains, in the case where the amplitude of the oscillations does not vanish as $\varepsilon \to 0$, are studied in [9], [11], [15, 16, 17, 18], and [27, 28, 29, 31]. Problems in domains with fragmented boundaries are treated in [21] and [26]. For general references about homogenization, we refer the reader to [6, 7, 8, 10, 13, 14, 23, 32, 33].

Let

$$D_\varepsilon = \left\{x = (x_1, x_3) \in \mathbf{R}^2 : x_1 \in \mathbf{R},\ b(x_1) < x_3 < \eta_\varepsilon(x_1)\right\},$$

and let, for $m \geq 0$, the space

$$H^m_{\mathrm{per}}(\Omega_\varepsilon) = \{u \in H^m_{\mathrm{loc}}(D_\varepsilon)\,;\, u \in H^m(\Omega_\varepsilon),\ u(x_1+l_1, x_3) = u(x_1, x_3),\ x_3 \in (b(0), l_3)\},$$

endowed with the norm of $H^m(\Omega_\varepsilon)$. In the present paper, we consider the slight generalization of problem (1.1),

$$\left\{\begin{array}{ll} -\Delta u_\varepsilon = f & \text{in } \Omega_\varepsilon, \\ u_\varepsilon = 0 & \text{on } R_\varepsilon, \\ u_\varepsilon = g & \text{on } P, \\ u_\varepsilon \in H^1_{\mathrm{per}}(\Omega_\varepsilon), \end{array}\right. \tag{1.2}$$

where f is a smooth function and g is a given constant. The paper is organized as follows. In section 2, we establish a convergence result for the sequence $\{u_\varepsilon\}_\varepsilon$: denoting $\Omega = \{(x_1, x_3)\,;\, 0 < x_1 < l_1,\, b(x_1) < x_3 < l_3'\}$, and $\widetilde{u_\varepsilon}$ being the zero extension of u_ε to Ω, we prove (Proposition 2.1) the convergence of $\{\widetilde{u_\varepsilon}\}_\varepsilon$ in $H^1(\Omega)$. Section 3 is devoted to decay estimates at infinity for the solution of the Laplace equation in an infinite vertical domain of $\mathbf{R}^2$ (Proposition 3.1). These estimates play a key role in the subsequent analysis. Section 4 contains the main result of the paper (Theorem 4.1). Using boundary layer correctors, we construct a nonoscillating approximation of u_ε in Ω_ε and an outer boundary layer of height 2ε. We show that this approximation is of order $\mathcal{O}(\varepsilon^{3/2})$ in the H^1-norm. This generalizes the result in [1] (see also [3]) which deals with the case where the frequency and the amplitude of the oscillations of the boundary are of same order ε.

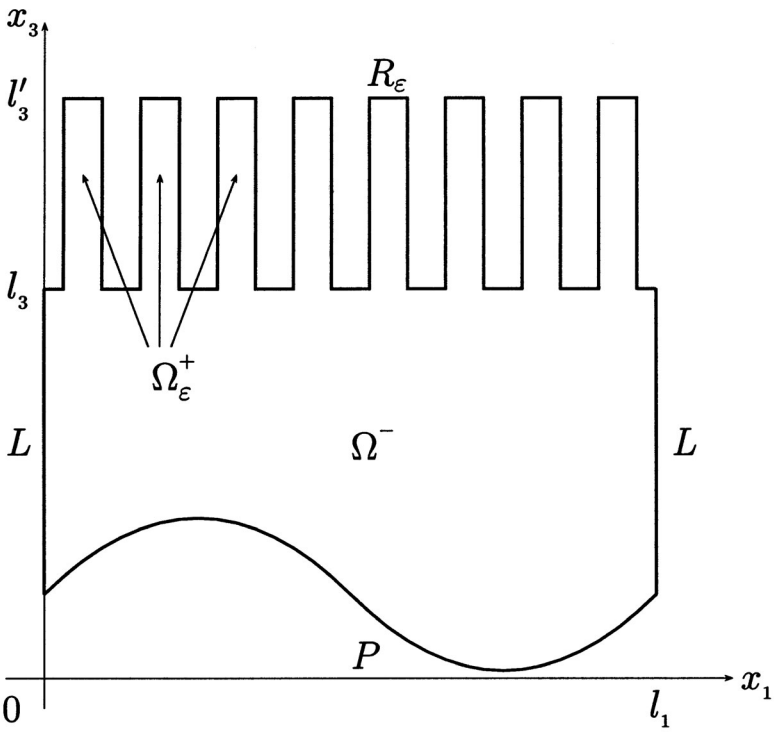


FIG. 2.1. *Domain* Ω_ε.

2. A convergence result. Let u_ε be the solution of (1.2). We denote

$$M = \max b, \tag{2.1}$$

and

$$\begin{cases} \Omega_\varepsilon^+ = \{(x_1,x_3) \in \Omega_\varepsilon : \ l_3 < x_3 < l_3'\}, \\ \Omega^+ = (0,l_1) \times (l_3, l_3'), \\ \Omega^- = \left\{(x_1,x_3) \in \mathbf{R}^2 : x_1 \in (0,l_1), \ b(x_1) < x_3 < l_3\right\}, \\ \Sigma = (0,l_1) \times \{l_3\}, \\ \Omega = \Omega^- \cup \Omega^+ \cup \Sigma; \end{cases}$$

see Figure 2.1 for a summary of the main notations. In what follows we will use the spaces $H^m_{\rm per}(\Omega)$ and $H^m_{\rm per}(\Omega^-)$ (for $m \geq 0$); the definition is similar to that of $H^m_{\rm per}(\Omega_\varepsilon)$ given in the previous section. The only regularity assumptions we make here are that b is Lipschitz-continuous and $f \in L^2(\Omega)$. Remark that

$$\chi_{\Omega_\varepsilon^+} \rightharpoonup \frac{b_1 - a_1}{l_1} \quad \text{weakly-}\star \text{ in } L^\infty\left(\Omega^+\right), \tag{2.2}$$

where $\chi_{\Omega_\varepsilon^+}$ denotes the characteristic function of Ω_ε^+. To describe the limit problem, as $\varepsilon \to 0$, of problem (1.2), we introduce the function

$$u = \begin{cases} 0 & \text{in } \Omega^+, \\ u^- & \text{in } \Omega^-, \end{cases} \tag{2.3}$$

where u^- is the unique solution of the following problem:

$$\begin{cases} -\Delta u^- = f & \text{in } \Omega^-, \\ u^- = 0 & \text{on } \Sigma, \\ u^- = g & \text{on } P, \\ u^- \in H^1_{\text{per}}(\Omega^-). \end{cases} \tag{2.4}$$

Let $\widetilde{u_\varepsilon}$ be the zero extension to Ω of u_ε. The following convergence result holds.

PROPOSITION 2.1. *Let u_ε be the solution of problem* (1.2)*, and let u be the function defined in* (2.3), (2.4)*. Then*

$$\widetilde{u_\varepsilon} \to u \quad \textit{strongly in } H^1(\Omega). \tag{2.5}$$

This convergence result was previously obtained in a more general framework by many authors who studied optimum design problems; see, e.g., [30], [31]. We give here a direct proof of the proposition.

Proof of Proposition 2.1. Let $s \in C^2(\mathbf{R})$ be such that

$$s(t) = \begin{cases} 0 & \text{if } t > \dfrac{M + l_3}{2}, \\ \\ 1 & \text{if } t < \dfrac{3M + l_3}{4}, \end{cases}$$

where M is given in (2.1), and let $h(x_1, x_3) = gs(x_3)$. Choosing $u_\varepsilon - h$ as a test function in problem (1.2), we get

$$\int_{\Omega_\epsilon} \nabla u_\varepsilon \left(\nabla u_\varepsilon - \nabla h\right) dx = \int_{\Omega_\epsilon} f\left(u_\varepsilon - h\right) dx. \tag{2.6}$$

Applying the Young inequality and the Poincaré inequality, we have

$$\begin{aligned} \int_{\Omega_\epsilon} |\nabla u_\varepsilon|^2 \, dx \le \alpha \int_{\Omega_\epsilon} |\nabla u_\varepsilon|^2 \, dx + \frac{1}{\alpha} \int_{\Omega^-} |\nabla h|^2 \, dx \\ + \alpha c \int_{\Omega_\epsilon} |\nabla u_\varepsilon|^2 \, dx + \frac{1}{\alpha} \int_{\Omega} f^2 dx + \int_{\Omega^-} fh \, dx, \end{aligned}$$

for any $\alpha > 0$, where c is a constant independent of ε and α, and, consequently,

$$\left(\int_{\Omega_\epsilon} |\nabla u_\varepsilon|^2 \right)^{\frac{1}{2}} dx \le c,$$

where c is a constant independent of ε. Then, by virtue of the Poincaré inequality, the sequence $\{\widetilde{u_\varepsilon}\}_\varepsilon$ is bounded in $H^1(\Omega)$. Therefore, up to a subsequence, not relabeled for convenience, there exists a function $u \in H^1_{\text{per}}(\Omega)$ (possibly depending on the subsequence) such that $u = g$ on P and

$$\begin{cases} \widetilde{u_\varepsilon} \rightharpoonup u & \text{weakly in } H^1(\Omega), \\ \widetilde{u_\varepsilon} \to u & \text{strongly in } L^2(\Omega). \end{cases} \tag{2.7}$$

Moreover, since

$$\widetilde{u_\varepsilon} = \widetilde{u_\varepsilon}\,\chi_{\Omega_\varepsilon^+} \quad \text{in } \Omega^+,$$

from (2.7) and (2.2), it follows that

$$u = 0 \quad \text{in } \Omega^+. \tag{2.8}$$

On the other hand, letting ε go to 0 in (1.2) with test functions $\varphi \in C^\infty(\Omega^-)$ such that φ is l_1-periodic with respect to x_1 for a.e. $x_3 \in (b(0), l_3)$, $\varphi_{|P} = 0$, and $\varphi_{|\Sigma} = 0$, it follows from (2.7) that $u_{|\Omega^-}$ solves problem (2.4). Since this problem admits a unique solution, convergences (2.7) hold for the whole sequence. To obtain the strong convergence (2.5), by virtue of (2.7) and (2.8), it is enough to prove that

$$\lim_{\varepsilon\to 0} \|\nabla\widetilde{u_\varepsilon}\|_{(L^2(\Omega))^2} = \|\nabla u\|_{(L^2(\Omega^-))^2}. \tag{2.9}$$

Relations (2.6), (2.7), and (2.8) provide that

$$\begin{aligned} \lim_{\varepsilon\to 0}\int_\Omega |\nabla\widetilde{u_\varepsilon}|^2\,dx &= \lim_{\varepsilon\to 0}\left(\int_\Omega \nabla\widetilde{u_\varepsilon}\nabla h\,dx + \int_\Omega f\,(\widetilde{u_\varepsilon} - h)\,dx\right) \\ &= \int_{\Omega^-} \nabla u\nabla h\,dx + \int_{\Omega^-} f\,(u-h)\,dx. \end{aligned} \tag{2.10}$$

Finally, choosing $u - h$ as a test function in problem (2.4), it follows that

$$\int_{\Omega^-} |\nabla u|^2\,dx = \int_{\Omega^-} \nabla u\nabla h\,dx + \int_{\Omega^-} f\,(u-h)\,dx. \tag{2.11}$$

The convergence (2.9) is then obtained by comparing (2.10) with (2.11). ∎

3. Decay estimates. The asymptotic approximation of u_ε will involve the solution of the Laplace equation in an infinite vertical domain of $\mathbf{R}^2$. Let $\Lambda^+ = (a_1, b_1)\times(0,+\infty)$, $\Lambda^- = (0,l_1)\times(-\infty,0)$ as displayed in Figure 3.1. Let $\psi^\pm$ be the functions defined by

$$\begin{cases} \psi^+ \in H^1(\Lambda^+), \\ \psi^- \in H^1_{\mathrm{loc,per}}(\Lambda^-), \quad \nabla\psi^- \in L^2(\Lambda^-), \end{cases} \tag{3.1}$$

$$\begin{cases} \Delta\psi^\pm = 0 & \text{in } \Lambda^\pm, \\ \psi^+ = 0 & \text{on } \partial\Lambda^+\backslash\Gamma, \\ \psi^- = 0 & \text{on } ((0,a_1)\cup(b_1,l_1))\times\{0\}, \\ \psi^+ = \psi^- & \text{on } \Gamma, \\ \dfrac{\partial\psi^+}{\partial y_3} = \dfrac{\partial\psi^-}{\partial y_3} + 1 & \text{on } \Gamma, \end{cases} \tag{3.2}$$

where $\Gamma = (a_1,b_1)\times\{0\}$. Here $\psi^- \in H^1_{\mathrm{loc,per}}(\Lambda^-)$ means $\psi^- \in H^1_{\mathrm{per}}(\Lambda')$ for any bounded domain $\Lambda' \subset \Lambda^-$. We denote by β the mean of ψ^- over a horizontal section

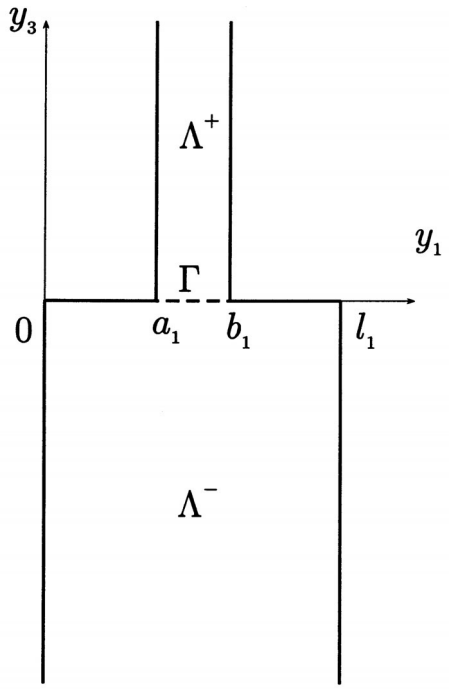


FIG. 3.1. *Infinite domain* Λ.

of Λ^-:

$$\beta = \frac{1}{l_1}\int_0^{l_1} \psi^-(y_1, -\delta)\, dy_1 \ \forall \delta \in (0, +\infty). \tag{3.3}$$

The following result is proved in [4].

PROPOSITION 3.1. *Problem* (3.1), (3.2) *admits a unique solution. Moreover,*

(i) *the constant* β *is independent of* δ*;*

(ii) *for any* $\alpha \in N^2$ *and for any* $\delta \in (0, +\infty)$*, there exist two positive constants* c *and* $c_{\alpha,\delta}$ *such that*

$$|\partial^\alpha \psi^+(y_1, y_3)| \leq c_{\alpha,\delta}\, e^{-cy_3} \ \forall (y_1, y_3) \in (a_1, b_1) \times (\delta, +\infty);$$

(iii) *for any* $\alpha \in \mathbb{N}^2$ *and for any* $\delta \in (0, +\infty)$*, there exist two positive constants* c *and* $c_{\alpha,\delta}$ *such that*

$$|\partial^\alpha (\psi^- - \beta)(y_1, y_3)| \leq c_{\alpha,\delta}\, e^{cy_3} \ \forall (y_1, y_3) \in (0, l_1) \times (-\infty, -\delta).$$

The above estimates are of the so-called de Saint-Venant type. The first is proved by means of Tartar's lemma (see [25, pp. 49–58]); see also [24]. The second is proved by adapting the proof of the first one. Let us remark that (3.1) and Proposition 3.1 provide that $(\psi^- - \beta) \in H^1_{\text{per}}(\Lambda^-)$.

Proposition 3.1 implies the following result.

COROLLARY 3.2. *Let $\psi^{\pm}$ be the functions satisfying* (3.1), (3.2). *Then there exist two positive constants c and C, independent of ε, such that*

$$\int_{\Omega_\varepsilon^+} \left|\psi^+\left(\frac{x_1}{\varepsilon}, \frac{x_3 - l_3}{\varepsilon}\right)\right|^2 dx \leq C\,\varepsilon,$$

$$\int_{\Omega^-} \left|\psi^-\left(\frac{x_1}{\varepsilon}, \frac{x_3 - l_3}{\varepsilon}\right) - \beta\right|^2 dx \leq C\,\varepsilon,$$

$$\int_{\Omega_\varepsilon \setminus B_\varepsilon} \left|\nabla\left(\psi\left(\frac{x_1}{\varepsilon}, \frac{x_3 - l_3}{\varepsilon}\right)\right)\right|^2 dx \leq C\,e^{-\frac{c}{\varepsilon}},$$

where $B_\varepsilon = (0, l_1) \times (l_3 - \varepsilon, l_3 + \varepsilon)$, and ψ is the function defined by $\psi = \psi^-$ in Λ^- and $\psi = \psi^+$ in Λ^+.

4. A corrector result. To build a corrector for the solution u_ε of problem (1.2), we need more regularity on the solution u^- of problem (2.4). We then assume the following regularity for f and b:

$$f \in H^4_{\rm per}(\Omega^-) \cap L^2(\Omega), \quad b \in H^6_{\rm per}(0, l_1). \tag{4.1}$$

Let $\mathcal{O}^- = \left\{(x_1, x_3) \in \mathbf{R}^2 \,:\, x_1 \in \mathbf{R},\ b(x_1) < x_3 < l_3\right\}$. The extension of u^- to $\mathcal{O}^-$ by l_1-periodicity is a solution in $H^1_{\rm per}(\Omega^-)$ of

$$\begin{cases} -\Delta u^- = f & \text{in } \mathcal{O}^-, \\ u^- = 0 & \text{on } \mathbf{R} \times \{l_3\}, \\ u^- = g & \text{on } \{(x_1, b(x_1)) \,:\, x_1 \in \mathbf{R}\}. \end{cases}$$

Using standard regularity results (see [19, 20]), we then have

$$u^- \in H^6_{\rm per}\left(\Omega^-\right) \subset C^4\left(\overline{\Omega^-}\right). \tag{4.2}$$

Let w now be the function defined by

$$w = \begin{cases} 0 & \text{in } \Omega^+, \\ w^- & \text{in } \Omega^-, \end{cases} \tag{4.3}$$

where w^- is the unique solution in $H^1_{\rm per}(\Omega^-)$ of

$$\begin{cases} \Delta w^- = 0 & \text{in } \Omega^-, \\ w^- = \beta\,\dfrac{\partial u^-}{\partial x_3} & \text{on } \Sigma, \\ w^- = 0 & \text{on } P, \end{cases} \tag{4.4}$$

where u^- is the solution of problem (2.4) and β is defined by (3.3). Let us point out that, due to assumptions (4.1),

$$w^- \in C^3(\overline{\Omega^-}). \tag{4.5}$$

Indeed, the functions w^- and u^- being extended by l_1-periodicity to $\mathcal{O}^-$, it follows that w^- is the solution in $H^1_{\rm per}(\Omega^-)$ of

$$\begin{cases} \Delta w^- = 0 & \text{in } \mathcal{O}^-, \\ w^- = \beta \dfrac{\partial u^-}{\partial x_3} & \text{on } \mathbf{R} \times \{l_3\}, \\ w^- = 0 & \text{on } \{(x_1, b(x_1)) \,:\, x_1 \in \mathbf{R}\}. \end{cases}$$

Due to (4.2), we have $\frac{\partial u^-}{\partial x_3} \in H^5_{\rm per}(\Omega^-)$, and, consequently, $w^- \in H^5_{\rm per}(\Omega^-) \subset C^3(\overline{\Omega^-})$.

This section is devoted to proving the following result.

THEOREM 4.1. *Assume* (4.1). *Let u_ε be the solution of problem* (1.2), *u be defined by* (2.3), (2.4), *and w be defined by* (4.3), (4.4). *Then there exists a positive constant c, independent of ε, such that*

$$\|u_\epsilon - u\|_{H^1(\Omega_\varepsilon \setminus B_\epsilon)} \le c\,\varepsilon \tag{4.6}$$

for ε small enough, where $B_\varepsilon = (0, l_1) \times (l_3 - \varepsilon, l_3 + \varepsilon)$. If in addition $f = 0$ in Ω^+, there exists a positive constant c, independent of ε, such that

$$\|u_\epsilon - u - \varepsilon w\|_{H^1(\Omega_\varepsilon \setminus B_\varepsilon)} \le c\,\varepsilon^{\frac{3}{2}} \tag{4.7}$$

for ε small enough.

To prove Theorem 4.1 we need to introduce some auxiliary functions. Let τ_ε be the function defined by

(4.8)
$$\tau_\varepsilon = \begin{cases} \tau_\varepsilon^+ = u_\varepsilon - \varepsilon w_\varepsilon^+ - \varepsilon \dfrac{\partial u^-}{\partial x_3}(x_1, l_3)\, \psi^+\left(\dfrac{x_1}{\varepsilon}, \dfrac{x_3 - l_3}{\varepsilon}\right) & \text{in } \Omega_\varepsilon^+, \\ \tau_\varepsilon^- = u_\varepsilon - u^- - \varepsilon w_\varepsilon^- - \varepsilon \dfrac{\partial u^-}{\partial x_3}(x_1, l_3) \left(\psi^-\left(\dfrac{x_1}{\varepsilon}, \dfrac{x_3 - l_3}{\varepsilon}\right) - \beta\right) & \text{in } \Omega^-, \end{cases}$$

and let ρ_ε be the function defined by

$$\rho_\varepsilon = \begin{cases} \rho_\varepsilon^+ = w_\varepsilon^+ - \varepsilon \dfrac{\partial w^-}{\partial x_3}(x_1, l_3)\, \psi^+\left(\dfrac{x_1}{\varepsilon}, \dfrac{x_3 - l_3}{\varepsilon}\right) & \text{in } \Omega_\varepsilon^+, \\ \rho_\varepsilon^- = w_\varepsilon^- - w^- - \varepsilon \dfrac{\partial w^-}{\partial x_3}(x_1, l_3)\, \psi^-\left(\dfrac{x_1}{\varepsilon}, \dfrac{x_3 - l_3}{\varepsilon}\right) & \text{in } \Omega^-, \end{cases} \tag{4.9}$$

where u_ε is the solution of problem (1.2), u^- is the solution of problem (2.4), w^- is the solution of problem (4.4), $\psi^\pm$ are the functions defined by (3.1), (3.2), and $w_\varepsilon^\pm$ are functions in $H^1(\Omega^+)$ and $H^1_{\rm per}(\Omega^-)$, respectively, satisfying

$$\begin{cases} \Delta w_\varepsilon^+ = 0 & \text{in } \Omega_\varepsilon^+, \\ \Delta w_\varepsilon^- = 0 & \text{in } \Omega^-, \\ w_\varepsilon^+ = 0 & \text{on } R_\varepsilon \setminus \Sigma, \\ w_\varepsilon^- = \beta \dfrac{\partial u^-}{\partial x_3} & \text{on } R_\varepsilon \cap \Sigma, \\ w_\varepsilon^- = 0 & \text{on } P, \\ w_\varepsilon^+ = w_\varepsilon^- - \beta \dfrac{\partial u^-}{\partial x_3} & \text{on } \Sigma \setminus R_\varepsilon, \\ \dfrac{\partial w_\varepsilon^+}{\partial x_3} = \dfrac{\partial w_\varepsilon^-}{\partial x_3} & \text{on } \Sigma \setminus R_\varepsilon, \end{cases} \tag{4.10}$$

β being defined by (3.3). Theorem 4.1 will be an immediate consequence of the two following propositions.

PROPOSITION 4.2. *Assume* (4.1). *Let* τ_ε *be the function defined by* (4.8). *Then there exists a positive constant* c, *independent of* ε, *such that*

$$\|\tau_\epsilon\|_{H^1(\Omega_\varepsilon)} \leq c\,\varepsilon, \tag{4.11}$$

and, if $f = 0$ *in* Ω^+,

$$\|\tau_\epsilon\|_{H^1(\Omega_\varepsilon)} \leq c\,\varepsilon^{\frac{3}{2}} \tag{4.12}$$

for ε *small enough.*

PROPOSITION 4.3. *Assume* (4.1). *Let* ρ_ε *be the function defined by* (4.9). *Then there exists a positive constant* c, *independent of* ε, *such that*

$$\|\rho_\epsilon\|_{H^1(\Omega_\varepsilon)} \leq c\,\varepsilon \tag{4.13}$$

for ε *small enough.*

Proof of Proposition 4.2. Obviously, $\tau_\varepsilon^+ \in H^1(\Omega_\varepsilon^+)$ and $\tau_\varepsilon^- \in H^1(\Omega^-)$. Due to the boundary conditions of u_ε, u, $\psi^\pm$, and $w_\varepsilon^\pm$, the functions τ_ε^+ and τ_ε^- have the same trace on $\overline{\Omega_\epsilon^+} \cap \overline{\Omega^-}$. Consequently, $\tau_\varepsilon \in H^1(\Omega_\varepsilon)$. Moreover, τ_ε is l_1-periodic with respect to x_1 for a.e. $x_3 \in (b(0), l_3)$ and $\tau_\varepsilon = 0$ on $R_\varepsilon \backslash \{(x_1, l_3') : x_1 \in (0, l_1)\}$. Furthermore, from the jump conditions in (3.2) and in (4.10) it follows that the normal derivatives of τ_ε^+ and τ_ε^- on $\overline{\Omega_\epsilon^+} \cap \overline{\Omega^-}$ are opposite as elements of $H^{-\frac{1}{2}}(\overline{\Omega_\epsilon^+} \cap \overline{\Omega^-})$. Consequently, $\Delta\tau_\varepsilon$ is weakly defined in Ω_ε by

$$\Delta\tau_\varepsilon = \begin{cases} -\varepsilon \dfrac{\partial^3 u^-}{\partial x_1^2 \partial x_3}(x_1, l_3)\, \psi^+\left(\dfrac{x_1}{\varepsilon}, \dfrac{x_3 - l_3}{\varepsilon}\right) \\ +\,-\,2\varepsilon \dfrac{\partial^2 u^-}{\partial x_1 \partial x_3}(x_1, l_3)\, \dfrac{\partial}{\partial x_1}\left(\psi^+\left(\dfrac{x_1}{\varepsilon}, \dfrac{x_3 - l_3}{\varepsilon}\right)\right) - f & \text{in } \Omega_\varepsilon^+, \\ -\varepsilon \dfrac{\partial^3 u^-}{\partial x_1^2 \partial x_3}(x_1, l_3)\left(\psi^-\left(\dfrac{x_1}{\varepsilon}, \dfrac{x_3 - l_3}{\varepsilon}\right) - \beta\right) \\ +\,-\,2\varepsilon \dfrac{\partial^2 u^-}{\partial x_1 \partial x_3}(x_1, l_3)\, \dfrac{\partial}{\partial x_1}\left(\psi^-\left(\dfrac{x_1}{\varepsilon}, \dfrac{x_3 - l_3}{\varepsilon}\right) - \beta\right) & \text{in } \Omega^-. \end{cases} \tag{4.14}$$

Since

$$\tau_{\varepsilon|_P} = -\varepsilon \frac{\partial u^-}{\partial x_3}(x_1, l_3)\left(\psi^-\left(\frac{x_1}{\varepsilon}, \frac{b(x_1) - l_3}{\varepsilon}\right) - \beta\right),$$

$$\tau_{\varepsilon|_{R_\varepsilon \cap ((0,l_1)\times l_3')}} = -\varepsilon \frac{\partial u^-}{\partial x_3}(x_1, l_3)\, \psi^+\left(\frac{x_1}{\varepsilon}, \frac{l_3' - l_3}{\varepsilon}\right),$$

setting

$$\tau_\varepsilon^1(x_1,x_3) = -\varepsilon \frac{\partial u^-}{\partial x_3}(x_1,l_3)\left(\psi^-\left(\frac{x_1}{\varepsilon},\frac{b(x_1)-l_3}{\varepsilon}\right)-\beta\right) m_1(x_3) \text{ in } \Omega_\varepsilon,$$

$$\tau_\varepsilon^2(x_1,x_3) = -\varepsilon \frac{\partial u^-}{\partial x_3}(x_1,l_3)\,\psi^+\left(\frac{x_1}{\varepsilon},\frac{l_3'-l_3}{\varepsilon}\right) m_2(x_3) \text{ in } \Omega_\varepsilon,$$

where $m_1,\, m_2 \in C^2(\mathbb{R};[0,1])$ and

$$m_1(t) = \begin{cases} 0 & \text{if } t > \dfrac{M+l_3}{2}, \\ 1 & \text{if } t < \dfrac{3M+l_3}{4}, \end{cases} \qquad m_2(t) = \begin{cases} 1 & \text{if } t > \dfrac{l_3+l_3'}{2}, \\ 0 & \text{if } t < \dfrac{3l_3+l_3'}{4}, \end{cases} \tag{4.15}$$

M being defined by (2.1), it results that $\tau_\epsilon - \tau_\varepsilon^1 - \tau_\varepsilon^2 \in H^1_{\mathrm{per}}(\Omega_\varepsilon)$ and vanishes on $R_\varepsilon \cup P$. Then, multiplying (4.14) by $\tau_\epsilon - \tau_\varepsilon^1 - \tau_\varepsilon^2$ and integrating on Ω_ε, it follows that

$$\begin{aligned}
&\int_{\Omega_\varepsilon} |\nabla \tau_\epsilon|^2\, dx \\
&= \int_{\Omega^-} \nabla\tau_\epsilon \nabla\tau_\epsilon^1 dx + \int_{\Omega_\varepsilon^+} \nabla\tau_\epsilon\nabla\tau_\epsilon^2 dx - \int_{\Omega_\varepsilon} \Delta\tau_\epsilon\left(\tau_\epsilon - \tau_\varepsilon^1 - \tau_\epsilon^2\right) dx \\
&= \int_{\Omega^-} \nabla\tau_\epsilon \nabla\tau_\epsilon^1 dx + \int_{\Omega_\varepsilon^+} \nabla\tau_\epsilon\nabla\tau_\epsilon^2 dx \\
&+ \varepsilon \int_{\Omega^-} \frac{\partial^3 u^-}{\partial x_1^2 \partial x_3}(x_1,l_3)\left(\psi^-\left(\frac{x_1}{\varepsilon},\frac{x_3-l_3}{\varepsilon}\right)-\beta\right)\left(\tau_\epsilon - \tau_\varepsilon^1\right) dx \\
&+ 2\varepsilon \int_{\Omega^-} \frac{\partial^2 u^-}{\partial x_1 \partial x_3}(x_1,l_3)\frac{\partial}{\partial x_1}\left(\psi^-\left(\frac{x_1}{\varepsilon},\frac{x_3-l_3}{\varepsilon}\right)-\beta\right)\left(\tau_\epsilon - \tau_\varepsilon^1\right) dx \\
&+ \varepsilon \int_{\Omega_\varepsilon^+} \frac{\partial^3 u^-}{\partial x_1^2 \partial x_3}(x_1,l_3)\,\psi^+\left(\frac{x_1}{\varepsilon},\frac{x_3-l_3}{\varepsilon}\right)\left(\tau_\epsilon - \tau_\epsilon^2\right) dx \\
&+ 2\varepsilon \int_{\Omega_\varepsilon^+} \frac{\partial^2 u^-}{\partial x_1 \partial x_3}(x_1,l_3)\frac{\partial}{\partial x_1}\left(\psi^+\left(\frac{x_1}{\varepsilon},\frac{x_3-l_3}{\varepsilon}\right)\right)\left(\tau_\epsilon - \tau_\epsilon^2\right) dx \\
&+ \int_{\Omega_\varepsilon^+} f\left(\tau_\epsilon - \tau_\epsilon^2\right) dx.
\end{aligned} \tag{4.16}$$

Let us estimate each term on the right-hand side of (4.16). We first compute the derivatives of τ_ϵ^1 and τ_ϵ^2:

$$\begin{aligned}
\frac{\partial \tau_\epsilon^1}{\partial x_1}(x_1,x_3) = &-\varepsilon \frac{\partial^2 u^-}{\partial x_1 \partial x_3}(x_1,l_3)\left(\psi^-\left(\frac{x_1}{\varepsilon},\frac{b(x_1)-l_3}{\varepsilon}\right)-\beta\right) m_1(x_3) \\
&- \frac{\partial u^-}{\partial x_3}(x_1,l_3)\frac{\partial \psi^-}{\partial y_1}\left(\frac{x_1}{\varepsilon},\frac{b(x_1)-l_3}{\varepsilon}\right) m_1(x_3) \\
&- \frac{\partial u^-}{\partial x_3}(x_1,l_3)\frac{\partial \psi^-}{\partial y_3}\left(\frac{x_1}{\varepsilon},\frac{b(x_1)-l_3}{\varepsilon}\right)\frac{db}{dx_1}(x_1)\, m_1(x_3) \quad \text{in } \Omega^-,
\end{aligned}$$

$$\frac{\partial \tau_\epsilon^1}{\partial x_3}(x_1,x_3) = -\varepsilon \frac{\partial u^-}{\partial x_3}(x_1,l_3)\left(\psi^-\left(\frac{x_1}{\varepsilon},\frac{b(x_1)-l_3}{\varepsilon}\right)-\beta\right)\frac{dm_1}{dx_3}(x_3) \quad \text{in } \Omega^-.$$

Then, from (4.2) and Proposition 3.1, we have

$$\left|\frac{\partial \tau_\epsilon^1}{\partial x_1}\right| \le Ce^{-\frac{c}{\varepsilon}}, \quad \left|\frac{\partial \tau_\epsilon^1}{\partial x_3}\right| \le Ce^{-\frac{c}{\varepsilon}} \quad \text{in } \Omega^-, \tag{4.17}$$

for ε small enough. Here and in what follows C and c denote positive constants independent of ε. Similarly,

$$\left|\frac{\partial \tau_\epsilon^2}{\partial x_1}\right| \le Ce^{-\frac{c}{\varepsilon}}, \quad \left|\frac{\partial \tau_\epsilon^2}{\partial x_3}\right| \le Ce^{-\frac{c}{\varepsilon}} \quad \text{in } \Omega_\varepsilon^+, \tag{4.18}$$

for ε small enough. For the first two terms on the right-hand side of (4.16), from the Cauchy–Schwarz inequality, (4.17) and (4.18), it follows that

$$\left|\int_{\Omega^-} \nabla\tau_\epsilon \nabla\tau_\epsilon^1 dx + \int_{\Omega_\varepsilon^+} \nabla\tau_\epsilon \nabla\tau_\epsilon^2 dx\right| \le Ce^{-\frac{c}{\varepsilon}} \left\|\nabla\tau_\epsilon\right\|_{(L^2(\Omega_\varepsilon))^2} \tag{4.19}$$

for ε small enough. For the third and fifth terms on the right-hand side of (4.16), Corollary 3.2, the Cauchy–Schwarz inequality, the Poincaré inequality, (4.2), (4.17), and (4.18) give

(4.20)

$$\begin{aligned}
&\left|\varepsilon \int_{\Omega^-} \frac{\partial^3 u^-}{\partial x_1^2 \partial x_3}(x_1, l_3) \left(\psi^-\left(\frac{x_1}{\varepsilon}, \frac{x_3 - l_3}{\varepsilon}\right) - \beta\right)\left(\tau_\epsilon - \tau_\varepsilon^1\right) dx\right. \\
&\left. + \varepsilon \int_{\Omega_\varepsilon^+} \frac{\partial^3 u^-}{\partial x_1^2 \partial x_3}(x_1, l_3)\, \psi^+\left(\frac{x_1}{\varepsilon}, \frac{x_3 - l_3}{\varepsilon}\right)\left(\tau_\epsilon - \tau_\epsilon^2\right) dx\right| \\
&\le C\varepsilon^{\frac{3}{2}} \left(\left\|\tau_\epsilon - \tau_\epsilon^1\right\|_{L^2(\Omega^-)} + \left\|\tau_\epsilon - \tau_\epsilon^2\right\|_{L^2(\Omega_\varepsilon^+)}\right) \\
&\le C\varepsilon^{\frac{3}{2}} \left(\left\|\nabla\left(\tau_\epsilon - \tau_\epsilon^1\right)\right\|_{(L^2(\Omega^-))^2} + \left\|\nabla\left(\tau_\epsilon - \tau_\epsilon^2\right)\right\|_{\left(L^2(\Omega_\varepsilon^+)\right)^2}\right) \\
&\le C\varepsilon^{\frac{3}{2}} \left(\left\|\nabla\tau_\epsilon\right\|_{(L^2(\Omega^-))^2} + \left\|\nabla\tau_\epsilon^1\right\|_{(L^2(\Omega^-))^2} + \left\|\nabla\tau_\epsilon\right\|_{\left(L^2(\Omega_\varepsilon^+)\right)^2} + \left\|\nabla\tau_\epsilon^2\right\|_{\left(L^2(\Omega_\varepsilon^+)\right)^2}\right) \\
&\le C\left(\varepsilon^{\frac{3}{2}} \left\|\nabla\tau_\epsilon\right\|_{(L^2(\Omega_\varepsilon))^2} + e^{-\frac{c}{\varepsilon}}\right)
\end{aligned}$$

for ε small enough. Integrating by parts the fourth and sixth terms on the right-hand side of (4.16), it follows that

$$\begin{aligned}
&2\varepsilon \int_{\Omega^-} \frac{\partial^2 u^-}{\partial x_1 \partial x_3}(x_1, l_3) \frac{\partial}{\partial x_1}\left(\psi^-\left(\frac{x_1}{\varepsilon}, \frac{x_3 - l_3}{\varepsilon}\right) - \beta\right)\left(\tau_\epsilon - \tau_\varepsilon^1\right) dx \\
&+ 2\varepsilon \int_{\Omega_\varepsilon^+} \frac{\partial^2 u^-}{\partial x_1 \partial x_3}(x_1, l_3) \frac{\partial}{\partial x_1}\left(\psi^+\left(\frac{x_1}{\varepsilon}, \frac{x_3 - l_3}{\varepsilon}\right)\right)\left(\tau_\epsilon - \tau_\epsilon^2\right) dx \\
&= -2\varepsilon \int_{\Omega^-} \frac{\partial^2 u^-}{\partial x_1 \partial x_3}(x_1, l_3)\left(\psi^-\left(\frac{x_1}{\varepsilon}, \frac{x_3 - l_3}{\varepsilon}\right) - \beta\right) \frac{\partial}{\partial x_1}\left(\tau_\epsilon - \tau_\varepsilon^1\right) dx \\
&\quad -2\varepsilon \int_{\Omega^-} \frac{\partial^3 u^-}{\partial x_1^2 \partial x_3}(x_1, l_3)\left(\psi^-\left(\frac{x_1}{\varepsilon}, \frac{x_3 - l_3}{\varepsilon}\right) - \beta\right)\left(\tau_\epsilon - \tau_\varepsilon^1\right) dx \\
&\quad -2\varepsilon \int_{\Omega_\varepsilon^+} \frac{\partial^2 u^-}{\partial x_1 \partial x_3}(x_1, l_3)\, \psi^+\left(\frac{x_1}{\varepsilon}, \frac{x_3 - l_3}{\varepsilon}\right) \frac{\partial}{\partial x_1}\left(\tau_\epsilon - \tau_\varepsilon^2\right) dx \\
&\quad -2\varepsilon \int_{\Omega_\varepsilon^+} \frac{\partial^3 u^-}{\partial x_1^2 \partial x_3}(x_1, l_3)\, \psi^+\left(\frac{x_1}{\varepsilon}, \frac{x_3 - l_3}{\varepsilon}\right)\left(\tau_\epsilon - \tau_\varepsilon^2\right) dx.
\end{aligned}$$

Consequently, Corollary 3.2, the Cauchy–Schwarz inequality, the Poincaré inequality, (4.2), (4.17), and (4.18) imply

(4.21)
$$
\begin{aligned}
&\left|2\varepsilon\int_{\Omega^-}\frac{\partial^2 u^-}{\partial x_1\partial x_3}(x_1,l_3)\,\frac{\partial}{x_1}\left(\psi^-\left(\frac{x_1}{\varepsilon},\frac{x_3-l_3}{\varepsilon}\right)-\beta\right)\left(\tau_\epsilon-\tau_\varepsilon^1\right)dx\right.\\
&\left.+2\varepsilon\int_{\Omega_\varepsilon^+}\frac{\partial^2 u^-}{\partial x_1\partial x_3}(x_1,l_3)\,\frac{\partial}{x_1}\left(\psi^+\left(\frac{x_1}{\varepsilon},\frac{x_3-l_3}{\varepsilon}\right)\right)\left(\tau_\epsilon-\tau_\epsilon^2\right)dx\right|\\
&\le C\varepsilon^{\frac{3}{2}}\left(\left\|\nabla\left(\tau_\epsilon-\tau_\epsilon^1\right)\right\|_{(L^2(\Omega^-))^2}+\left\|\tau_\epsilon-\tau_\epsilon^1\right\|_{L^2(\Omega^-)}\right.\\
&\qquad\left.+\left\|\nabla\left(\tau_\epsilon-\tau_\epsilon^2\right)\right\|_{\left(L^2(\Omega_\varepsilon^+)\right)^2}+\left\|\tau_\epsilon-\tau_\epsilon^2\right\|_{L^2(\Omega_\varepsilon^+)}\right)\\
&\le C\varepsilon^{\frac{3}{2}}\left(\left\|\nabla\left(\tau_\epsilon-\tau_\epsilon^1\right)\right\|_{(L^2(\Omega^-))^2}+\left\|\nabla\left(\tau_\epsilon-\tau_\epsilon^2\right)\right\|_{\left(L^2(\Omega_\varepsilon^+)\right)^2}\right)\\
&\le C\varepsilon^{\frac{3}{2}}\left(\left\|\nabla\tau_\epsilon\right\|_{(L^2(\Omega^-))^2}+\left\|\nabla\tau_\epsilon^1\right\|_{(L^2(\Omega^-))^2}+\left\|\nabla\tau_\epsilon\right\|_{\left(L^2(\Omega_\varepsilon^+)\right)^2}+\left\|\nabla\tau_\epsilon^2\right\|_{\left(L^2(\Omega_\varepsilon^+)\right)^2}\right)\\
&\le C\left(\varepsilon^{\frac{3}{2}}\left\|\nabla\tau_\epsilon\right\|_{(L^2(\Omega_\varepsilon))^2}+e^{-\frac{c}{\varepsilon}}\right)
\end{aligned}
$$

for ε enough small. For the last term on the right-hand side of (4.16), we observe, using the Cauchy–Schwarz inequality, that

$$
\begin{aligned}
&\int_{\Omega_\varepsilon^+}\left|\tau_\epsilon(x_1,x_3)-\tau_\epsilon^2(x_1,x_3)\right|^2dx\\
&=\sum_{k=0}^{1/\varepsilon-1}\int_{l_3}^{l_3'}\int_{\varepsilon(a_1+kl_1)}^{\varepsilon(b_1+kl_1)}\left|\tau_\epsilon(x_1,x_3)-\tau_\epsilon^2(x_1,x_3)\right|^2dx\\
&=\sum_{k=0}^{1/\varepsilon-1}\int_{l_3}^{l_3'}\int_{\varepsilon(a_1+kl_1)}^{\varepsilon(b_1+kl_1)}\left|\int_{\varepsilon(a_1+kl_1)}^{x_1}\frac{\partial\left(\tau_\epsilon-\tau_\epsilon^2\right)}{\partial t}(t,x_3)\,dt\right|^2dx\\
&\le\varepsilon^2(b_1-a_1)^2\sum_{k=0}^{1/\varepsilon-1}\int_{l_3}^{l_3'}\int_{\varepsilon(a_1+kl_1)}^{\varepsilon(b_1+kl_1)}\left|\frac{\partial\left(\tau_\epsilon-\tau_\epsilon^2\right)}{\partial x_1}\right|^2dx\\
&\le\varepsilon^2(b_1-a_1)^2\int_{\Omega_\varepsilon^+}\left|\frac{\partial\left(\tau_\epsilon-\tau_\epsilon^2\right)}{\partial x_1}\right|^2dx.
\end{aligned}
$$

From (4.18) it then follows that

$$
\left|\int_{\Omega_\varepsilon^+}f\left(\tau_\epsilon-\tau_\epsilon^2\right)dx\right|\le C\left(\varepsilon\left\|\nabla\tau_\epsilon\right\|_{(L^2(\Omega_\varepsilon))^2}+e^{-\frac{c}{\varepsilon}}\right)\tag{4.22}
$$

for ε small enough. Combining (4.16) with (4.19) $\div$ (4.22), we have

$$
\left\|\nabla\tau_\epsilon\right\|^2_{(L^2(\Omega_\varepsilon))^2}\le C\left(\varepsilon\left\|\nabla\tau_\epsilon\right\|_{(L^2(\Omega_\varepsilon))^2}+e^{-\frac{c}{\varepsilon}}\right)
$$

for ε small enough. Therefore

$$
\left\|\nabla\tau_\epsilon\right\|_{(L^2(\Omega_\varepsilon))^2}\le C\varepsilon\tag{4.23}
$$

for ε small enough. Estimate (4.11) follows from (4.23), using the Poincaré inequality.

If $f = 0$ in Ω^+, from identity (4.16) combined with (4.19) $\div$ (4.21), we have

$$\|\nabla\tau_\epsilon\|^2_{(L^2(\Omega_\varepsilon))^2} \leq C\left(\varepsilon^{\frac{3}{2}}\|\nabla\tau_\epsilon\|_{(L^2(\Omega_\varepsilon))^2} + e^{-\frac{c}{\varepsilon}}\right)$$

for ε small enough. Therefore

$$\|\nabla\tau_\epsilon\|_{(L^2(\Omega_\varepsilon))^2} \leq C\varepsilon^{\frac{3}{2}}$$

for ε small enough, from which estimate (4.12) follows, due to the Poincaré inequality. □

The proof of Proposition 4.3 has the same framework as the proof of Proposition 4.2.

Proof of Proposition 4.3. Obviously, $\rho_\varepsilon^+ \in H^1(\Omega_\varepsilon^+)$ and $\rho_\varepsilon^- \in H^1(\Omega^-)$. Due to the boundary conditions of $w_\varepsilon^\pm$, w^-, and $\psi^\pm$, the functions ρ_ε^+ and ρ_ε^- have the same trace on $\overline{\Omega_\epsilon^+} \cap \overline{\Omega^-}$. Consequently, $\rho_\varepsilon \in H^1(\Omega_\varepsilon)$. Moreover, ρ_ε is l_1-periodic with respect to x_1 for a.e. $x_3 \in (b(0), l_3)$ and $\rho_\varepsilon = 0$ on $R_\varepsilon \setminus \{(x_1, l_3') : x_1 \in (0, l_1)\}$. Furthermore, from the jump condition in (3.2) and in (4.10), it follows that the normal derivatives of ρ_ε^+ and ρ_ε^- on $\overline{\Omega_\epsilon^+} \cap \overline{\Omega^-}$ are opposite as elements of $H^{-\frac{1}{2}}(\overline{\Omega_\epsilon^+} \cap \overline{\Omega^-})$. Consequently, $\Delta\rho_\varepsilon$ is weakly defined in Ω_ε and satisfies

(4.24)
$$\Delta\rho_\varepsilon = \begin{cases} -\varepsilon\dfrac{\partial^3 w^-}{\partial x_1^2 \partial x_3}(x_1, l_3)\,\psi^+\left(\dfrac{x_1}{\varepsilon}, \dfrac{x_3 - l_3}{\varepsilon}\right) \\ +\!- 2\varepsilon\dfrac{\partial^2 w^-}{\partial x_1 \partial x_3}(x_1, l_3)\,\dfrac{\partial}{x_1}\left(\psi^+\left(\dfrac{x_1}{\varepsilon}, \dfrac{x_3 - l_3}{\varepsilon}\right)\right) & \text{in } \Omega_\varepsilon^+, \\ -\varepsilon\dfrac{\partial^3 w^-}{\partial x_1^2 \partial x_3}(x_1, l_3)\,\psi^-\left(\dfrac{x_1}{\varepsilon}, \dfrac{x_3 - l_3}{\varepsilon}\right) \\ +\!- 2\varepsilon\dfrac{\partial^2 w^-}{\partial x_1 \partial x_3}(x_1, l_3)\,\dfrac{\partial}{x_1}\left(\psi^-\left(\dfrac{x_1}{\varepsilon}, \dfrac{x_3 - l_3}{\varepsilon}\right)\right) & \text{in } \Omega^-. \end{cases}$$

We have

$$\rho_{\varepsilon|_P} = -\varepsilon\frac{\partial w^-}{\partial x_3}(x_1, l_3)\,\psi^-\left(\frac{x_1}{\varepsilon}, \frac{b(x_1) - l_3}{\varepsilon}\right),$$

$$\rho_{\varepsilon|_{R_\varepsilon \cap ((0,l_1)\times l_3')}} = -\varepsilon\frac{\partial w^-}{\partial x_3}(x_1, l_3)\,\psi^+\left(\frac{x_1}{\varepsilon}, \frac{l_3' - l_3}{\varepsilon}\right),$$

and then, setting

$$\rho_\varepsilon^1(x_1, x_3) = -\varepsilon\frac{\partial w^-}{\partial x_3}(x_1, l_3)\,\psi^-\left(\frac{x_1}{\varepsilon}, \frac{b(x_1) - l_3}{\varepsilon}\right) m_1(x_3) \text{ in } \Omega_\varepsilon,$$

$$\rho_\varepsilon^2(x_1, x_3) = -\varepsilon\frac{\partial w^-}{\partial x_3}(x_1, l_3)\,\psi^+\left(\frac{x_1}{\varepsilon}, \frac{l_3' - l_3}{\varepsilon}\right) m_2(x_3) \text{ in } \Omega_\varepsilon,$$

the functions m_1 and m_2 being defined by (4.15), it follows that $\rho_\epsilon - \rho_\varepsilon^1 - \rho_\varepsilon^2 \in H^1_{\mathrm{per}}(\Omega_\varepsilon)$ and vanishes on $R_\varepsilon \cup P$. Then, multiplying (4.24) by $\rho_\epsilon - \rho_\varepsilon^1 - \rho_\varepsilon^2$ and integrating on Ω_ε, we find

$$
\begin{aligned}
&\int_{\Omega_\varepsilon} |\nabla \rho_\epsilon|^2 \, dx \\
&= \int_{\Omega^-} \nabla \rho_\epsilon \nabla \rho_\epsilon^1 dx + \int_{\Omega_\varepsilon^+} \nabla \rho_\epsilon \nabla \rho_\epsilon^2 dx - \int_{\Omega_\varepsilon} \Delta \rho_\epsilon \left(\rho_\epsilon - \rho_\varepsilon^1 - \rho_\epsilon^2 \right) dx \\
&= \int_{\Omega^-} \nabla \rho_\epsilon \nabla \rho_\epsilon^1 dx + \int_{\Omega_\varepsilon^+} \nabla \rho_\epsilon \nabla \rho_\epsilon^2 dx \\
&+ \varepsilon \int_{\Omega^-} \frac{\partial^3 w^-}{\partial x_1^2 \partial x_3}(x_1, l_3)\, \psi^- \left(\frac{x_1}{\varepsilon}, \frac{x_3 - l_3}{\varepsilon} \right) \left(\rho_\epsilon - \rho_\varepsilon^1 \right) dx \\
&+ 2\varepsilon \int_{\Omega^-} \frac{\partial^2 w^-}{\partial x_1 \partial x_3}(x_1, l_3) \frac{\partial}{\partial x_1} \left(\psi^- \left(\frac{x_1}{\varepsilon}, \frac{x_3 - l_3}{\varepsilon} \right) \right) \left(\rho_\epsilon - \rho_\varepsilon^1 \right) dx \\
&+ \varepsilon \int_{\Omega_\varepsilon^+} \frac{\partial^3 w^-}{\partial x_1^2 \partial x_3}(x_1, l_3)\, \psi^+ \left(\frac{x_1}{\varepsilon}, \frac{x_3 - l_3}{\varepsilon} \right) \left(\rho_\epsilon - \rho_\epsilon^2 \right) dx \\
&+ 2\varepsilon \int_{\Omega_\varepsilon^+} \frac{\partial^2 w^-}{\partial x_1 \partial x_3}(x_1, l_3) \frac{\partial}{\partial x_1} \left(\psi^+ \left(\frac{x_1}{\varepsilon}, \frac{x_3 - l_3}{\varepsilon} \right) \right) \left(\rho_\epsilon - \rho_\epsilon^2 \right) dx.
\end{aligned} \tag{4.25}
$$

Let us estimate each term on the right-hand side of (4.25). First, the derivatives of ρ_ϵ^1 and ρ_ϵ^2 are

$$
\begin{aligned}
\frac{\partial \rho_\epsilon^1}{\partial x_1}(x_1, x_3) = &-\varepsilon \frac{\partial^2 w^-}{\partial x_1 \partial x_3}(x_1, l_3)\, \psi^- \left(\frac{x_1}{\varepsilon}, \frac{b(x_1) - l_3}{\varepsilon} \right) m_1(x_3) \\
&- \frac{\partial w^-}{\partial x_3}(x_1, l_3) \frac{\partial \psi^-}{\partial y_1} \left(\frac{x_1}{\varepsilon}, \frac{b(x_1) - l_3}{\varepsilon} \right) m_1(x_3) \\
&- \frac{\partial w^-}{\partial x_3}(x_1, l_3) \frac{\partial \psi^-}{\partial y_3} \left(\frac{x_1}{\varepsilon}, \frac{b(x_1) - l_3}{\varepsilon} \right) \frac{db}{dx_1}(x_1)\, m_1(x_3) \quad \text{in } \Omega^-,
\end{aligned}
$$

$$
\frac{\partial \rho_\epsilon^1}{\partial x_3}(x_1, x_3) = -\varepsilon \frac{\partial w^-}{\partial x_3}(x_1, l_3)\, \psi^- \left(\frac{x_1}{\varepsilon}, \frac{b(x_1) - l_3}{\varepsilon} \right) \frac{dm_1}{dx_3}(x_3) \quad \text{in } \Omega^-.
$$

Then, Proposition 3.1 and (4.5) imply

$$
\left| \frac{\partial \rho_\epsilon^1}{\partial x_1} \right| \le C\, \varepsilon, \quad \left| \frac{\partial \rho_\epsilon^1}{\partial x_3} \right| \le C\, \varepsilon \quad \text{in } \Omega^-, \tag{4.26}
$$

for ε small enough. Similarly,

$$
\left| \frac{\partial \rho_\epsilon^2}{\partial x_1} \right| \le C e^{-\frac{c}{\varepsilon}}, \quad \left| \frac{\partial \rho_\epsilon^2}{\partial x_3} \right| \le C e^{-\frac{c}{\varepsilon}} \quad \text{in } \Omega_\varepsilon^+, \tag{4.27}
$$

for ε small enough. From the Cauchy–Schwarz inequality, (4.26) and (4.27), the first

two terms on the right-hand side of (4.25) satisfy

$$\left|\int_{\Omega^-} \nabla\rho_\epsilon \nabla\rho_\epsilon^1 dx + \int_{\Omega_\varepsilon^+} \nabla\rho_\epsilon \nabla\rho_\epsilon^2 dx\right| \leq C\,\varepsilon\,\|\nabla\rho_\epsilon\|_{(L^2(\Omega_\varepsilon))^2} \tag{4.28}$$

for ε small enough. For the third and fifth terms on the right-hand side of (4.25), Corollary 3.2, the Cauchy–Schwarz inequality, the Poincaré inequality, (4.5), (4.26), and (4.27) give

(4.29)

$$\begin{aligned}
&\left|\varepsilon\int_{\Omega^-} \frac{\partial^3 w^-}{\partial x_1^2 \partial x_3}(x_1,l_3)\,\psi^-\left(\frac{x_1}{\varepsilon},\frac{x_3-l_3}{\varepsilon}\right)\left(\rho_\epsilon-\rho_\varepsilon^1\right)dx\right.\\
&\left.+\varepsilon\int_{\Omega_\varepsilon^+} \frac{\partial^3 w^-}{\partial x_1^2 \partial x_3}(x_1,l_3)\,\psi^+\left(\frac{x_1}{\varepsilon},\frac{x_3-l_3}{\varepsilon}\right)\left(\rho_\epsilon-\rho_\epsilon^2\right)dx\right|\\
&\leq C\varepsilon\left(\left\|\rho_\epsilon-\rho_\epsilon^1\right\|_{L^2(\Omega^-)}+\left\|\rho_\epsilon-\rho_\epsilon^2\right\|_{L^2(\Omega_\varepsilon^+)}\right)\\
&\leq C\varepsilon\left(\left\|\nabla\left(\rho_\epsilon-\rho_\epsilon^1\right)\right\|_{(L^2(\Omega^-))^2}+\left\|\nabla\left(\rho_\epsilon-\rho_\epsilon^2\right)\right\|_{\left(L^2(\Omega_\varepsilon^+)\right)^2}\right)\\
&\leq C\varepsilon\left(\|\nabla\rho_\epsilon\|_{(L^2(\Omega^-))^2}+\left\|\nabla\rho_\epsilon^1\right\|_{(L^2(\Omega^-))^2}+\|\nabla\rho_\epsilon\|_{\left(L^2(\Omega_\varepsilon^+)\right)^2}+\left\|\nabla\rho_\epsilon^2\right\|_{\left(L^2(\Omega_\varepsilon^+)\right)^2}\right)\\
&\leq C\left(\varepsilon\,\|\nabla\rho_\epsilon\|_{(L^2(\Omega_\varepsilon))^2}+\varepsilon^2\right)
\end{aligned}$$

for ε small enough. Integrating by parts the fourth and sixth terms on the right-hand side of (4.25), it follows that

$$\begin{aligned}
&2\varepsilon\int_{\Omega^-} \frac{\partial^2 w^-}{\partial x_1 \partial x_3}(x_1,l_3)\,\frac{\partial}{x_1}\left(\psi^-\left(\frac{x_1}{\varepsilon},\frac{x_3-l_3}{\varepsilon}\right)\right)\left(\rho_\epsilon-\rho_\varepsilon^1\right)dx\\
&+2\varepsilon\int_{\Omega_\varepsilon^+} \frac{\partial^2 w^-}{\partial x_1 \partial x_3}(x_1,l_3)\,\frac{\partial}{x_1}\left(\psi^+\left(\frac{x_1}{\varepsilon},\frac{x_3-l_3}{\varepsilon}\right)\right)\left(\rho_\epsilon-\rho_\epsilon^2\right)dx\\
&=\quad -2\varepsilon\int_{\Omega^-} \frac{\partial^2 w^-}{\partial x_1 \partial x_3}(x_1,l_3)\,\psi^-\left(\frac{x_1}{\varepsilon},\frac{x_3-l_3}{\varepsilon}\right)\frac{\partial}{x_1}\left(\rho_\epsilon-\rho_\varepsilon^1\right)dx\\
&\qquad -2\varepsilon\int_{\Omega^-} \frac{\partial^3 w^-}{\partial x_1^2 \partial x_3}(x_1,l_3)\,\psi^-\left(\frac{x_1}{\varepsilon},\frac{x_3-l_3}{\varepsilon}\right)\left(\rho_\epsilon-\rho_\varepsilon^1\right)dx\\
&\qquad -2\varepsilon\int_{\Omega_\varepsilon^+} \frac{\partial^2 w^-}{\partial x_1 \partial x_3}(x_1,l_3)\,\psi^+\left(\frac{x_1}{\varepsilon},\frac{x_3-l_3}{\varepsilon}\right)\frac{\partial}{x_1}\left(\rho_\epsilon-\rho_\varepsilon^2\right)dx\\
&\qquad -2\varepsilon\int_{\Omega_\varepsilon^+} \frac{\partial^3 w^-}{\partial x_1^2 \partial x_3}(x_1,l_3)\,\psi^+\left(\frac{x_1}{\varepsilon},\frac{x_3-l_3}{\varepsilon}\right)\left(\rho_\epsilon-\rho_\varepsilon^2\right)dx.
\end{aligned}$$

Then, Corollary 3.2, the Cauchy–Schwarz inequality, the Poincaré inequality, (4.5),

(4.26), and (4.27) imply

(4.30)

$$
\begin{aligned}
&\left|2\varepsilon \int_{\Omega^-} \frac{\partial^2 w^-}{\partial x_1 \partial x_3}(x_1, l_3)\, \frac{\partial}{x_1}\left(\psi^-\left(\frac{x_1}{\varepsilon}, \frac{x_3 - l_3}{\varepsilon}\right)\right)\left(\rho_\epsilon - \rho_\varepsilon^1\right) dx\right. \\
&\qquad \left. + 2\varepsilon \int_{\Omega_\varepsilon^+} \frac{\partial^2 w^-}{\partial x_1 \partial x_3}(x_1, l_3)\, \frac{\partial}{x_1}\left(\psi^+\left(\frac{x_1}{\varepsilon}, \frac{x_3 - l_3}{\varepsilon}\right)\right)\left(\rho_\epsilon - \rho_\epsilon^2\right) dx\right| \\
&\le C\varepsilon \left(\left\|\nabla\left(\rho_\epsilon - \rho_\epsilon^1\right)\right\|_{(L^2(\Omega^-))^2} + \left\|\rho_\epsilon - \rho_\epsilon^1\right\|_{L^2(\Omega^-)}\right. \\
&\qquad \left. + \left\|\nabla\left(\rho_\epsilon - \rho_\epsilon^2\right)\right\|_{\left(L^2(\Omega_\varepsilon^+)\right)^2} + \left\|\rho_\epsilon - \rho_\epsilon^2\right\|_{L^2(\Omega_\varepsilon^+)}\right) \\
&\le C\varepsilon \left(\left\|\nabla\left(\rho_\epsilon - \rho_\epsilon^1\right)\right\|_{(L^2(\Omega^-))^2} + \left\|\nabla\left(\rho_\epsilon - \rho_\epsilon^2\right)\right\|_{\left(L^2(\Omega_\varepsilon^+)\right)^2}\right) \\
&\le C\varepsilon \left(\|\nabla\rho_\epsilon\|_{(L^2(\Omega^-))^2} + \left\|\nabla\rho_\epsilon^1\right\|_{(L^2(\Omega^-))^2} + \|\nabla\rho_\epsilon\|_{\left(L^2(\Omega_\varepsilon^+)\right)^2} + \left\|\nabla\rho_\epsilon^2\right\|_{\left(L^2(\Omega_\varepsilon^+)\right)^2}\right) \\
&\le C\left(\varepsilon \|\nabla\rho_\epsilon\|_{(L^2(\Omega_\varepsilon))^2} + \varepsilon^2\right)
\end{aligned}
$$

for ε enough small. Combining (4.25) with (4.28) $\div$ (4.30), we obtain

$$\|\nabla\rho_\epsilon\|^2_{(L^2(\Omega_\varepsilon))^2} \le C\left(\varepsilon \|\nabla\rho_\epsilon\|_{(L^2(\Omega_\varepsilon))^2} + \varepsilon^2\right)$$

for ε small enough. Therefore

$$\|\nabla\rho_\epsilon\|_{(L^2(\Omega_\varepsilon))^2} \le C\varepsilon \tag{4.31}$$

for ε small enough. Finally, making use of the Poincaré inequality, estimate (4.13) follows from (4.31). □

Proof of Theorem 4.1. Let $\psi^\pm$ be the functions satisfying (3.1), (3.2), and τ_ε and ρ_ε be the functions defined in (4.8) and (4.9), respectively. Since

$$u_\epsilon - u - \varepsilon w = \tau_\varepsilon + \varepsilon\rho_\varepsilon + g_\varepsilon \quad \text{in } \Omega_\varepsilon,$$

where

$$
g_\varepsilon = \begin{cases}
\varepsilon \dfrac{\partial u^-}{\partial x_3}(x_1, l_3)\, \psi^+\left(\dfrac{x_1}{\varepsilon}, \dfrac{x_3 - l_3}{\varepsilon}\right) \\
+ \varepsilon^2 \dfrac{\partial w^-}{\partial x_3}(x_1, l_3)\, \psi^+\left(\dfrac{x_1}{\varepsilon}, \dfrac{x_3 - l_3}{\varepsilon}\right) & \text{in } \Omega_\varepsilon^+, \\
\varepsilon \dfrac{\partial u^-}{\partial x_3}(x_1, l_3)\left(\psi^-\left(\dfrac{x_1}{\varepsilon}, \dfrac{x_3 - l_3}{\varepsilon}\right) - \beta\right) \\
+ \varepsilon^2 \dfrac{\partial w^-}{\partial x_3}(x_1, l_3)\, \psi^-\left(\dfrac{x_1}{\varepsilon}, \dfrac{x_3 - l_3}{\varepsilon}\right) & \text{in } \Omega^-,
\end{cases}
$$

estimates (4.6) and (4.7) follow from Propositions 4.2 and 4.3, Corollary 3.2, and estimates (4.2) and (4.5). □

REFERENCES

[1] Y. Achdou and O. Pironneau, *Domain decomposition and wall laws*, C. R. Acad. Sci. Paris Sér. I Math., 320 (1995), pp. 541–547.

[2] Y. Achdou, O. Pironneau, and F. Valentin, *Effective boundary conditions for laminar flows over rough boundaries*, J. Comput. Phys., 147 (1998), pp. 187–218.

[3] G. Allaire and M. Amar, *Boundary layer tails in periodic homogenization*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 209–243.

[4] Y. Amirat and O. Bodart, *Boundary layer correctors for the solution of Laplace equation in a domain with oscillating boundary*, Z. Anal. Anwendungen, 20 (2001), pp. 929–940.

[5] Y. Amirat, D. Bresch, J. Lemoine, and J. Simon, *Effect of rugosity on a flow governed by Navier-Stokes equations*, Quart. Appl. Math., 59 (2001), pp. 769–785.

[6] H. Attouch, *Variational Convergence for Functions and Operators*, Applicable Mathematics Series, Pitman, London, 1984.

[7] N. Bakhvalov and G. Panasenko, *Homogenisation: Averaging Processes in Periodic Media*, Math. Appl. (Soviet Ser.) 36, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1989.

[8] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North–Holland, Amsterdam, 1978.

[9] D. Blanchard, L. Carbone, and A. Gaudiello, *Homogenization of a monotone problem in a domain with oscillating boundary*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1057–1070.

[10] A. Braides and A. Defranceshi, *Homogenization of Multiple Integrals*, Oxford Lecture Ser. Math. Appl. 12, Clarendon Press, Oxford University Press, New York, 1998.

[11] R. Brizzi and J. P. Chalot, *Boundary homogenization and Neumann boundary value problem*, Ricerche Mat., 46 (1997), pp. 341–387.

[12] G. A. Chechkin, A. Friedman, and A. L. Piatniski, *The boundary value problem in a domain with very rapidly oscillating boundary*, J. Math. Anal. Appl., 231 (1999), pp. 213–234.

[13] D. Cioranescu and P. Donato, *An Introduction to Homogenization*, Oxford Lecture Ser. Math. Appl. 17, Clarendon Press, Oxford University Press, New York, 1999.

[14] D. Cioranescu and J. Saint Jean Paulin, *Homogenization of Reticuled Structures*, Appl. Math. Sc. 136, Springer-Verlag, New York, 1999.

[15] A. Corbo Esposito, P. Donato, A. Gaudiello, and C. Picard, *Homogenization of the p-Laplacian in a domain with oscillating boundary*, Comm. Appl. Nonlinear Anal., 4 (1997), pp. 1–23.

[16] A. Gaudiello, *Asymptotic behaviour of non-homogeneous Neumann problems in domains with oscillating boundary*, Ricerche Mat., 43 (1994), pp. 239–292.

[17] A. Gaudiello, *Homogenization of an elliptic transmission problem*, Adv. Math. Sci. Appl., 5 (1995), pp. 639–657.

[18] A. Gaudiello, R. Hadiji, and C. Picard, *Homogenization of the Ginzburg-Landau equation in a domain with oscillating boundary*, Commun. Appl. Anal., 7 (2003), pp. 209–223.

[19] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of the Second Order*, Springer-Verlag, Berlin, 1977.

[20] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, London, 1985.

[21] E. Ja. Hruslov, *The method of orthogonal projections and the Dirichlet problem in domains with a fine-grained boundary*, Mat. Sb. (N.S.), 88 (1972), pp. 37–59.

[22] W. Jäger and A. Mikelić, *Homogenization of the Laplace equation in a partially perforated domain*, in Homogenization, Ser. Adv. Math. Appl. Sci. 50, World Scientific, River Edge, NJ, 1999, pp. 259–284.

[23] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994.

[24] E. M. Landis and G. P. Panasenko, *A theorem of the asymptotics of solutions of elliptic equations with coefficients periodic in all variables except one*, Soviet Math. Dokl., 18 (1977), pp. 1140–1143.

[25] J. L. Lions, *Some Methods in the Mathematical Analysis of Systems and their Control*, Kexue Chubanshe (Science Press), Beijing, Gordon and Breach Science Publishers, New York, 1981.

[26] V. A. Marčenko and E. Ja. Hruslov, *Boundary Value Problems in Domains with a Fine-Grained Boundary*, Naukova Dumka, Kiev, 1974.

[27] T. A. Mel'nyk, *Homogenization of the Poisson equation in a thick periodic junction*, Z. Anal. Anwendungen, 18 (1999), pp. 953–975.

[28] T. A. Mel'nyk and S. A. Nazarov, *The asymptotic structure of the spectrum in the problem*

of harmonic oscillations in a hub with heavy spokes, Russ. Acad. Sci. Dokl. Math, 48 (1994), pp. 428–432.

[29] J. Nevard and J. B. Keller, *Homogenization of rough boundaries and interfaces*, SIAM J. Appl. Math., 57 (1997), pp. 1660–1686.

[30] O. Pironneau, *Optimal Shape Design for Elliptic Systems*, Springer Ser. Comput. Phys., Springer, New York, 1984.

[31] O. Pironneau and C. Saguez, *Asymptotic Behavior with Respect to the Domain of Solutions of PDE*, Laboria Report 218, 1977.

[32] E. Sánchez-Palencia, *Nonhomogeneous Media and Vibration Theory*, Lecture Notes in Phys., 127, Springer-Verlag, Berlin, New York, 1980.

[33] L. Tartar, *Cours Peccot, Collège de France (March* 1977*)*, partially written in F. Murat, H-Convergence, Séminaire d'analyse fonctionnelle et numérique de l'Université d'Alger (1977–78). English translation in Mathematical Modeling of Composite Materials, A. Cherkaev and R. V. Kohn, eds., Progr. Nonlinear Differential Equations Appl. 31, Birkhäuser-Verlag, Boston, 1997, pp. 21–44.

SIAM J. MATH. ANAL.
Vol. 35, No. 6, pp. 1617–1635

THE DYNAMICS OF A PLANE DIODE*

YAN GUO†, CHI-WANG SHU†, AND TIE ZHOU‡

Abstract. The dynamics of a plane diode is described by the Vlasov–Poisson system over an interval with inflow boundary conditions at two ends. In this article, the uniqueness and regularity of such dynamics are investigated. It is shown that a rather general initial and boundary datum leads to a unique solution with bounded variations (BV). Moreover, such a solution becomes discontinuous if the external voltage is large enough, while it can remain C^1 if the external voltage is sufficiently small or absent.

Key words. Vlasov–Poisson system, boundary problem, outward pointing electric field, BV estimate, Harten's lemma, C^1 solution

AMS subject classification. 35Q72

DOI. 10.1137/S0036141003421133

1. Introduction and notation.

The construction of particle accelerators and free electron lasers requires electron guns which produce relativistic electron beams of high quality (low emission and high current). Numerical simulations of various sophisticated models are used in the design of electron guns. The most complete mathematical model is the boundary value problem for the Vlasov–Maxwell evolution equations in a geometrically complicated spatially three-dimensional domain, for which global weak solutions were constructed in [8]. The boundary in such a problem is always characteristic, so the question of uniqueness and regularity of the solution in the presence of a boundary is very challenging mathematically. In this paper, we study the one-dimensional Vlasov–Poisson equation over an interval, which is the evolutional model for classical electronic conduction in a plane diode. Even in this simplest case, the uniqueness and regularity of such dynamics have been open.

Let a dilute electron gas be emitted at $x = 0$ and absorbed at $x = 1$. Under an external voltage, the dynamics of such a plane diode is modeled by the Vlasov–Poisson system [7] as follows:

$$\partial_t f + v\partial_x f + \partial_x \phi \partial_v f = 0, \tag{1.1}$$

$$\partial_{xx}\phi(t,x) = \rho(t,x), \tag{1.2}$$

$$f(0,x,v) = f_0(x,v), \tag{1.3}$$

where the macroscopic charged density $\rho(t,x)$ and related current density $j(t,x)$ are given by

$$\rho(t,v) = \int_{-\infty}^{\infty} f(t,x,v)dv, \qquad j(t,x) = \int_{-\infty}^{\infty} vf(t,x,v)dv. \tag{1.4}$$

*Received by the editors January 9, 2003; accepted for publication September 5, 2003; published electronically April 7, 2004.
http://www.siam.org/journals/sima/35-6/42113.html

†Division of Applied Mathematics, Brown University, Providence, RI 02912 (guoy@dam.brown.edu, shu@dam.brown.edu). The research of the first author was supported in part by an A.P. Sloan Fellowship. The research of the second author was supported by ARO grant DAAD19-00-1-0405, NSF grant DMS-0207451, NASA Langley grant NCC1-01035, and AFOSR grant F49620-02-1-0113.

‡LMAM, School of Mathematical Sciences, Peking University, Beijing 100871, China (tzhou@math.pku.edu.cn). The research of this author was supported by the China State Major Key Project for Basic Research (G1999032803).

Here the boundary conditions for the electron distribution $f(t,x,v) \geq 0$ are given by

$$f(t,0,v) = g(t,v), \quad v > 0; \qquad f(t,1,v) \equiv 0, \quad v < 0; \tag{1.5}$$

and external voltages are given at $x = 0$ and $x = 1$:

$$\phi(t,0) = 0, \quad \phi(t,1) = \lambda(t) \geq 0. \tag{1.6}$$

Mathematical study of such nonlinear boundary value problems was initiated in the pioneering work in [7], in which stationary (independent of time t) solutions are constructed. Such stationary solutions are generically not continuous. A higher dimensional generalization was given in [13] and [6]. On the other hand, for the dynamical problem of such a plane diode (1.1)–(1.6), a weak solution can be constructed as in [3] and [1] (see also the related works in [2], [4], and [5]).

In this article, we show that under rather general conditions on the initial and boundary data $f_0(x,v)$ as well as $g(t,v)$, for every fixed t, the solution $f(t,x,v)$ is a bounded variation (BV) function of x and v. This implies its uniqueness for the nonlinear plane diode (Theorem 4.2). Moreover, given smooth initial and boundary electron distributions, we investigate the influence of the voltage λ on the smoothness of the solutions $f(t,x,v)$. We demonstrate that $f(t,x,v)$ can be discontinuous if the voltage $\lambda(0)$ is large (Lemma 4.3). In the absence of the external voltage (i.e., $\lambda(t) \equiv 0$), the solutions always remain C^1 inside (Theorem 4.4). Furthermore, if $g(t,v)$ is nonvanishing and the voltage $\lambda(t)$ is sufficiently small, the solution also remains C^1 inside (Theorem 4.5). Our results are in interesting contrast to the stationary case.

In general, as in the stationary case, it is expected that singularity (discontinuity) should develop at the boundary and then propagate into the region. This kind of singularity was first characterized by a BV estimate in [9] for a half line problem, via a rather complicated procedure, for which a delicate damping term was introduced. Such a BV regularity is crucial in proving the uniqueness for the full nonlinear problem, at least in the case of one space dimension. Somewhat surprisingly, unlike in the stationary case, it was discovered in [9] that the sign of the electric field $E(t,x) \equiv \partial_x \phi$ at the boundary can ensure the smoothness of the solution of a dynamical problem. Both BV and C^1 estimates in [9] are very delicate and depend on a so-called velocity lemma, which is the only tool so far to estimate the bouncing particles reflected specularly at the boundary. In general, solutions to the Vlasov–Poisson system can be classical if the electric field $E(t,\mathbf{x})$ (for the multidimensional case $E(t,\mathbf{x}) \equiv \nabla_{\mathbf{x}} \phi$) points strictly outward at the boundary. In the absence of an external voltage, such an outward condition is automatically true for a nonvanishing electron gas, which satisfies the nonlinear Vlasov–Poisson system. Classical solutions can therefore be constructed in [9] and [10], in a half space. Recently, regularity for the full Vlasov–Poisson system in a smooth three-dimensional convex domain was obtained in [12]. In the presence of an external potential $\lambda(t)$, however, an outward condition for $\partial_x \phi$ is not true in general. A nonvanishing g and a small voltage $\lambda(t)$ guarantee that the electric field $\partial_x \phi$ always points outward at both $x = 0$ and $x = 1$ for the nonlinear problem (1.1)–(1.6).

The main mathematical difficulty in this article lies in the linear analysis with a given, external electric field $E(t,x)$. The novelty of the C^1 estimate here extends the previous results to treat the case if $E(t,x)$ points outward at $x = 0$ and $x = 1$, but *not* strictly. This case has been left open in [9], which was out of reach of the velocity lemma. A new regularity estimate is established in Theorem 2.4, via Lemma 2.3, a weak version of the velocity lemma (see also [12]). This directly leads to Theorem 4.4

for the nonlinear case naturally, in the absence of an external voltage. On the other hand, based on Lemma 2.1 with an additional nonvanishing assumption for f_0 and g, an explicit upper bound (4.15) is given in Theorem 4.5 for the external voltage to guarantee the regularity of f. As for the BV estimate, we use a more direct approach of constructing a positive finite difference scheme. We introduce an error function to absorb the complicated boundary contribution. By using Harten's lemma [11], [14], we are able to give a much simpler and straightforward BV estimate (Theorem 3.4). We expect that our method here is very general to treat more complicated domains in higher dimensions.

Notation. We now introduce some notation. Let T be an arbitrary positive constant. We denote the region

$$\Pi = [0,T]\times[0,1]\times\mathbf{R}, \quad \Pi_s = \Pi\cap\{t=s\}, \quad 0\le s\le T,$$

and the incoming sets at the boundaries $\{x=0\}$ and $\{x=1\}$ as

$$\gamma_0^+ = \{(t,0,v)\mid v>0,\ 0\le t\le T\}, \qquad \gamma_1^- = \{(t,1,v)\mid v<0,\ 0\le t\le T\}.$$

We denote the singular set

$$\gamma_S \equiv \{x=0,\ v=0\}\cup\{x=1,\ v=0\}.$$

We use $\|\cdot\|_p$ to denote the standard L^p norms for $1\le p\le\infty$, and let $C^{0,1}=W^{1,\infty}$ be the space of Lipschitzly continuous functions. Let $\boldsymbol{\partial}=[\partial_t,\partial_x,\partial_v]$, and let

$$\begin{aligned}|||[f_0,g]||| \equiv{}& \|f_0\|_{C^1}+\|g\|_{C^1}+\|v\partial_x f_0\|_\infty+\|v\partial_v f_0\|_\infty\\ &+\|v^{-1}\partial_t g\|_\infty+\|v^{-1}\partial_v g\|_\infty.\end{aligned} \tag{1.7}$$

The last two terms above are needed for our regularity analysis.

2. Linear C^1 estimate. We consider the following linear problem with a given, external electric field $E(t,x)$:

$$\begin{aligned}&\partial_t f + v\partial_x f + E(t,x)\partial_v f = 0,\\ &f|_{\gamma_0^+} = g, \qquad f|_{\gamma_1^-}=0,\\ &f(0,x,v)=f_0(x,v).\end{aligned} \tag{2.1}$$

For any point $(t,x,v)\in\Pi$, we define $\Gamma(\tau;t,x,v)=(\tau,X(\tau;t,x,v),V(\tau;t,x,v))$ to be the unique trajectory of

$$\frac{dX(\tau)}{d\tau}=V(\tau), \qquad \frac{dV(\tau)}{d\tau}=E(\tau,X(\tau)), \tag{2.2}$$

such that $X(t;t,x,v)=x$, $V(t;t,x,v)=v$. Equivalently,

$$\begin{aligned}V(s)&=v+\int_t^s E(\tau,X(\tau))d\tau,\\ X(s)&=x-v(t-s)+\int_t^s\int_t^{t_1}E(\tau,X(\tau))d\tau dt_1.\end{aligned} \tag{2.3}$$

We define the *starting point* of (t,x,v), $(t_0(t,x,v),x_0(t,x,v),v_0(t,x,v))$, to be the (unique) first point at $\partial\Pi$ on the backward-in-time trajectory (2.2). We begin with two basic facts for such a starting point (t_0,x_0,v_0).

LEMMA 2.1. *Let $E \in C^{0,1}([0,T]\times[0,1])$. Let $(t,x,v)\in\Pi$ with $v>0$ and $2x\|E\|_\infty < v^2$. Then the starting point $(t_0(t,x,v), x_0(t,x,v), v_0(t,x,v))$ satisfies*

$$0 \le x_0 \le x, \tag{2.4}$$

$$0 \le t - t_0 \le \frac{v - \sqrt{v^2 - 2\|E\|_\infty x}}{\|E\|_\infty}, \tag{2.5}$$

$$0 \le v_0 \le v + \sqrt{v^2 - 2\|E\|_\infty x}. \tag{2.6}$$

Proof. We construct a $C^{0,1}$ extension $\bar{E}(t,x)$ of $E(t,x)$ to $t\le 0$ and $x\le 0$ so that $\|\bar{E}\|_\infty = \|E\|_\infty$.

Now we fix $(t,x,v)\in\Pi$ with $v>0$ and $2x\|E\|_\infty < v^2$. Let the trajectory $\Gamma(\tau;t,x,v)$ satisfy (2.2) with E replaced by $\bar{E}$ such that

$$\begin{aligned} V(s) &= v + \int_t^s \bar{E}(\tau, X(\tau))d\tau, \\ X(s) &= x - v(t-s) + \int_t^s\int_t^{t_1} \bar{E}(\tau, X(\tau))d\tau dt_1, \end{aligned} \tag{2.7}$$

where $X(\tau) = X(\tau;t,x,v)$. Notice that for

$$t - s \le \frac{v - \sqrt{v^2 - 2\|E\|_\infty x}}{\|E\|_\infty},$$

we have

$$\left|\int_t^s \bar{E}(\tau, X(\tau))d\tau\right| \le \|E\|_\infty \times \frac{v - \sqrt{v^2 - 2\|E\|_\infty x}}{\|E\|_\infty} < v.$$

Therefore, $V(s) > 0$ and $X(s) \le x$. To show (2.5), we look for a root θ^* of the function

$$X(\theta;t,x,v) = x - v(t-\theta) - \int_t^\theta\int_t^{t_1} \bar{E}(\tau, X(\tau))d\tau dt_1. \tag{2.8}$$

Clearly, $X(t) = X(t;t,x,v) = x > 0$. Define a quadratic function

$$X_1(\theta) = x - v(t-\theta) + \frac{\|E\|_\infty}{2}(t-\theta)^2,$$

so that $X(\theta)\le X_1(\theta)$. In the case $2\|E\|_\infty x < v^2$, an elementary analysis on the roots of $X_1(\theta)$ shows that for $t-\theta$ greater than but arbitrarily close to its first positive root

$$\frac{v - \sqrt{v^2 - 2\|E\|_\infty x}}{\|E\|_\infty},$$

we have

$$X(\theta) \le X_1(\theta) < 0.$$

By the continuity of $X(\theta)$, this implies that there exists a root θ^* of $X(\theta)$ which satisfies

$$t - \frac{v - \sqrt{v^2 - 2\|E\|_\infty x}}{\|E\|_\infty} < \theta^* < t.$$

But $X(\theta)$ is strictly increasing over such an interval since $X'(\theta) = V(\theta) > 0$; hence θ^* is unique. Because $X(s) \geq 0$, we recover the original (2.3) with $\bar{E} \equiv E$.

Now if $\theta^* > 0$, then $t_0 = \theta^*$, which leads to (2.5) and (2.6) with $x_0 \geq 0$. On the other hand, if $\theta^* = 0$, then $t_0 = 0$, and (2.5) and (2.6) with $x_0 \geq 0$ are again valid. ∎

LEMMA 2.2. *Let $E \in C^1([0,T] \times [0,1])$. Fix $(t,x,v) \in \Pi$ such that $v_0(t,x,v) > 0$. Then $[t_0, v_0]$ are C^1 functions near (t,x,v). Let $\boldsymbol{\partial} t_0 = [\partial_t t_0, \partial_x t_0, \partial_v t_0]$; then $v_0 \boldsymbol{\partial} t_0$ is*

$$\begin{bmatrix} v_0 + E(t,x)(t_0 - t) + \displaystyle\int_{t_0}^{t} E + \int_{t_0}^{t}\int_{t}^{s} \partial_x E \partial_t X, \\ t - t_0 + \displaystyle\int_{t_0}^{t}\int_{t}^{s} \partial_x E \partial_v X, \\ -1 + \displaystyle\int_{t_0}^{t}\int_{t}^{s} \partial_x E \partial_x X \end{bmatrix}. \tag{2.9}$$

Proof. We now consider a C^1 extension $\bar{E}$ of E so that $\|\bar{E}\|_{C^1} \leq C\|E\|_{C^1}$. We still consider (2.7) with such a new $\bar{E}$. We again look for a root for $X(\theta; t, x, v)$ in (2.8). Notice that by our assumption,

$$\left.\frac{\partial X(\theta;t,x,v)}{\partial \theta}\right|_{\theta = t_0} = v + \int_{t}^{t_0} \bar{E}(\tau) d\tau = v_0 > 0.$$

By the implicit function theorem, there is a unique C^1 function $\theta(t,x,v)$ locally, which satisfies $X(\theta,t,x,v) = 0$. We now claim that $\theta(t',x',v') = t_0(t',x',v')$ for (t',x',v') near (t,x,v). It suffices to show that

$$X(s;t',x',v') > 0$$

for $\theta(t',x',v') < s \leq t'$, for (t',x',v') near (t,x,v), for this implies $\bar{E} = E$ in (2.3). In fact, since $\theta(t',x',v') \in C^1$ and $\frac{\partial X}{\partial \theta} > 0$ at $\theta = t_0$, it follows that there is a $\delta > 0$, such that

$$X(s;t',x',v') > 0$$

for $\theta < s < \theta + \delta$. Since $X(s;t,x,v) \geq \epsilon_0 > 0$ for $t_0 + \delta/2 \leq s \leq t$, it follows that by further choosing (t',x',v') closer to (t,x,v),

$$X(s;t',x',v') \geq \epsilon_0/2 > 0$$

for $\theta + \delta \leq s \leq t$. Therefore $X(s;t',x',v') > 0$ for $\theta(t',x',v') < s \leq t'$ and $\theta(t',x',v') = t_0(t',x',v')$. Furthermore, (2.9) follows from differentiating (2.3) with $\bar{E} = E$. ∎

We now establish a weak version of the velocity lemma (see [9]), with E pointing outward at $x = 0$ and $x = 1$, but not strictly. On the other hand, no explicit estimate of (x_0, v_0) in terms of (x,v) can be given, as in the previous velocity lemma [9].

LEMMA 2.3. *Let $E \in C^{0,1}([0,T] \times [0,1])$. Assume $E(t,0) \leq 0$ for all $0 \leq t \leq T$. For any $(t,x,v) \in \Pi$, the following hold:*

(a) *If its characteristic $\Gamma(\tau;t,x,v) \in \Pi$ passes through $(t^*,0,0)$ for some $t^* \geq 0$, then $x = 0$ and $v = 0$.*

(b) *Moreover, let (t_0,x_0,v_0) be the starting point of (t,x,v); then*

$$\lim_{(x,v)\to(0,0)} (|v_0(t,x,v)| + |x_0(t,x,v)|) = 0. \tag{2.10}$$

Proof. We first define a constant extension $\bar{E}(t,x)$ of the electric field $E(t,x)$ to $x \le 0$ as $\bar{E}(t,x) \equiv E(t,0)$. Clearly, $\bar{E}(t,x) \in C^{0,1}$.

To prove (a), fix (t,x,v) and consider the trajectory $(\tau, X(\tau;t^*,0,0), V(\tau;t^*,0,0))$ emanating from $(t^*,0,0)$. Notice that such a curve also satisfies

$$\frac{dX(\tau)}{d\tau} = V(\tau), \qquad \frac{dV(\tau)}{d\tau} = \bar{E}(\tau, X(\tau)),$$

or, equivalently,

$$V(s) = \int_{t^*}^{t} \bar{E}(\tau, X(\tau)) d\tau ds, \qquad X(s) = \int_{t^*}^{s}\int_{t^*}^{t_1} \bar{E}(\tau, X(\tau)) d\tau ds t_1. \tag{2.11}$$

Since (t,x,v) is on the curve, it follows that $X(\tau) \ge 0$ between t and t^* and

$$V(t) = v, \qquad X(t) = x.$$

On the other hand, for any $\epsilon > 0$, we define an approximate field

$$E_\epsilon \equiv \bar{E} - \epsilon$$

and denote the approximate curve $(\tau, X_\epsilon(\tau;t^*,0,0), V_\epsilon(\tau;t^*,0,0))$ to satisfy

$$\frac{dX_\epsilon(\tau)}{d\tau} = V_\epsilon(\tau), \qquad \frac{dV_\epsilon(\tau)}{d\tau} = \bar{E}_\epsilon(\tau, X_\epsilon(\tau))$$

or

$$V_\epsilon(t) = \int_{t^*}^{t} \bar{E}_\epsilon(\tau, X_\epsilon(\tau)) d\tau, \qquad X_\epsilon(t) = \int_{t^*}^{t}\int_{t^*}^{s} \bar{E}_\epsilon(\tau, X_\epsilon(\tau)) d\tau ds. \tag{2.12}$$

Since $E_\epsilon(\tau, y) \le -\epsilon < 0$ for all $y \le 0$ and $0 \le \tau \le T$, it follows from (2.12) that

$$V_\epsilon(\tau) \le 0 \quad \text{for } \tau \ge t^*, \qquad V_\epsilon(\tau) \ge 0 \quad \text{for } \tau \le t^*.$$

Therefore, for all $0 \le \tau \le T$, we have

$$X_\epsilon(\tau) \le 0.$$

Since $E_\epsilon \to \bar{E}$ as $\epsilon \to 0$, we have $V_\epsilon(\tau) \to V(\tau)$ and $X_\epsilon(\tau) \to X(\tau)$. In particular,

$$X(\tau) \le 0$$

for all τ between t^* and t. Therefore $X(\tau) \equiv 0$. We deduce that either $t = t^*$ or

$$E(\tau, 0) \equiv 0$$

between t and t^* in (2.11). Both cases imply that $x = v = 0$. We therefore conclude (a).

We now turn to part (b). If (2.10) were false, there would exist (t, x_n, v_n) such that $|x_n| + |v_n| \to 0$, but their corresponding starting points would satisfy

$$|v_0(t, x_n, v_n)| + |x_0(t, x_n, v_n)| \ge \delta > 0.$$

Up to a subsequence, let $t_0(t, x_n, v_n) \to t_0^*$, $v_0(t, x_n, v_n) \to v_0^*$, and $x_0(t_n, x_n, v_n) \to x_0^*$ such that

$$|v_0^*| + x_0^* \ge \delta > 0.$$

We deduce that $(t,0,0)$ connects with (t_0^*, x_0^*, v_0^*) through a characteristic curve. Therefore $v_0^* = x_0^* = 0$ by part (a), which is a contradiction. ∎

We are now ready to establish the C^1 estimate.

THEOREM 2.4. *Let $E(t,x)$ be $C^1([0,T]\times[0,1])$ and point outward at $x=0$ and $x=1$; that is,*

$$E(t,0) \le 0 \qquad \text{and} \qquad E(t,1) \ge 0. \tag{2.13}$$

Let f_0 and g be C^1. Assume the following compatibility conditions are valid:

$$\begin{aligned} &f_0(0,v) = g(0,v) \quad \text{for } v>0, \qquad f_0(1,v) = 0 \quad \text{for } v<0,\\ &\partial_t g(0,v) + v\partial_x f_0(0,v) + E(0,0)\partial_v f_0(0,v) = 0 \quad \text{for } v>0,\\ &v\partial_x f_0(1,v) + E(0,1)\partial_v f_0(1,v) = 0 \quad \text{for } v<0. \end{aligned} \tag{2.14}$$

Then the solution $f(t,x,v)$ to (2.1) belongs to $C^1(\Pi\backslash\gamma_S)\cap C^0(\Pi)$. Furthermore, there is a numerical constant $C>0$ such that

$$\|f\|_{C^{0,1}(\Pi)} \le Ce^{T\|E\|_{C^1}} |||[f_0,g]|||, \tag{2.15}$$

where $|||[f_0,g]|||$ is defined in (1.7).

Proof. For any $(t,x,v)\in\Pi\backslash\gamma_S$, we consider its backward trajectory $\Gamma(\tau;t,x,v)$ to (2.2) with respect to the external field $E(t,x)$ for $0\le\tau\le t$. Let (t_0,x_0,v_0) be its starting point. We first consider only the case of $0\le x_0<1$. The case $x_0=1$ at the right boundary will be treated via a reflection at the end of the proof.

(a) If $0<x_0<1$ so that $t_0=0$, we define

$$f(t,x,v) = f(t_0,x_0,v_0) = f_0(X(0;t,x,v), V(0;t,x,v)). \tag{2.16}$$

By standard ODE theory, we easily deduce that $f\in C^1$ near such a point (t,x,v). Moreover, differentiating with respect to (t,x,v) of (2.2) yields

$$|\partial X| + |\partial V| \le Ce^{T\|E\|_{C^1}} \tag{2.17}$$

for $\partial = [\partial_t, \partial_x, \partial_v]$. Hence (2.15) follows directly from (2.16).

(b) If $x_0=0$, we now separate two cases: If $t_0>0$, then by (2.3)

$$f(t,x,v) = f(t_0,x_0,v_0) = g\left(t_0(t,x,v), v + \int_{t_0}^t E(\tau, X(\tau))d\tau\right). \tag{2.18}$$

By Lemma 2.3, $v_0>0$. Therefore, for $\partial = \partial_t$, ∂_x and ∂_v, we have

$$\partial f(t,x,v) = \partial t_0 \partial_t g(t_0,v_0) - \partial t_0 E(t_0,0)\partial_v g(t_0,v_0).$$

By Lemma 2.2, since $E\in C^1$, it follows from (2.17) that

$$|\partial t_0| \le Ce^{T\{\|E\|_{C^1}+1\}} v_0^{-1}.$$

Therefore, for $v_0>0$, $f\in C^1$ at (t,x,v), and

$$|f(t,x,v)| \le Ce^{T\{\|E\|_{C^1}+1\}} v_0^{-1}\{|\partial_t g(t_0,v_0)| + |\partial_v g(t_0,v_0)|\}.$$

Moreover, if $v>0$, then $v_0>0$ for $t-t_0$ small by (2.5) in Lemma 2.1, and $\partial f(t,x,v)$ is continuous across the boundary $\{x=0\}$.

On the other hand, if $x_0 = 0$ but $t_0 = 0$, to show f is C^1 at (t, x, v), which is on the surface of $(t, X(t; 0, 0, v_0), V(t; 0, 0, v_0))$, it suffices to prove that f is C^1 at $(0, 0, v_0)$ by using the C^1 mapping (2.2). We choose (t_1, x_1, v_1) and (t_2, x_2, v_2) which are close to $(0, 0, v_0)$ but satisfy different expressions, (2.16) and (2.18), respectively. Clearly, by $f_0(0, v_0) = g(0, v_0)$, $f \in C^0$ at $(0, 0, v_0)$. Let

$$\delta \equiv t_1 + t_2 + x_1 + x_2 + |v_1 - v_0| + |v_2 - v_0|.$$

Expanding (2.16) and (2.18) around $(0, 0, v_0)$ and $(0, 0, v_0)$, respectively, leads to

$$\begin{aligned}
&\partial f(t_1, x_1, v_1) = \partial_x f \partial X|_{(0,0,v_0)} + \partial_v f \partial V|_{(0,0,v_0)} + o(\delta),\\
&\partial f(t_2, x_2, v_2) = \partial_t g \partial t_0|_{(0,0,v_0)} + \partial_v g\{[E, 0, 1] - \partial t_0 E\}|_{(0,0,v_0)} + o(\delta).
\end{aligned}$$

Notice that $\partial X|_{(0,0,v_1)} = (v_0, 1, 0)$, and $\partial V|_{(0,0,v_0)} = (E(0, 0), 0, 1)$. On the other hand, by (2.9)

$$\partial t_0 = (1, -v_0^{-1}, 0)$$

for $\partial = [\partial_t, \partial_x, \partial_v]$. By using the compatibility condition (2.14), $f \in C^1$ at $(0, 0, v_0)$.

(c) We show that f is continuous at $x = 0, v = 0 \in \gamma_S$. We define

$$f(t, 0, 0) \equiv g(0, 0) = f_0(0, 0)$$

for $0 \le t \le T$. Choose any point (t, x, v) near γ_S and $(x, v) \ne (0, 0)$. Let (t_0, x_0, v_0) be its starting point so that

$$f(t, x, v) = f(t_0, x_0, v_0).$$

By $|||[f_0, g]||| < +\infty$, it follows that $\partial_t g(t, 0) \equiv 0$ so that

$$g(\tau, 0) = g(0, 0)$$

for all $0 \le \tau \le T$. By (2.13) and the weak velocity lemma (Lemma 2.3), it follows that $(x_0, v_0) \to (0, 0)$ so that either

$$\lim_{x+|v|\to 0} f(t, x, v) = \lim_{v_0 \to 0} g(t_0, v_0) \equiv g(0, 0) = f(t, 0, 0),$$

or

$$\lim_{x+|v|\to 0} f(t, x, v) = \lim_{x_0+|v_0|\to 0} f_0(x_0, v_0) = f_0(0, 0) = f(t, 0, 0).$$

Both imply f is continuous at $(t, 0, 0)$.

(d) The last case of $x_0 = 1$ can be reduced to the case of $x_0 = 0$ via a simple reflection. We define

$$\bar{f}(t, x, v) \equiv f(t, 1 - x, -v), \qquad \bar{E}(t, x) \equiv E(t, 1 - x). \tag{2.19}$$

The system (2.1) now becomes

$$\begin{aligned}
&\bar{f}_t + v\bar{f}_y + \bar{E}(t, y)\bar{f}_v = 0,\\
&\bar{f}(t, 0, v) = 0 \quad \text{at } \gamma_0^+, \qquad \bar{f}(t, 1, v) = g(t, v) \quad \text{at } \gamma_1^-,\\
&\bar{f}(0, y, v) = \bar{f}_0(y, v),
\end{aligned} \tag{2.20}$$

and its characteristic $\bar{\Gamma}(\tau;t,y,v)=(\tau,\bar{X}(\tau;t,y,v),\bar{V}(\tau;t,y,v))$ satisfies

$$\frac{dy}{dt}=v,\qquad \frac{dv}{dt}=-E(t,1-y).$$

Comparing with (2.2), we deduce $X(\tau;t,x,v)$, $V(\tau;t,x,v)$, and

$$1-\bar{X}(\tau;t,1-x,-v),\ -\bar{V}(\tau;t,1-x,-v)$$

satisfies the same ODE with the same initial condition (t,x,v). Therefore

$$\begin{aligned}X(\tau;t,x,v)&\equiv 1-\bar{X}(\tau;t,1-x,-v),\\ V(\tau;t,x,v)&\equiv -\bar{V}(\tau;t,1-x,-v).\end{aligned}$$

Hence $x_0=1$ implies that $\bar{\Gamma}(\tau;t,1-x,-v)$ will first intersect with $\{x=0\}$, and applying step (b) gives that $\bar{f}$ is C^1 near $(t,1-x,v)$. Thus f is C^1 near (t,x,v) by (2.19). This concludes the proof of the theorem. □

COROLLARY 2.5. *Let $E(t,x)\in C^{0,1}$ instead of C^1 in the assumptions in Theorem 2.4. Then $f\in C^{0,1}$ and satisfies (2.15).*

Proof. We construct a C^1 approximation for E. Let E_ϵ be a family of C^1 functions such that $E_\epsilon\to E$ in C^0 and $\|E_\epsilon\|_{W^{1,\infty}}\le 2\|E\|_{W^{1,\infty}}$. We finally choose an approximation of E as

$$E_\epsilon(t,x)+\{E(t,0)-E_\epsilon(t,0)\}(1-x)+x\{E(t,1)-E_\epsilon(t,1)\}, \tag{2.21}$$

which keeps the same boundary values as $E(t,x)$ so that the compatibility conditions (2.14) still hold. Applying Theorem 2.4 to the above electric field (2.21), we deduce the corollary as $\epsilon\to 0$. □

By the same proof of Theorem 2.4 with extra flatness assumptions for both f_0 and g near $\{v=0\}$, we easily have the following.

Remark 2.6. Let $f_0(x,v)$, $g(t,v)$, $E(t,x)$ be C^1. Assume (2.14) and (2.13). Assume there is a constant $C>0$, such that for $|v|\le 1$,

$$\begin{aligned}&|\nabla g(t,v)|\le C|v|^{1+\delta},\\ &\nabla f_0(x,v)|\le C(|x|+|v|)^{1+\delta},\qquad |\nabla f_0(x,v)|\le C(|1-x|+|v|)^{1+\delta}\end{aligned} \tag{2.22}$$

for some $\delta>0$. Then the solution to (2.1), $f(t,x,v)$, is $C^1(\Pi)$ and satisfies (2.15).

3. Linear *BV* estimate. For $E(t,x)$ without sign conditions (2.13), the solution f to (2.1) is not continuous in general (see [9]). To characterize such singularity, we now turn to a BV estimate of discontinuous solutions to (2.1). By using a finite difference scheme, we first establish a BV estimate for $f(t,\cdot,\cdot)$ under rather general conditions on the data f_0 and g. We introduce rectangular meshes in which the grid points are

$$\begin{aligned}t_n&=n\Delta t,\quad n=0,1,\dots,N,\\ x_i&=i\Delta x,\quad i=0,1,\dots,I,\\ v_j&=j\Delta v,\quad j=0,\pm1,\dots,\end{aligned}$$

where

$$\Delta t=T/N,\qquad h\equiv\Delta x=1/I=\Delta v>0.$$

We first define the initial boundary conditions as

$$\begin{aligned} f_{ij}^0 &= f_0(x_i, v_j); \\ f_{0,j}^n &= g_j^n, \quad j \geq 1; \\ f_{I,j}^n &= 0, \quad j \leq -1. \end{aligned} \tag{3.1}$$

We construct the following explicit up-wind scheme for $n \geq 0$.

$$\begin{aligned} &\frac{f_{ij}^{n+1} - f_{ij}^n}{\Delta t} + v_j^+ \frac{f_{ij}^n - f_{i-1,j}^n}{\Delta x} + v_j^- \frac{f_{i+1,j}^n - f_{ij}^n}{\Delta x} \\ &\quad + [E_i^n]^+ \frac{f_{ij}^n - f_{i,j-1}^n}{\Delta v} + [E_i^n]^- \frac{f_{i,j+1}^n - f_{ij}^n}{\Delta v} = 0 \end{aligned} \tag{3.2}$$

for both the interior region of $i = 1, \ldots, I-1,\ j \in Z$, and the two parts of outgoing boundaries: for $i = 0,\ j \leq 0$ and for $i = I,\ j \geq 0$. Here $[\cdot]^{\pm}$ denotes the positive (negative) part. It is important to note that by (3.1), both $f_{0,0}^n$ and $f_{I,0}^n$ are not explicitly given in terms of g, and they are determined only through (3.2) via f_{ij}^{n-1}.

If $\mu = \Delta t/h$, the scheme (3.2) can be written equivalently:

$$\begin{aligned} f_{ij}^{n+1} &= (1 - \mu v_j^+ + \mu v_j^- - \mu [E_i^n]^+ + \mu [E_i^n]^-) f_{ij}^n \\ &\quad + \mu v_j^+ f_{i-1,j}^n - \mu v_j^- f_{i+1,j}^n + \mu [E_i^n]^+ f_{i,j-1}^n - \mu [E_i^n]^- f_{i,j+1}^n. \end{aligned} \tag{3.3}$$

LEMMA 3.1. *Let f_0 and g have compact support in v,*

$$\{f_0(x, v) = 0,\ g(t, v) = 0,\ |v| \geq A\}. \tag{3.4}$$

If

$$\mu \leq \mu_0 = \frac{1}{2[A + \|E\|_\infty]}, \tag{3.5}$$

then (3.3) *is a positive scheme for* $0 \leq t \leq 1/2$.

Proof. For the first time-step (i.e., from f_{ij}^0 to f_{ij}^1), support in v of f_{ij}^1 is expanded to $[-A-h, A+h]$. In general, after k steps, the v-support of f_{ij}^k grows to

$$|v_j| \leq A + kh.$$

Notice that $\mu = \Delta t/h$. Hence for $0 \leq t \leq 1/2$, the number of time-steps is $\frac{1}{2h\mu}$, and support of v is bounded by

$$A + \frac{1}{2\mu}.$$

Therefore, (3.3) is a positive scheme if

$$\mu \left[A + \frac{1}{2\mu} \right] + \mu \|E\|_\infty \leq 1,$$

which is equivalent to $\mu \leq \mu_0$. We thus deduce our lemma. ∎

From the basic property of a positive scheme, we have the following lemma.

LEMMA 3.2. *Let $f_0 \geq 0$ and $g \geq 0$, and assume* (3.4). *If $\mu \leq \mu_0$ in* (3.5), *then $\|f^{n+1}\|_\infty \leq \|f^n\|_\infty$. Moreover, for $0 \leq t \leq 1/2$,*

$$\max_{i,j} |f_{ij}^n| \leq \{\max |f_0| + \max |g|\} \quad \textit{and } f_{ij}^n \geq 0.$$

In order to take into account the complicated boundary contributions at $x = 0$ and $x = 1$, we extend the boundary data to $i = -1$ or $x = -\Delta x$ as

$$f^n_{-1,j} = f^n_{0,j} = g^n_j, \quad j \geq 1, \qquad f^n_{-1,0} = f^n_{0,0}, \tag{3.6}$$

and to $i = I + 1$ or $x = 1 + \Delta x$ as

$$f^n_{I+1,j} = f^n_{I,j} = 0, \quad j \leq -1, \qquad f^n_{I+1,0} = f^n_{I,0}. \tag{3.7}$$

We can rewrite formula (3.3) for all $0 \leq i \leq I$ and $j \in Z$ as follows:

$$\begin{aligned} f^{n+1}_{ij} &= (1 - \mu v^+_j + \mu v^-_j - \mu[E^n_i]^+ + \mu[E^n_i]^-) f^n_{ij} \\ &\quad + \mu v^+_j f^n_{i-1,j} - \mu v^-_j f^n_{i+1,j} + \mu[E^n_i]^+ f^n_{i,j-1} - \mu[E^n_i]^- f^n_{i,j+1} + e^n_{ij}. \end{aligned} \tag{3.8}$$

Here (3.8) is free of boundary conditions, and the error term e^n_{ij} is constructed to contain the boundary information. In fact, e^n_{ij} is nonvanishing only if either $i = 0$ and $j \geq 1$; that is,

$$e^n_{0j} \equiv g^{n+1}_j - (1 - \mu[E^n_0]^+ + \mu[E^n_0]^-) g^n_j - \mu[E^n_0]^+ f^n_{0,j-1} + \mu[E^n_0]^- g^n_{j+1}, \tag{3.9}$$

or $i = I$ and $j = -1$; that is,

$$e^n_{I,-1} = \mu[E^n_I]^- f^n_{I,0}. \tag{3.10}$$

We denote the forward difference as

$$\Delta h(l) = h(l+1) - h(l)$$

for any function h, with suitable superscripts for different variables. We then estimate e^n_{ij} as follows.

LEMMA 3.3.

$$\sum_{i,j} |e^n_{ij}| \leq \sum_{j=1}^{\infty} |\Delta^t g^n_j| + 2\mu \|E\|_\infty \left\{ \sum_{j=1}^{\infty} |\Delta^v g^n_j| + \|g\|_\infty + \max_{i,j} |f^n_{ij}| \right\}.$$

Proof. By definitions (3.9) and (3.10),

$$\sum_{i,j} |e^n_{ij}| = \sum_{j \geq 1} |e^n_{0j}| + |e^n_{I,-1}|.$$

For $j \geq 2$, $f^n_{0,j-1} = g^n_{j-1}$ and we have

$$\begin{aligned} |e^n_{0j}| &= |g^{n+1}_j - (1 - \mu[E^n_0]^+ + \mu[E^n_0]^-) g^n_j - \mu[E^n_0]^+ g^n_{j-1} + \mu[E^n_0]^- g^n_{j+1}| \\ &\leq |g^{n+1}_j - g^n_j| + \mu |E^n_0| \{ |g^n_j - g^n_{j-1}| + |g^n_{j+1} - g^n_j| \}. \end{aligned}$$

On the other hand, for $j = 1$ we have

$$\begin{aligned} |e^n_{01}| &= |g^{n+1}_1 - (1 - \mu[E^n_0]^+ + \mu[E^n_0]^-) g^n_1 + \mu[E^n_0]^+ f^n_{00} - \mu[E^n_0]^- g^n_2| \\ &\leq |g^{n+1}_1 - g^n_1| + \mu |E^n_0| \{ |g^n_2 - g^n_1| + |g^n_1| + |f^n_{00}| \}. \end{aligned}$$

Moreover, at $i = I$ we have

$$|e^n_{I,-1}| \leq \mu |E^n_I| f^n_{I0}.$$

Therefore, summing over $j \geq 1$, we deduce the lemma. ∎

We now estimate the total variation of f in i and j, which is defined by

$$TV[f^n] := TV_{x,v}(f^n) = \sum_{-\infty}^{\infty}\sum_{i=0}^{I-1} h|\Delta^x f_{ij}^n| + \sum_{-\infty}^{\infty}\sum_{i=0}^{I} h|\Delta^v f_{ij}^n|.$$

The main idea is based on Harten's lemma [11].

THEOREM 3.4. *Let $E(t,x) \in C^{0,1}$. Assume $0 \le g(t,v) \in BV \cap L^\infty(\gamma_0^+)$, $g_t, vg_v \in L^1(\gamma_0^+)$, and $0 \le f_0(x,v) \in BV \cap L^\infty(\Pi_0)$. Then*

$$\begin{aligned} TV[f(t)] \le C(\|E\|_{C^{0,1}})\Big\{ & TV[f_0] + \|f_0\|_\infty \\ & + \int_{\gamma_0^+} ([1+v]|g_v| + |g_t|)dvdt + \|g\|_\infty \Big\}. \end{aligned} \tag{3.11}$$

Proof. Without loss of generality, we assume $f_0, g \in C_c^1$ satisfying (3.4). Let $\mu \le \mu_0$ (3.5) and $0 \le t \le 1/2$ so that (3.3) is a positive scheme.

Taking Δ^v onto scheme (3.8), we first estimate the variation along the v direction. Notice that E_i^n does not depend on j. By the product rule,

$$\Delta^v(v_j^\pm \cdot f_{ij}) = v_j^\pm \Delta^v f_{ij} + \Delta^v[v_j^\pm] f_{i,j+1},$$

and we deduce from (3.8)

$$\begin{aligned} |\Delta^v f_{ij}^{n+1}| \le & \{1 - \mu v_j^+ + \mu v_j^- - \mu[E_i^n]^+ + \mu[E_i^n]^-\}|\Delta^v f_{ij}^n| \\ & + \mu v_j^+ |\Delta^v f_{i-1,j}^n| - \mu v_j^- |\Delta^v f_{i+1,j}^n| \\ & + \mu[E_i^n]^+ |\Delta^v f_{i,j-1}^n| - \mu[E_i^n]^- |\Delta^v f_{i,j+1}^n| \\ & + \mu|\Delta^v v_j^+||f_{i,j+1}^n - f_{i-1,j+1}^n| + \mu|\Delta^v v_j^-||f_{i+1,j+1}^n - f_{i,j+1}^n| + |\Delta^v e_{ij}^n|. \end{aligned}$$

We first fix j, sum over $0 \le i \le I$, and then sum over j. By making changes of dummy indices of i and j to make cancellations, we obtain

$$\begin{aligned} \sum_{j=-\infty}^{\infty}\sum_{i=0}^{I}|\Delta^v f_{ij}^{n+1}| \le & \sum_{j=-\infty}^{\infty}\sum_{i=0}^{I}|\Delta^v f_{ij}^n| + \mu\sum_{j=-\infty}^{\infty} v_j^+\{|\Delta^v f_{-1,j}^n| - |\Delta^v f_{Ij}^n|\} \\ & + \mu\sum_{j=-\infty}^{\infty} v_j^-\{|\Delta^v f_{0,j}^n| - |\Delta^v f_{I+1,j}^n|\} \\ & + \sum_{j=-\infty}^{\infty}\sum_{i=0}^{I}\{\mu h|\Delta^x f_{i,j}^n| + |\Delta^v e_{ij}^n|\}. \end{aligned}$$

But by (3.6) and (3.7),

$$\begin{aligned} & \mu\sum_{j=-\infty}^{\infty} v_j^+|\Delta^v f_{-1,j}^n| \le \mu\sum_{j\ge 1}^{\infty} v_j^+|\Delta^v g_j^n|, \\ & \sum_{j=-\infty}^{\infty} v_j^-|\Delta^v f_{I+1,j}^n| = h f_{I,0}^n. \end{aligned}$$

Moreover, applying Lemma 3.3 yields

$$\sum_{j=-\infty}^{\infty}\sum_{i=0}^{I}|\Delta^v e_{ij}^n| \le 2\sum_{j=-\infty}^{\infty}\sum_{i=0}^{I}|e_{ij}^n|$$
$$\le 2\sum_j|\Delta^t g_j^n| + 4\mu\|E\|_\infty\left\{\sum_j|\Delta^v g_j^n| + \|g\|_\infty + \|f_0\|_\infty\right\}.$$

We thus conclude that

$$\begin{aligned}&\sum_{j=-\infty}^{\infty}\sum_{i=0}^{I}|\Delta^v f_{ij}^{n+1}| - \mu\sum_{j=-\infty}^{0}v_j|\Delta^v f_{0j}^n| + \mu\sum_{j=0}^{\infty}v_j|\Delta^v f_{Ij}^n|\\ &\le \sum_{j=-\infty}^{\infty}\sum_{i=0}^{I}|\Delta^v f_{ij}^n| + \mu h\sum_{j=-\infty}^{\infty}\sum_{i=0}^{I-1}|\Delta^x f_{i,j}^n|\\ &\quad +(4\mu\|E\|_\infty + h)\left[\|f_0\|_\infty + \|g\|_\infty + \sum_{j=0}^{\infty}v_j|\Delta^v g_j^n|\right] + 2\sum_{j=0}^{\infty}|\Delta^t g_j^n|.\end{aligned} \tag{3.12}$$

Next we consider the variation along x direction. We now operate Δ^x onto (3.8). By the product rule,

$$\Delta^x(E_i^n f_i^n) = E_i^n(\Delta^x f_i^n) + (\Delta^x E_i^n)f_{i+1}^n,$$

and we have, for all $0 \le i \le I-1$ and $j \in Z$,

$$\begin{aligned}|\Delta^x f_{ij}^{n+1}| \le \{&1 - \mu v_j^+ + \mu v_j^- - \mu[E_i^n]^+ + \mu[E_i^n]^-\}|\Delta^x f_{ij}^n|\\ &+\mu v_j^+|\Delta^x f_{i-1,j}^n| - \mu v_j^-|\Delta^x f_{i+1,j}^n|\\ &+\mu[E_i^n]^+|\Delta^x f_{i,j-1}^n| - \mu[E_i^n]^-|\Delta^x f_{i,j+1}^n|\\ &+\mu|\Delta^x E_i^n|(|\Delta^v f_{i+1,j-1}^n| + |\Delta^v f_{i+1,j}^n|) + |e_{ij}^n|.\end{aligned}$$

Notice that since $E \in C^{0,1}$, $|\Delta^x E_i^n| \le \|E\|_{C^{0,1}}h$. Summing the above over $1 \le i \le I-1$, $-\infty < j < \infty$, we deduce

$$\begin{aligned}\sum_{-\infty}^{\infty}\sum_{i=1}^{I-1}|\Delta^x f_{ij}^{n+1}| \le &\sum_{-\infty}^{\infty}\sum_{i=1}^{I-1}|\Delta^x f_{ij}^n| + \sum_{-\infty}^{\infty}\mu v_j^+|\Delta^x f_{-1,j}^n| - \sum_{-\infty}^{\infty}\mu v_j^+|\Delta^x f_{I-1,j}^n|\\ &+\sum_{-\infty}^{\infty}\mu v_j^-|\Delta^x f_{0,j}^n| - \sum_{-\infty}^{\infty}\mu v_j^-|\Delta^x f_{I,j}^n|\\ &+\mu h\|E\|_{C^{0,1}}\sum_{-\infty}^{\infty}\sum_{i=1}^{I-1}(|\Delta^v f_{i+1,j-1}^n| + |\Delta^v f_{i+1,j}^n|) + \sum_{-\infty}^{\infty}\sum_{i=1}^{I-1}|e_{ij}^n|.\end{aligned}$$

Notice that from our construction in (3.6) and (3.7),

$$\sum_{j=-\infty}^{\infty}\mu v_j^+|\Delta^x f_{-1,j}^n| = \sum_{j=-\infty}^{\infty}\mu v_j^-|\Delta^x f_{I,j}^n| = 0.$$

We thus deduce

$$\begin{aligned}
&\sum_{-\infty}^{\infty}\sum_{i=0}^{I-1}|\Delta^x f_{ij}^{n+1}| - \sum_{j=-\infty}^{-1}\mu v_j|\Delta^x f_{0,j}^n| + \sum_{j=1}^{\infty}\mu v_j|\Delta^x f_{I-1,j}^n| \\
&\quad\le \sum_{-\infty}^{\infty}\sum_{i=0}^{I-1}|\Delta^x f_{ij}^n| + 2\mu h\|E\|_{C^{0,1}}\sum_{-\infty}^{\infty}\sum_{i=0}^{I}|\Delta^v f_{i,j}^n| \\
&\qquad + C(\|E\|_\infty + 1)\sum_{j=1}^{\infty}\{|\Delta^v g_j^n| + |\Delta^t g_j^n| + \|g\|_\infty + \|f_0\|_\infty\}.
\end{aligned} \tag{3.13}$$

Combining (3.12) and (3.13) and summing over n, we deduce (3.11) for $0 \le t \le 1/2$, since constants in (3.11) do not depend on A. By approximating f_0 and g by C_c^1 functions, we deduce our theorem for the time interval $[0, 1/2]$. For the general $[0, T]$, we repeat the above estimate a finite number of times. ∎

4. Nonlinear plane diode. We are ready to study the nonlinear plane diode problem (1.1)–(1.6). We construct an iterating sequence $E^m \equiv \partial_x\phi^m$ and f^m for $m = 1, 2, 3, \ldots$ as follows:

$$\partial_t f^{m+1} + v\partial_x f^{m+1} + \partial_x\phi^m\partial_v f^{m+1} = 0, \tag{4.1}$$

$$\begin{aligned}
&\partial_{xx}\phi^m = \int_{-\infty}^{\infty} f^m(t,x,v)\,dv, \\
&\phi^m(t,0) \equiv 0, \qquad \phi^m(t,1) \equiv \lambda(t), \\
&f^{m+1}(t,0,v) = g(t,v), \quad v > 0, \qquad f^{m+1}(t,1,v) = 0, \quad v < 0, \\
&f^{m+1}(0,x,v) = f_0(x,v)
\end{aligned} \tag{4.2}$$

with $\partial_{xx}\phi^0 = \int_{-\infty}^{\infty} f_0(x,v)\,dv$, starting with $f^1(t,x,v) \equiv f_0(x,v)$. We have the following uniform estimates in m.

LEMMA 4.1. *The following uniform estimates are valid.*

$$f^{m+1}(t,x,v) \le \max\{\|f_0\|_\infty, \|g\|_\infty\}, \tag{4.3}$$

$$\|f^{m+1}(t)\|_1 \le \|f_0\|_1 + \|vg\|_1, \tag{4.4}$$

$$\|\partial_x\phi^m(t)\|_\infty \le \lambda(t) + \|f_0\|_1 + \int_0^t\int_0^\infty vg, \tag{4.5}$$

$$\|v^p f^{m+1}(t)\|_\infty \le Ce^{p\{\|\lambda\|_\infty + \|f_0\|_1 + \|vg\|_1\}t}\{\|v^p f_0\|_\infty + \|v^p g\|_\infty + 1\}, \tag{4.6}$$

$$\|\partial\partial_x\phi^m(t)\|_\infty \le C_p\left[\|\lambda'\|_\infty + \|f^{m+1}(t)\|_\infty + \|v^p f_0\|_\infty + \|v^p g\|_\infty\right], \tag{4.7}$$

$$\begin{aligned}
TV[f^{m+1}(t)] \le C(\|\partial_x\phi^m\|_{C^{0,1}})\Bigg[&TV[f_0] + \int_0^t\int_0^\infty([1+v]|g_v| + |g_t|) \\
&+\|f_0\|_\infty + \|g\|_\infty\Bigg],
\end{aligned} \tag{4.8}$$

where $\partial = \partial_t$ and ∂_x in (4.7)*, and $p > 2$.*

Proof. Both (4.3) and (4.4) are straightforward from the Vlasov equation (4.1). The BV estimate (4.8) directly follows from Theorem 3.4.

We now consider (4.5). Notice that from (4.2),

$$\partial_x\phi^m(t,x) = \partial_x\phi^m(t,0) + \int_0^x \rho^m(t,y)\,dy. \tag{4.9}$$

Since $\phi^m(t,1)=\lambda(t)$ and $\phi^m(t,0)=0$, we have

$$\lambda(t)=\int_0^1 \partial_x\phi^m(t,x)dx=\int_0^1\Big\{\partial_x\phi^m(t,0)+\int_0^x\rho^m(t,y)dy\Big\}dx.$$

Solving $\partial_x\phi^m(t,0)$ above and inserting into (4.9) yield

$$\partial_x\phi^m(t,x)=\lambda(t)-\int_0^1\int_0^x\rho^m(t,y)dydx+\int_0^x\rho^m(t,y)\,dy. \tag{4.10}$$

Therefore,

$$\lambda(t)-\int_0^1\int_0^x\rho^m(t,y)dydx\le\partial_x\phi^m(t,x)\le\lambda(t)+\int_0^x\rho^m(t,y)\,dy.$$

We thus deduce (4.5) by estimating two integrals of $\rho^m(t,x)$, again by (4.4).

We now turn to (4.6). Multiplying (4.1) with $|v|^p$ gives

$$[\partial_t+v\partial_x+\partial_x\phi^m\partial_v]\{|v|^pf^{m+1}\}\le p|v|^{p-1}|\partial_x\phi^m|f^{m+1}.$$

By using (4.5) and a simple inequality, $|v|^{p-1}\le 1+|v|^p$ for $p>2$, we get

$$[\partial_t+v\partial_x+\partial_x\phi^m\partial_v]\{|v|^pf^{m+1}\}\le p\{\|\lambda\|_\infty+\|f_0\|_1+\|vg\|_1\}(1+|v|^pf^{m+1}).$$

Taking the L^∞ norms along the characteristic (2.2) with $E=\partial_x\phi^m$ on both sides and applying the Gronwall lemma, we deduce (4.6).

Now we prove (4.7). We first treat $\partial_{xx}\phi^m$ as

$$\begin{aligned}\partial_{xx}\phi^m&=\int_{-\infty}^{\infty}f^m(t,x,v)dv=\int_{|v|\le1}+\int_{|v|\ge1}\\&\le\|f^m\|_\infty+\|v^pf^m(t)\|_\infty\int_{|v|\ge1}|v|^{-p}dv\\&\le\|f^m\|_\infty+\frac{1}{p-1}\|v^pf^m(t)\|_\infty.\end{aligned}$$

Moreover, recalling (1.4) and (4.1), we have the charge continuity $\partial_t\rho^m+\partial_xj^m=0$. This implies $\partial^3_{txx}\phi^m=-\partial_xj^m$ and

$$\partial_{tx}\phi^m(t,x)=\partial_{tx}\phi^m(t,0)+j^m(t,0)-j^m(t,x). \tag{4.11}$$

We first estimate $j(t,x)$ on the left of (4.11) by

$$\begin{aligned}|j^m(t,x)|&\le\int_{|v|\le1}+\int_{|v|\ge1}\\&\le\|f^m\|_\infty+\|v^pf^m(t)\|_\infty\int_{|v|\ge1}|v|^{1-p}dv\\&\le\|f^m\|_\infty+\frac{1}{p-2}\|v^pf^m(t)\|_\infty.\end{aligned}$$

Taking the t derivative of (4.10) yields

$$\partial_{tx}\phi^m(t,0)=\lambda'(t)-\int_0^1\int_0^x\rho^m(t,y)dydx=\lambda'(t)+\int_0^1(j^m(t,x)-j^m(t,0))dx.$$

Plugging this back into (4.11) implies (4.7). ∎

We now prove the uniqueness of the nonlinear dynamics of a diode, modeled by equations (1.1), (1.5), and (1.6).

THEOREM 4.2. *Assume that for some $p > 2$, $f_0(x,v)$ and $g(t,v)$ satisfy*

$$TV[f_0] + \int_{\gamma_0^+} ([1+v]|g_v| + |g_t|) + \|v^p f_0\|_\infty + \|v^p g\|_\infty < \infty. \tag{4.12}$$

There exists the unique solution $f(t,x,v)$ to (1.1)–(1.6), which satisfies (4.3)–(4.8).

Proof. We construct the approximate solutions f^m as in (4.1) and (4.2). Taking $m \to \infty$, by (4.3)–(4.8) we obtain a weak solution $f(t,x,v)$ to the nonlinear plane diode problem (1.1)–(1.6), which also satisfies the same estimates as in (4.3)–(4.8). Assume there is another such weak solution $f_1(t,x,v)$ satisfying the same estimates. Subtracting $f - f_1$ yields

$$[\partial_t + v\partial_x + \partial_x\phi_f]\{f - f_1\} = -\{\partial_x\phi_f - \partial_x\phi_{f_1}\}f_v, \tag{4.13}$$

$\{f - f_1\}|_{\gamma_0^+} \equiv \{f - f_1\}|_{\gamma_1^-} \equiv 0$, $\{f - f_1\}(0,x,v) \equiv 0$, and $\{\phi_f - \phi_{f_1}\}(t,0) \equiv \{\phi_f - \phi_{f_1}\}(t,1) \equiv 0$. Notice that from the Poisson equation,

$$\|\partial_x\phi_f - \partial_x\phi_{f_1}\|_\infty \le C\|f - f_1\|_1.$$

Moreover, for any test function $\eta(x,v)$, since $\partial_x\phi_f(t) - \partial_x\phi_{f_1}(t)$ does not depend on v, we deduce for any fixed t

$$\begin{aligned}
&\left|\int_0^1 \int_{-\infty}^{\infty} \{\partial_x\phi_f(t) - \partial_x\phi_{f_1}(t)\} f\partial_v\eta\right| \\
&\quad\le \|\partial_x\phi_f(t) - \partial_x\phi_{f_1}(t)\|_\infty \int_0^1 \left|\int_{-\infty}^{\infty} f\partial_v\eta\right| \\
&\quad\le \|\partial_x\phi_f(t) - \partial_x\phi_{f_1}(t)\|_\infty \int_0^1 \|\eta(x,\cdot)\|_{C^0}\|f(t,x,\cdot)\|_{BV}dx \\
&\quad\le \|\partial_x\phi_f(t) - \partial_x\phi_{f_1}(t)\|_\infty \|\eta\|_{C^0}\|f(t)\|_{BV}.
\end{aligned}$$

Therefore, $-\{\partial_x\phi_f - \partial_x\phi_{f_1}\}f_v$ is a measure over $[0,1]\times\mathbf{R}$, and

$$\|-\{\partial_x\phi_f - \partial_x\phi_{f_1}\}f_v\|_m \le \|\partial_x\phi_f(t) - \partial_x\phi_{f_1}(t)\|_\infty \|f(t)\|_{BV}.$$

Since taking the measure norm m on both sides of (4.13), we obtain

$$\begin{aligned}
\frac{d}{dt}\|f(t) - f_1(t)\|_1 &\le \|\{\phi_f - \phi_{f_1}\}f_v\|_m \le \|\phi_f(t) - \phi_{f_1}(t)\|_\infty \|f(t)\|_{BV} \\
&\le C\|f(t) - f_1(t)\|_1.
\end{aligned}$$

Therefore, the uniqueness follows from the Gronwall lemma. ∎

Now we study the question of higher regularity of $f(t,x,v)$. We are particularly interested in the conditions under which the solution $f(t,x,v)$ could become classical, without discontinuity.

LEMMA 4.3. *Assume f_0, $g \in C_c^1$ such that $|||[f_0,g]||| < +\infty$, which also satisfy (2.14) and (4.12). Furthermore, let*

$$f_0(x,v) \ne f_0(0,0)$$

for $(x,v) \neq (0,0)$ and x and v small. If the external voltage further satisfies

$$\lambda(0) > \int_0^1 \int_{-\infty}^{\infty} (1-x) f_0(x,v) dv dx,$$

then $f(t,x,v)$ is not continuous.

Proof. Since by (4.9) and (4.10),

$$\partial_x \phi(0,0) = \lambda(0) - \int_0^1 \int_{-\infty}^{\infty} (1-x) f_0(x,v) dv dx > 0,$$

the backward trajectory $\Gamma(\tau; t, 0, 0)$ hits on initial plane $\{t = 0\}$ for t sufficiently small. Therefore

$$f(t,0,0) = f_0(0, X(\tau;t,0,0), V(\tau;t,0,0)) \neq f_0(0,0).$$

But from (2.22), we have that $f(t,0,0) = f(0,0,0) = f_0(0,0)$, which is a contradiction. ∎

Next we consider the case when the external voltage $\lambda(t)$ is absent.

THEOREM 4.4. *Let $\lambda(t) \equiv 0$. Assume $f_0, g \in C^1$, such that $|||[f_0, g]||| < \infty$, which also satisfy* (2.14) *and* (4.12). *Then $f(t,x,v) \in C^1(\Pi \backslash \gamma_S) \cap C^0(\Pi)$.*

Proof. Notice that $\partial_{xx}\phi = \rho \geq 0$, and $\phi(t,0) = \phi(t,1) = 0$. It follows that $\partial_x \phi(t,0) \leq 0$, and $\partial_x \phi(t,1) \geq 0$. By Corollary 2.5, it follows that $f \in C^{0,1}(\Pi)$. To show $\partial_x \phi$ is C^1, we use (4.6) to get

$$\begin{aligned} |\rho(t,x) - \rho(t_1,x_1)| &= \left| \int f(t,x,v) dv - \int f(t_1,x_1,v) dv \right| \\ &\leq \int_{|v| \leq N} |f(t,x,v) - f(t_1,x_1,v)| dv + 2\|v^p f(t)\|_\infty \int_{|v| \geq N} |v|^{-p} dv \\ &\leq \int_{|v| \leq N} |f(t,x,v) - f(t_1,x_1,v)| dv + C N^{-p+1} \|v^p f(t)\|_\infty. \end{aligned}$$

By choosing N large and then letting (t_1, x_1) tend to (t,x), we deduce that $\rho(t,x)$ is continuous. So is $j(t,x)$, and we deduce $\partial_x \phi(t,x) \in C^1$ by the Poisson equation $(\partial_x^2 \phi = \rho)$ and (4.11). Our theorem follows from Theorem 2.4. ∎

The next theorem shows that for some given f_0 and g, $f \in C^1$ if the external voltage is small enough with respect to f_0 and g.

THEOREM 4.5. *Assume $f_0, g \in C^1$, such that $|||[f_0, g]||| < \infty$, which also satisfy* (2.14) *and* (4.12). *Furthermore, assume that f_0 and g satisfy the following nonvanishing condition: there are positive constants g_0, d, and $d_1 < 1$, such that*

$$\min_{0 \leq v \leq d} g(t,v) \geq g_0, \qquad \min_{0 \leq v \leq d, 0 \leq x \leq d_1} f_0(x,v) \geq g_0. \tag{4.14}$$

If we further assume

$$\|\lambda\|_\infty \leq \frac{(1-d_1)}{24} \min \left\{ \frac{g_0 d^3}{\sqrt{a^2 + \frac{(1-d_1) g_0 d^3}{12}} + a}, \frac{64}{3}(1-d_1) g_0^2 d_1^3 \right\}, \tag{4.15}$$

where $a = \|f_0\|_1 + \|vg\|_1$, then $f(t,x,v) \in C^1(\Pi \backslash \gamma_S) \cap C^0(\Pi)$.

Proof. By the proof of Theorem 4.4, it suffices to show that $\partial_x\phi$ points outward at $x = 0$ and $x = 1$. Notice that from $\partial_{xx}\phi = \rho \geq 0$ and $\phi(t, 0) = 0$, it follows that $\partial_x\phi(t, 1) \geq \dot{0}$.

To consider $\partial_x\phi(t, 0)$, by integration by parts we have

$$\begin{aligned}\partial_x\phi(t,0) &= \lambda(t) - \int_0^1 \int_0^x \rho(t,y)\,dydx \\ &= \lambda(t) - \int_0^1 \int_{-\infty}^{\infty} (1-x) f(t,x,v)\,dvdx.\end{aligned} \tag{4.16}$$

We consider the starting point (t_0, x_0, v_0) of (t, x, v) by separating two cases. If

$$\frac{d}{4\|\partial_x\phi\|_\infty} \leq d_1 < 1,$$

we consider the following region of $\{0 \leq v \leq \frac{d}{2}, 0 \leq x \leq \frac{v^2}{2\|\partial_x\phi\|_\infty}\}$. Clearly, from Lemma 2.1, the starting point (t_0, x_0, v_0) of such (t, x, v) satisfies

$$0 \leq x_0 \leq x \leq d_1, \qquad 0 \leq v_0 \leq 2v \leq d$$

so that by (4.14) $f(t_0, x_0, v_0) \geq g_0$. Restricting the integration region in the left-hand side (LHS) of (4.16) accordingly yields

$$\begin{aligned}\partial_x\phi(t,0) &\leq \lambda(t) - (1-d_1)\int_0^{d/2} \int_0^{\frac{v^2}{2\|\partial_x\phi\|_\infty}} f(t_0,x_0,v_0)\,dvdx \\ &\leq \lambda(t) - (1-d_1)g_0 \int_0^{d/2} \int_0^{\frac{v^2}{2\|\partial_x\phi\|_\infty}} dxdv \\ &\leq \lambda(t) - \frac{(1-d_1)g_0 d^3}{48\|\partial_x\phi\|_\infty} \\ &\leq \lambda(t) - \frac{(1-d_1)g_0 d^3}{48(\lambda(t) + \|f_0\|_1 + \|vg\|_1)},\end{aligned}$$

where we have used (4.5) in the last line. By solving a quadratic inequality, we deduce that if

$$0 \leq \lambda(t) \leq \frac{(1-d_1)g_0 d^3}{24\Big(\sqrt{a^2 + \frac{(1-d_1)g_0 d^3}{12}} + a\Big)}, \tag{4.17}$$

then $\partial_x\phi(t, 0) \leq 0$, where $a = \|f_0\|_1 + \|vg\|_1$.

On the other hand, if

$$\frac{d}{4\|\partial_x\phi\|_\infty} > d_1,$$

then we are restricted to the region $\{0 \leq x \leq d_1, 0 \leq v \leq \sqrt{2\|\partial_x\phi\|_\infty x}\}$. It follows that the starting point (t_0, x_0, v_0) of such (t, x, v) again satisfies

$$0 \leq x_0 \leq x \leq d_1, \qquad 0 \leq v_0 \leq 2v \leq d,$$

so that $f(t_0, x_0, v_0) \geq g_0$. Restricting the integration region in the LHS of (4.16) accordingly yields

$$\begin{aligned}
\partial_x \phi(t,0) &\leq \lambda(t) - (1-d_1) \int_0^{d_1} \int_0^{\sqrt{2\|\partial_x \phi\|_\infty x}} f(t_0, x_0, v_0)\, dv dx \\
&\leq \lambda(t) - (1-d_1) g_0 \int_0^{d_1} \int_0^{\sqrt{2\|\partial_x \phi\|_\infty x}} dv dx \\
&\leq \lambda(t) - \frac{2^{3/2}(1-d_1) g_0 \|\partial_x \phi\|_\infty^{1/2} d_1^{3/2}}{3} \\
&\leq \lambda(t) - \frac{2^{3/2}(1-d_1) g_0 \lambda(t)^{1/2} d_1^{3/2}}{3}.
\end{aligned}$$

Here $\|\partial_x \phi\|_\infty \geq \lambda(t)$, from $\phi(t,0) = 0$ and $\phi(t,1) = \lambda(t)$ as well as the mean value theorem. Therefore, $\partial_x \phi(t,0)$ is nonpositive if

$$\lambda(t) \leq \frac{8}{9}(1-d_1)^2 g_0^2 d_1^3.$$

Combining this with (4.17) concludes the theorem. □

REFERENCES

[1] N. B. Abdallah, *Weak solutions of the initial-boundary value problem for the Vlasov-Poisson system*, Math. Methods Appl. Sci., 17 (1994), pp. 451–476.

[2] F. Alabau, K. Hamdache, and Y. J. Peng, *Asymptotic analysis of the transient Vlasov-Poisson system for a plane diode*, Asymptot. Anal., 16 (1998), pp. 25–48.

[3] R. Alexandre, *Weak solutions of the Vlasov-Poisson initial-boundary value-problem*, Math. Methods Appl. Sci., 16 (1993), pp. 587–607.

[4] R. Beals and V. Protopopescu, *Abstract time-dependent transport equations*, J. Math. Anal. Appl., 121 (1987), pp. 370–405.

[5] J. Cooper and A. Klimas, *Boundary value problem for the Vlasov-Maxwell equations in one dimension*, J. Math. Anal. Appl., 75 (1980), pp. 306–329.

[6] P. Degond and P.-A. Raviart, *An asymptotic analysis of the Vlasov-Poisson system: The Child-Langmuir law*, Asymptot. Anal., 4 (1991), pp. 187–214.

[7] C. Greengard and P.-A. Raviart, *A boundary-value problem for the stationary Vlasov-Poisson equation: The plane diode*, Comm. Pure. Appl. Math., 43 (1990), pp. 472–507.

[8] Y. Guo, *Global weak solutions of the Vlasov-Maxwell system with boundary conditions*, Comm. Math. Phys., 154 (1993), pp. 254–263.

[9] Y. Guo, *Singular solutions of the Vlasov-Maxwell system on a half line*, Arch. Ration. Mech. Anal., 131 (1995), pp. 241–304.

[10] Y. Guo, *Regularity for the Vlasov equations in a half space*, Indiana Univ. Math. J., 43 (1994), pp. 225–320.

[11] A. Harten, *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys., 49 (1983), pp. 357–393.

[12] H.-J. Hwang, *Regularity of Vlasov–Poisson system in a convex domain*, SIAM J. Math. Anal., to appear.

[13] F. Poupaud, *Boundary value problems for the stationary Vlasov-Maxwell system*, Forum Math., 4 (1992), pp. 499–527.

[14] C.-W. Shu, *TVB boundary treatment for numerical solutions of conservation laws*, Math. Comp., 49 (1987), pp. 123–134.

SIAM J. MATH. ANAL.
Vol. 35, No. 6, pp. 1636–1643

REGULARITY OF AXIALLY SYMMETRIC FLOWS IN A HALF-SPACE IN THREE DIMENSIONS*

KYUNGKEUN KANG†

Abstract. We study axially symmetric solutions with no swirl of the three-dimensional Navier–Stokes equations in a half-space. We prove that *suitable weak solutions* in this case are Hölder continuous up to the boundary at all points except for the origin. For interior points this implies smoothness in the spatial variables. Hölder continuity at the origin remains as an open problem.

Key words. axially symmetric flow, Navier–Stokes equations, no swirl

AMS subject classifications. 76D03, 76D05

DOI. 10.1137/S0036141002414421

1. Introduction. We consider a vector field in a half-space $\mathbb{R}^3_+$ which vanishes at $\partial\mathbb{R}^3_+$. If a vector field u is invariant under rotation around the x_3-axis, we say that it is axially symmetric; in other words, $u(\mathbf{R}(x)) = \mathbf{R}(u(x))$ for every rotation $\mathbf{R}$ about the x_3-axis. If, moreover, an axially symmetric vector field v is invariant under reflection by every plane containing an x_3-axis, we say it is axially symmetric with no swirl; that is to say, v is axially symmetric and $v(\mathbf{T}(x)) = \mathbf{T}(v(x))$ for every reflection $\mathbf{T}$ as above. In this paper we study the regularity of axially symmetric solutions with no swirl of the Navier–Stokes equations in $\mathbb{R}^3_+$ with zero boundary condition on $\{x_3 = 0\}$. When $\mathbb{R}^3_+$ is replaced by $\mathbb{R}^3$, it is known that such solutions are regular (see [4], [8], and [18]).

Our main result is that *suitable weak solutions* of the Navier–Stokes equations are locally Hölder continuous up to the boundary $(\bar{\mathbb{R}}^3_+ \setminus \{\mathbf{0}\}) \times (0, \infty)$. It follows that in $\mathbb{R}^3_+ \times (0, \infty)$ the solutions are smooth in spatial variables x.

The main tools are the partial regularity results of *suitable weak solutions* of the Navier–Stokes equations (see [1], [5], and [11] for the interior case and see [14] for the boundary case) and the maximum principle for the azimuthal component of vorticity, which was also used to prove full regularity in the case of $\mathbb{R}^3$. Our result implies that the only possible singular point for axially symmetric solutions with no swirl in $\mathbb{R}^3_+$ would be the origin. It seems to be open whether or not singularity may occur at the origin, and so we leave it as an open problem. Our result could be also deduced from [14] combined with [12]. However, the method of proof presented in this paper, in order to use the maximum principle, is different from the one in [12].

The plan of the paper is as follows: In section 2, we introduce notation and definitions, review some well-known facts for our proof, and, finally, state our main theorem. In section 3, we present the proof of the main theorem.

2. Preliminaries and main result. In this section, we introduce notation and definitions, recall some well-known results used later, and, finally, state our main theorem. Let us begin with notation.

*Received by the editors September 12, 2002; accepted for publication (in revised form) October 3, 2003; published electronically April 7, 2004. This research was supported in part by NSF grant DMS-9877055.

http://www.siam.org/journals/sima/35-6/41442.html

†Department of Mathematics, University of British Columbia, Vancouver, BC V6T 1Z2 Canada (kkang@pims.math.ca).

• We denote by $\mathbb{R}^3_+$ a half-space of three dimension $\mathbb{R}^3$ and write the origin of $\mathbb{R}^3$ as $\mathbf{0}$.

• For a given point $(x,t) \in \mathbb{R}^3_+ \times I$, we denote by $B_{x,r} \subset \mathbb{R}^3_+$ the ball of radius r centered at x, where $0 < r < \operatorname{dist}(x, \partial\mathbb{R}^3_+)$. We also denote a parabolic ball by $Q_{(x,t),r} = B_{x,r} \times (t-r^2, t)$, where $0 < r < \min\{\operatorname{dist}(x, \partial\mathbb{R}^3_+), \sqrt{t}\,\}$.

• If x is located on the boundary of $\mathbb{R}^3_+$, then we write a half ball of radius r as $B^+_{x,r} = B_{x,r} \cap \mathbb{R}^3_+$, where $B_{x,r} \subset \mathbb{R}^3$. Similarly, if $(x,t) \in \partial\mathbb{R}^3_+ \times I$, a parabolic half ball at (x,t) is defined by $Q^+_{(x,t),r} = B^+_{x,r} \times (t-r^2, t)$ for $0 < r < \sqrt{t}$.

• Let $\Omega \subset \mathbb{R}^3$ be a domain. For $1 \le q \le \infty$, $W^{k,q}(\Omega)$ denotes the usual Sobolev space, i.e., $W^{k,q}(\Omega) = \{u \in L^q(\Omega) : D^\alpha u \in L^q(\Omega), 0 \le |\alpha| \le k\}$. As usual, $W^{k,q}_0(\Omega)$ is defined as the completion of $\mathcal{C}^\infty_0(\Omega)$ in $W^{k,q}(\Omega)$. We also denote by $W^{-k,q'}(\Omega)$ the dual space of $W^{k,q}_0(\Omega)$, where q and q' are Hölder conjugates.

• Let $1 \le q, r \le \infty$ and $I = (0,\infty)$. $L^r(I; W^{k,q}(\Omega))$ is the Banach space consisting of all measurable functions with a finite norm

$$||u||_{L^r(I;W^{k,q}(\Omega))} = \left(\int_I ||u(\cdot,t)||^r_{W^{k,q}(\Omega)} \, \mathrm{d}t \right)^{\frac{1}{r}}.$$

The Navier–Stokes equations are expressed in the Cartesian coordinates x, y, and z in a half-space $\mathbb{R}^3_+$ as follows:

$$\left.\begin{array}{c} u_t - \nu\Delta u + (u\cdot\nabla)u + \nabla p = f \\ \nabla \cdot u = 0 \end{array}\right\} \quad \text{in} \quad \mathbb{R}^3_+ \times (0,\infty) \tag{2.1}$$

with initial and boundary conditions

$$\left\{\begin{array}{cl} u(x,0) = u_0 & \text{when} \quad t = 0, \\ u = 0 & \text{on} \quad \partial\mathbb{R}^3_+ \times (0,\infty), \end{array}\right. \tag{2.2}$$

where $u : \mathbb{R}^3_+ \times (0,\infty) \to \mathbb{R}^3$ and $p : \mathbb{R}^3_+ \times (0,\infty) \to \mathbb{R}$ are unknown vector field and pressure, respectively, ν is the kinematic viscosity, and f and u_0 are prescribed external force and initial condition, respectively. From now on, f and u_0 are, for simplicity, assumed to be smooth and compactly supported, and we denote $I = (0,\infty)$ for simplicity. A solution u of (2.1) and (2.2) is called a *suitable weak solution* if u is a Leray–Hopf weak solution satisfying local energy inequality. In other words, u satisfies the following:

1. u belongs to the class

$$u \in L^\infty(I; L^2(\mathbb{R}^3_+)^3) \cap L^2(I; W^{1,2}_0(\mathbb{R}^3_+)^3), \quad u_t \in L^{\frac{4}{3}}(I; W^{-1,2}(\mathbb{R}^3_+)^3),$$

which is continuous in $t \in I$ in the weak topology in $L^2(\mathbb{R}^3_+)^3$, and solves (2.1) in a weak sense:

$$\int_{\mathbb{R}^3_+ \times I} (u\xi_t - \nu\nabla u : \nabla\xi - (u\nabla)u\xi + f\xi) \, \mathrm{d}x \, \mathrm{d}t = 0$$

for all $\xi \in \mathcal{C}^\infty_0(\mathbb{R}^3_+ \times I; \mathbb{R}^3)$ with $\nabla \cdot \xi = 0$ and

$$\int_{\mathbb{R}^3_+} u(\cdot,t)\nabla\phi \, \mathrm{d}x = 0$$

for all $\phi \in \mathcal{C}^\infty_0(\mathbb{R}^3_+)$ and a.e. $t \in I$.

2. The pressure p belongs locally to the class $p \in L^{\frac{3}{2}}_{\mathrm{loc}}(\mathbb{R}^+_3 \times I)$, i.e., $p \in L^{\frac{3}{2}}(Q^+_{z,r})$, where $z = (x,t) \in \mathbb{R}^+_3 \times I$ and $Q^+_{z,r} = B^+_{x,r} \times (t - r^2, t)$ for any $0 < r < \sqrt{t}$, such that u and p satisfy the following local energy inequality:

$$\begin{aligned} &\int_{\mathbb{R}^3_+} |u(x,t)|^2 \phi(x,t)\,\mathrm{d}x + 2\nu \int_{\mathbb{R}^3_+ \times I} |\nabla u|^2 \phi\,\mathrm{d}x\,\mathrm{d}t \\ &\le \int_{\mathbb{R}^3_+ \times I} |u|^2(\partial_t \phi + \nu \Delta \phi) + (|u|^2 + 2p) u \cdot \nabla \phi + 2f \cdot u\phi\,\mathrm{d}x\,\mathrm{d}t \end{aligned} \tag{2.3}$$

for a.e. $t \in I$ and for all nonnegative functions $\phi \in \mathcal{C}^\infty_0(Q^+_{z,r})$.

Before we go further, we make several remarks on "suitable weak solutions."

Remark 2.1. Our weak solutions are the Leray–Hopf solutions satisfying the local energy inequality. Therefore, these solutions also satisfy the global energy inequality (see, e.g., [2, pp. 71–72]); i.e.,

$$\frac{1}{2} \int_{\mathbb{R}^3_+} |u(x,t)|^2\,\mathrm{d}x + \nu \int_0^t \int_{\mathbb{R}^3_+} |\nabla u(x,t)|^2\,\mathrm{d}x\,\mathrm{d}t$$

$$\le \frac{1}{2} \int_{\mathbb{R}^3_+} |u_0(x)|^2\,\mathrm{d}x + \int_0^t \int_{\mathbb{R}^3_+} f(x,t) \cdot u(x,t)\,\mathrm{d}x\,\mathrm{d}t$$

for almost all $t \in I$.

Remark 2.2. For the regularity property of the pressure p, it is known that, under the reasonable assumption on f and u_0, the pressure p is in $L^{\frac{5}{3}}(Q)$, where Q is parabolic domain, which is $\mathbb{R}^3$, bounded, exterior, or half-space (see [16, Theorems 3.3 and 3.4] and [3, Theorem 3.1]). Therefore, it seems reasonable to assume that the pressure p is in $L^{\frac{3}{2}}_{\mathrm{loc}}(\mathbb{R}^+_3 \times I)$.

Remark 2.3. The existence of *suitable weak solutions* among weak solutions was proved in [1], and a slightly modified definition of it was observed in other contexts (see [5], [11], and [14]). In this paper, we follow the definition shown in [14]. As indicated in [1, Remark 4, p. 823], we do not know whether weak solutions are "suitable weak solutions"; in other words, it seems to be an open question whether weak solutions obtained by Galerkin approximation satisfy the local energy inequality (2.3).

Next we first define a regular or singular point of a *suitable weak solution* u.

DEFINITION 2.4. *We say a point $(x,t) \in \overline{\mathbb{R}}^3_+ \times I$ is a regular point when a* suitable weak solution *u is bounded in a neighborhood $Q_{(x,t),r}$ (or $Q^+_{(x,t),r}$) for some $0 < r < \min\{\mathit{dist}(x, \partial\mathbb{R}^3_+),\ \sqrt{t}\}$ (or $0 < r < \sqrt{t}$) for $x \in \mathbb{R}^3_+$ (or $x \in \partial\mathbb{R}^3_+$). Otherwise it is called a singular point. In addition, we say u is regular at (x,t) if it is a regular point. Similarly, we say u is singular at (x,t) if (x,t) is a singular point.*

It is well known that weak solutions are smooth in spatial variables and Hölder continuous in time in a neighborhood of an interior regular point (see [15]). At the boundary such solutions are Hölder continuous at each regular point, while the higher regularity seems to be open (see [14]).

On the other hand, it is also well known that weak solutions are smooth and unique for a short time for given smooth data f and u_0 (see, e.g., Theorem 3.2 in [17, p. 22] or Theorem 9.3 in [2, p. 80]). Here we recall a well-known result regarding a Hausdorff measure of possible singular set of time (see, e.g., [2], [9], [13], and [17]).

THEOREM 2.5. *Let u be a weak solution of the Navier–Stokes equations* (2.1). *Then there exists a closed set $\mathcal{S} \subset I$ whose $\frac{1}{2}$ dimensional Hausdorff measure vanishes such that u is regular in $\bar{\mathbb{R}}^3_+ \times (I \setminus \mathcal{S})$.*

Proof. See, for example, the proof of Theorem 10.8 in [2]. □

Remark 2.6. It should be mentioned that if *suitable weak solutions* which are axially symmetric in $\mathbb{R}^3_+$ have a singular point, then singularity can occur only on the x_3-axis. This argument is based on the result that the one-dimensional parabolic Hausdorff measure of a singular set is zero for the interior case proved in [1] (see also [5] and [11]) and for the boundary case proved recently in [14]. Therefore, it suffices to investigate the behavior of solutions near the x_3-axis, provided that it is axially symmetric.

We conclude this section by stating our main theorem, and its proof will be given in the next section.

MAIN THEOREM. *Let u be a* suitable weak solution *of* (2.1) *which is axially symmetric with no swirl in a half-space $\mathbb{R}^3_+$. Then u is regular for every $(x,t) \in \bar{\mathbb{R}}^3_+ \times I$ unless $x = \mathbf{0}$. Therefore, it is Hölder continuous in $(\bar{\mathbb{R}}^3_+ \setminus \{\mathbf{0}\}) \times I$.*

3. The proof of the main theorem. We recall that the partial regularity results imply that all points $(\mathbf{x}, t)$ with $x_1^2 + x_2^2 \neq 0$ are regular because, as mentioned earlier, singularity cannot happen away from the axis of symmetry for axially symmetric solutions. Therefore, we have only to prove that every point in $\{\, (\mathbf{x},t) \in \bar{\mathbb{R}}^3_+ \times I \,:\, x_3 > 0 \,\}$ is regular. In what follows, we consider only a fixed *suitable weak solution* u of (2.1) which is axially symmetric with no swirl. Here we assume that f and u_0 are smooth, compactly supported, and axially symmetric with no swirl. For convenience, we denote $\mathcal{Z}^+ = \{\, \vec{z} = (0,0,z) \in \mathbb{R}^3_+ \,:\, z > 0 \,\}$. Let us start with a simple observation.

LEMMA 3.1. *There exist two sequences $(z_i')_{i=1}^\infty$ and $(z_i'')_{i=1}^\infty$ such that $z_i' \searrow 0$ and $z_i'' \nearrow \infty$, and every point in $\{\, (\mathbf{x},t) \in \mathcal{Z}^+ \times I \,:\, x_3 = z_i' \text{ or } x_3 = z_i'' \,\}$ is regular.*

Proof. We show only the validity in the case of a decreasing sequence $(z_i')_{i=1}^\infty$. The other part can be proved by a similar argument. Suppose there is no such sequence. Then there is an interval $J_{\mathcal{Z}^+} \equiv (0,\delta)$ such that for every $\vec{z} = (0,0,z) \in \mathcal{Z}^+$ with $z \in J_{\mathcal{Z}^+}$, u is singular at $(\vec{z}, t_z)$ for some time $t_z \in I$. We collect all such points and denote them by

$$\mathcal{S}_\delta = \{(\vec{z}, t_z) \in \mathcal{Z}^+ \times I : u \text{ is singular at } (\vec{z}, t_z), \text{ where } \vec{z} = (0,0,z), z \in J_{\mathcal{Z}^+}\}.$$

Note that $\mathcal{S}_\delta$ is a subset of possible singular set and it can be easily checked that the one-dimensional parabolic Hausdorff measure of $\mathcal{S}_\delta$ is finite, not zero. In fact, $\mathcal{P}^1(\mathcal{S}_\delta) \geq \delta > 0$. However, in [1], it was proved that the one-dimensional parabolic Hausdorff measure of a possible singular set is zero, which leads to a contradiction. Therefore, such a sequence must exist. The existence of an increasing sequence $(z_i'')_{1=1}^\infty$ can be proved by a similar argument, and therefore we omit the details. This completes the proof. □

Let $\{\vec{z}_i^0 = (0,0,z_i')\}_{i=1}^\infty$ and $\{\vec{z}_i^\infty = (0,0,z_i'')\}_{i=1}^\infty$ be the sequences obtained in the previous lemma. Without loss of generality, we assume that $z_1' < z_1''$ because $z_i' \searrow 0$ and $z_i'' \nearrow \infty$ as $i \to \infty$. Next, we define a set $\mathcal{Z}_l^+ \subset \mathbb{R}^3_+$ as follows:

$$\mathcal{Z}_l^+ \equiv \{\vec{x} \in \mathcal{Z}^+ \,:\, z_l' < x_3 < z_l''\} \quad \text{for each} \quad l \in \mathbb{N}.$$

We also consider a set $\mathcal{S}_l \subset I$, which is related to $\mathcal{Z}_l^+$ and defined as follows:

$$\mathcal{S}_l \equiv \{t \in I \,:\, u \text{ is singular for some } (\vec{x}, t) \in \mathcal{Z}_l^+ \times I\} \quad \text{for each} \quad l \in \mathbb{N}.$$

Our aim is to show that $\mathcal{S}_l = \emptyset$ for all $l \in \mathbb{N}$, which implies our main result. Suppose that this is not the case. Then there exists $m \in \mathbb{N}$ such that $\mathcal{S}_m \neq \emptyset$, and then we consider

$$t_m \equiv \inf_{t\in I} \mathcal{S}_m, \quad \mathcal{S}_m \equiv \{t \in I : u \text{ is singular for some } (\vec{z}, t) \in \mathcal{Z}_m^+ \times I\}. \tag{3.1}$$

LEMMA 3.2. *Suppose $\mathcal{S}_m \neq \emptyset$ for some $m \in \mathbb{N}$. Let t_m be the number defined in* (3.1). *Then t_m is a strictly positive number in I and, moreover, there exists $\vec{z}_m \in \mathcal{Z}_m^+$ such that u is singular at $(\vec{z}_m, t_m)$.*

Proof. We note first that t_m is strictly bigger than 0 because u is smooth for a short time interval depending on given smooth data f and u_0 (see, e.g., Theorem 3.2 in [17, p. 22] or Theorem 9.3 in [2, p. 80]). We claim that there exists $\vec{z}_m = (0, 0, z_m) \in \mathcal{Z}_m^+$ such that u is singular at $(\vec{z}_m, t_m)$. Indeed, if t_m is isolated in $\mathcal{S}_m$, then obviously there is a point $\vec{z}_m \in \mathcal{Z}_m^+$ such that u is singular at that point. On the other hand, if t_m is a limit point in $\mathcal{S}_m$, then there is a sequence of point $(\vec{z}_{m,j}, t_{m,j})_{j=1}^{\infty} \in \mathcal{Z}_m^+ \times I$ such that $t_{m,j} \searrow t_m$. On the other hand, since $[z'_m, z''_m]$ is compact, $\{\vec{z}_{m,j}\}$ must have a limit point, say $\vec{z}_m$, and therefore an appropriate subsequence of $(\vec{z}_{m,j}, t_{m,j})$, which we relabel as $(\vec{z}_{m,j}, t_{m,j})$, converges to $(\vec{z}_m, t_m)$. We note that z_m must be located in $[z'_m, z''_m]$, that is, $z'_m \le z_m \le z''_m$. However, z_m cannot be z'_m nor z''_m, because z'_m and z''_m were chosen at the beginning to satisfy that u is regular at (z'_m, t) and (z''_m, t) for all $t \in I$. Our assertion is completed by noting that the singular set is closed. This completes the proof. ∎

Remark 3.3. It is worth noting that there exists a positive number r such that u is bounded in $B_{\vec{z}_m^0, r} \times (0, t_m]$ and $B_{\vec{z}_m^\infty, r} \times (0, t_m]$. Indeed, according to Theorem 2.5, there exists $t_0 > 0$ such that $u(\cdot, t)$ is regular everywhere, provided that $t < t_0$. Combining the facts that u is regular for all time at $\vec{z}_m^0$ and $\vec{z}_m^\infty$ and $[t_0, t_m]$ is a compact set, we can say that there exist positive numbers r_1, r_2 such that u is bounded in $B_{\vec{z}_m^0, r_1} \times (0, t_m]$ and $B_{\vec{z}_m^\infty, r_2} \times (0, t_m]$, respectively. By choosing $r = \min\{r_1, r_2\}$, we complete our claim.

The next step is to investigate the vorticity equation. If flow is axially symmetric with no swirl, then with the aid of the cylindrical coordinates

$$x = r\cos\theta, \quad y = r\sin\theta, \quad z = z,$$

velocity vector u is converted as follows:

$$u^r e_r + u^z e_z = u^x e_x + u^y e_y + u^z e_z,$$

where e_x, e_y, and e_z are the basis vectors with unit length in the Cartesian coordinates and e_r and e_z are the basis vectors with unit length in the cylindrical coordinates (note that azimuthal component u^θ vanishes because it has no swirl). In addition, each component satisfies the following relation:

$$u^r(r, z) = u^x \cos\theta + u^y \sin\theta, \quad u^z(r, z) = u^z(x, y, z).$$

We can also see that system (2.1) can be written in the cylindrical coordinates (see, e.g., [7, pp. 48–49]).

Now we consider the vorticity vector $w = \nabla \times u$. For simplicity, we assume that the outer force $f = 0$. The advantage of a flow with no swirl is that the vector w has only an azimuthal component, i.e., $w^\theta = u^r_z - u^z_r$. More precisely, $w = \nabla \times u = (0, w^\theta, 0) = (0, u^r_z - u^z_r, 0)$, and it solves the following single equation in $\mathbb{R}^3_+ \times I$:

$$\frac{\partial w}{\partial t} + u^r \frac{\partial w}{\partial r} + u^z \frac{\partial w}{\partial z} - \frac{u^r}{r} w - \nu \left[\frac{\partial^2 w}{\partial r^2} + \frac{\partial^2 w}{\partial z^2} + \frac{1}{r}\frac{\partial w}{\partial r} - \frac{w}{r^2} \right] = 0 \tag{3.2}$$

with smooth initial condition $w_0 = \nabla \times u_0$ and boundary $w = u_z^r - u_r^z$ on $\partial\mathbb{R}^3_+ \times I$. Now we define an axially scalar function $\xi \equiv w/r$, and simple calculations show that it satisfies the following equation in $\mathbb{R}^3_+ \times I$:

$$\xi_t - \nu\left(\frac{\partial^2\xi}{\partial r^2} + \frac{\partial^2\xi}{\partial z^2}\right) + u^r\frac{\partial\xi}{\partial r} + u^z\frac{\partial\xi}{\partial z} - 3\nu\frac{1}{r}\frac{\partial\xi}{\partial r} = 0. \tag{3.3}$$

The last term in (3.3) has a singular coefficient, and thus we cannot apply the maximum principle directly. The main trick is to extend ξ to be a function defined in $\mathbb{R}^3_+ \times \mathbb{R}^2 \times I$ by introducing another two variables z_1, z_2 such that the extended one is radial with respect to the variables x, y, z_1, and z_2. For clarity, we denote the extended function by $\tilde{\xi}$, which is defined as follows:

$$\tilde{\xi}(x, y, z_1, z_2, z, t) = \xi(r, z, t), \tag{3.4}$$

where $r^2 = x^2 + y^2 + z_1^2 + z_2^2$. In the same manner, we can also extend u^r, u^z, denoted by $\tilde{u}^r, \tilde{u}^z$, into $\mathbb{R}^3_+ \times \mathbb{R}^2 \times I$. Then, from (3.3), $\tilde{\xi}$ satisfies the following equation:

$$\tilde{\xi}_t - \nu\left(\frac{\partial^2\tilde{\xi}}{\partial r^2} + \frac{\partial^2\tilde{\xi}}{\partial z^2}\right) + \tilde{u}^r\frac{\partial\tilde{\xi}}{\partial r} + \tilde{u}^z\frac{\partial\tilde{\xi}}{\partial z} - 3\nu\frac{1}{r}\frac{\partial\tilde{\xi}}{\partial r} = 0. \tag{3.5}$$

This can be rewritten as follows:

$$\tilde{\xi}_t - \nu\tilde{\Delta}\tilde{\xi} + \tilde{u}^r\partial_r\tilde{\xi} + \tilde{u}^z\partial_z\tilde{\xi} = 0, \tag{3.6}$$

where $\tilde{\Delta}$ indicates the Laplace operator in five dimensions. To sum up, (3.3) is converted to five-dimensional parabolic equation (3.6) in $\mathbb{R}^3_+ \times \mathbb{R}^2 \times I$ with "good" coefficients if it is considered in a neighborhood of regular points. Note that (3.5) (or (3.6)) is reduced to (3.3) when $\mathbb{R}^3_+ \times \mathbb{R}^2 \times I$ is restricted to $\mathbb{R}^3_+ \times I$.

We argue as follows. We first show that $\tilde{\xi}$ is regular at $(\vec{z}_m, 0, 0, t_m)$, which implies that ξ is regular at $(\vec{z}_m, t_m)$, too. Therefore, u is also regular at $(\vec{z}_m, t_m)$, which is contrary to the assumption that u is singular at $(\vec{z}_m, t_m)$. Therefore, $\mathcal{S}_m$ must be empty, which makes our argument complete. Without any confusion, we denote $\vec{z}_m = (\vec{z}_m, 0, 0) \in \mathbb{R}^3_+ \times \mathbb{R}^2$ and $\mathbb{R}^3_+ \times \mathbb{R}^2 \times I = \mathbb{R}^4 \times \mathbb{R}_+ \times I$ by interchanging coordinates. Now we are ready to prove the main theorem.

Proof of the main theorem. We first show that $\tilde{\xi}$ is regular at $(\vec{z}_m, t_m)$. As mentioned in Remark 3.3, there exists a positive number r_1 such that u is bounded in $B_{\vec{z}^0_m, r} \times (0, t_m]$ and $B_{\vec{z}^\infty_m, r} \times (0, t_m]$ for all $0 < r \le r_1$. On the other hand, there exists $r_2 > 0$ such that u is smooth at $t = t_m - r_2^2$, which is due to Theorem 2.5 because the set of possible singular time is of $\frac{1}{2}$ Hausdorff measure zero. Without loss of generality, we may take $r_2 < r_1$. We denote r_2 by r and define $\Omega = [0, r) \times (z'_m, z''_m) \subset \mathbb{R}^4 \times \mathbb{R}_+$, where $[0, r) = \{\, y \in \mathbb{R}^4 : |y| < r \,\}$, and consider parabolic domains

$$Q = \Omega \times (t_m - r^2, t_m), \quad Q^\epsilon = \Omega \times (t_m - r^2, t_m - \epsilon),$$

where ϵ is an arbitrary small positive number with $\epsilon < r^2/4$.

We note first that $\tilde{\xi}$ is regular on $Q_0 \equiv \Omega \times \{t_m - r^2\}$ because of our choice of r. Hence $\tilde{\xi}$ is bounded on $\Omega \times \{t_m - r^2\}$. For convenience we denote $M_0 = \sup_{Q_0} |\tilde{\xi}|$. We also show that $\tilde{\xi}$ is bounded on another parabolic boundary of Q. Note that the other parabolic boundary of Q is composed of three parts, which are denoted by ∂Q_i for $i = 1, 2, 3$,

$$\partial Q_1 \equiv [0,r) \times \{z_m'\} \times (t_m - r^2, t_m),$$

$$\partial Q_2 \equiv [0,r) \times \{z_m''\} \times (t_m - r^2, t_m),$$

$$\partial Q_3 \equiv \{r\} \times (z_m', z_m'') \times (t_m - r^2, t_m).$$

It is obvious that $\tilde{\xi}$ is bounded on ∂Q_i for $i = 1, 2$ because z_m' and z_m'' were chosen in such a way that $\tilde{\xi}$ is regular on each ∂Q_i, $i = 1, 2$. In addition, since ∂Q_3 is the part strictly away from the z-axis and boundary, $\tilde{\xi}$ is also regular at every point on ∂Q_3, which implies that $\tilde{\xi}$ is bounded on ∂Q_3. Let $M_i = \sup_{Q_i} |\tilde{\xi}|$ for $i = 1, 2, 3$ and $M = \max\{M_i : i = 0, 1, 2, 3\}$. Now we claim that $\tilde{\xi}$ is bounded by M in a parabolic domain Q^ϵ. Indeed, $\tilde{\xi}$ is bounded on each parabolic boundary of Q^ϵ, denoted by $\partial_p Q^\epsilon$, because they are a subset of parabolic boundaries $\partial_p Q$.

On the other hand, since u is smooth in spatial variable and each spatial derivative is regular in Q^ϵ, so are $\tilde{\xi}$ and each spatial derivative of $\tilde{\xi}$, where we used that $\tilde{\xi}$ is axially symmetric. Therefore, $\tilde{\xi}_t$ is also bounded in Q^ϵ by (3.6), which enables us to apply the maximum principle to $\tilde{\xi}$ in Q^ϵ (see, e.g., Theorem 7.1 in [10, p. 156] and Chapter 3.7 in [6]). Therefore, we obtain $\sup_{Q^\epsilon} |\tilde{\xi}| \le \sup_{\partial_p Q^\epsilon} |\tilde{\xi}|$, which is bounded by M. Since the upper bound M is independent of ϵ, passing to the limit, we obtain $\operatorname{ess\,sup}_{\mathrm{Q}} |\tilde{\xi}| \le \mathrm{M}$. Therefore, there exists $\rho > 0$ such that $\tilde{\xi}$ is bounded by M in $Q_{(\vec{z}_m, t_m), \rho} \subset Q$, which means $\tilde{\xi}$ is regular at $(\vec{z}_m, t_m) \in \mathbb{R}^4 \times \mathbb{R}_+ \times I$. Thus, automatically ξ is regular at $(\vec{z}_m, t_m) \in \mathbb{R}^3_+ \times I$, which immediately implies that u is also regular at $(\vec{z}_m, t_m) \in \mathbb{R}^3_+ \times I$. However, this is contrary to the statement given in Lemma 3.2 under the assumption that $\mathcal{S}_m \neq \emptyset$ for some $m \in \mathbb{N}$. Hence $\mathcal{S}_l$ must be an empty set for all $l \in \mathbb{N}$. Hölder continuity follows that u is locally bounded. This completes the proof. □

Acknowledgment. The author expresses sincere gratitude to his thesis advisor, Professor Vladimír Šverák, for helpful discussions.

REFERENCES

[1] L. Caffarelli, R. Kohn, and L. Nirenberg, *Partial regularity of suitable weak solutions of the Navier-Stokes equations*, Comm. Pure Appl. Math., 35 (1982), pp. 771–831.
[2] P. Constantin and C. Foias, *Navier-Stokes Equations*, Chicago Lectures in Math., University of Chicago Press, Chicago, IL, 1988.
[3] Y. Giga and H. Sohr, *Abstract L^p estimates for the Cauchy problem with applications to the Navier-Stokes equations in exterior domains*, J. Funct. Anal., 102 (1991), pp. 72–94.
[4] O. A. Ladyzenskaja, *Unique global solvability of the three-dimensional Cauchy problem for the Navier-Stokes equations in the presence of axial symmetry*, Zap. Naucn. Sem. LOMI, 7 (1968), pp. 155–177 (in Russian).
[5] O. A. Ladyzenskaja and G. A. Seregin, *On partial regularity of suitable weak solutions to the three-dimensional Navier-Stokes equations*, J. Math. Fluid Mech., 1 (1999), pp. 356–387.
[6] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Uralceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr., 23, AMS, Providence, RI, 1968.
[7] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics*, Addison-Wesley, Reading, MA, 1959.
[8] S. Leonardi, J. Málek, J. Nečas, and M. Pokorný, *On axially symmetric flows in $\mathbf{R}^3$*, Z. Anal. Anwendungen, 18 (1999), pp. 639–649.
[9] J. Leray, *Sur le mouvement d'un liquide visqueux emplissant l'espace*, Acta Math., 63 (1934), pp. 193–248.
[10] G. M. Lieberman, *Second Order Parabolic Differential Equations*, World Scientific, River Edge, NJ, 1996.

[11] F.-H. Lin, *A new proof of the Caffarelli-Kohn-Nirenberg theorem*, Comm. Pure Appl. Math., 51 (1998), pp. 241–257.
[12] J. Neustupa and M. Pokorný, *An interior regularity criterion for an axially symmetric suitable weak solution to the Navier-Stokes equations*, J. Math. Fluid Mech., 2 (2000), pp. 381–399.
[13] V. Scheffer, *Partial regularity of solutions to the Navier-Stokes equations*, Pacific J. Math., 66 (1976), pp. 535–552.
[14] G. A. Seregin, *Local regularity of suitable weak solutions to the Navier-Stokes equations near the boundary*, J. Math. Fluid Mech., 4 (2002), pp. 1–29.
[15] J. Serrin, *On the interior regularity of weak solutions of the Navier-Stokes equations*, Arch. Rational Mech. Anal., 9 (1962), pp. 187–195.
[16] H. Sohr and W. Von Wahl, *On the regularity of the pressure of weak solutions of Navier-Stokes equations*, Arch. Math. (Basel), 46 (1986), pp. 428–439.
[17] R. Temam, *Navier-Stokes Equations and Nonlinear Functional Analysis*, 2nd ed., CBMS-NSF Reg. Conf. Ser. in Appl. Math. 66, SIAM, Philadelphia, 1995.
[18] M. R. Ukhovskii and V. I. Iudovich, *Axially symmetric flows of ideal and viscous fluids filling the whole space*, J. Appl. Math. Mech., 32 (1968), pp. 52–61.